1962 IRE International Convention Record

3

PART 4

Sessions Sponsored by

IRE Professional Groups on

Electronic Computers

Information Theory

at

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

March 26-29, 1962

The Institute of Radio Engineers

1962 IRE INTERNATIONAL CONVENTION RECORD

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

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

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

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

1962 IRE INTERNATIONAL CONVENTION RECORD

PART 4 - ELECTRONIC COMPUTERS; INFORMATION THEORY

TABLE OF CONTENTS

Page

Digital Computers: Components, Circuits, and Techniques
(Session 4: sponsored by PGEC)

Computers in Control and Simulation
(Session 12: sponsored by PGEC)

Artificial Intelligence: Recent Developments in Concepts and Hardware
(Session 20: sponsored by PGEC)

Panel: Status Session on Information Theory
(Session 34: sponsored by PGIT)

Page

Information Theory-II
(Session 49: sponsored by PGIT)

A NOVEL MICROWAVE COMPUTING TECHNIQUE

B. M. Mindes
ITT Fed. Labs.
Nutley, N. J.

R. T. Adams
Sichak Associates
Nutley, N. J.

Abstract

This paper describes a novel microwave computing technique with the potential ability to process data at a rate in excess of 1000 Mc. In this system the binary digits are represented by two UHF signals. Logic operations are performed by frequency conversions in a logic element composed of simple mixers and filters. This element will form Boolean functions of two, three or four variables and can change the function at the full clock rate, e.g., serve as an AND gate in one clock interval and a NOR gate in the next. The memory element is an oscillator capable of supporting oscillations at either logic frequency. Test circuits of the memory and logic elements have been constructed using tunnel diodes.

NCR 315 CURRENT MODE DIODE LOGIC BUILDING BLOCKS

G. H. Goldstick
T. T. Dao*
F. L. Ashford
The National Cash Register Company
Electronics Division
1401 East El Segundo Boulevard
Hawthorne, California

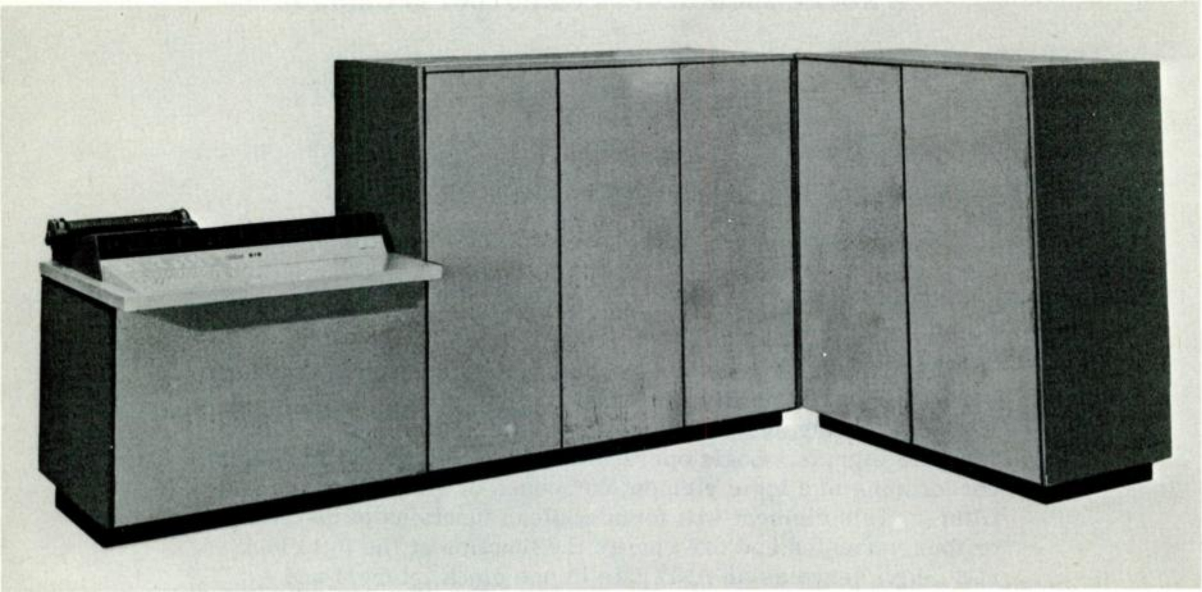


Fig. 1. National 315 electronic data processing system.

Summary

The logic of the NCR-315 Data Processing System is mechanized with two-level, current switching, negative, and-or gates, and saturated transistor inverting amplifiers. A schematic of the logical block is presented and the circuit operation qualitatively discussed.

The NCR-315 logic is mechanized with four principal building blocks, namely: 1) logic driver, 2) power driver, 3) double driver, 4) flip-flop. The operation of each of these building blocks is explained and their design procedure outlined.

The advantages of current mode logic over other systems are explained, and comparisons of the selected building block circuits with alternative approaches are presented.

Introduction

(1)The NCR C-315 Electronic Data Processing System (Figure 1) is medium size (40,000 word

* Mr. Dao is now employed by General Electric Computer Laboratory, Mountain View, California

memory capacity), medium speed (basic cycle time is six microseconds), and is designed for commercial data processing applications. Input and output equipment, including magnetic character sorter readers, high speed line printers, card punches and readers, magnetic tape and card machines, and a control console, are packaged as separate peripheral units to the main memory and processor. Versatility and flexibility are achieved by employing specific quantities of each of the peripheral units with the main memory and processor, and time sharing the input output information. Thus, every system is "custom assembled" to fulfill all of the requirements of each application.

The system is capable of reading or writing 100,000 alpha-numeric characters per second; two characters comprise a word, each word consists of twelve bits plus a parity bit. The system employs more than sixty basic commands to access the information programmed in the memory; one word at a time is accessed by each command. The control console continuously monitors the system operation, and permits manual interrogation of the memory, data storage in the memory, transmittance of programs to the processor, and modification of programs.

Standard circuit building blocks are utilized in the system design, and modular construction is employed in all equipment.

Basic Logic Block

Figure 2 shows the basic configuration of "negative-and" current mode diode logic mechanized with pnp transistors.[2, 3, 4] The operation of the logic gate is briefly as follows. The anodes of the product diodes are driven by output stages, such as Q_1, of the various building blocks. A proposition is defined as false when the output transistor, Q_1, of the associated building block is "on" (saturated), and true when the output transistor is "off" (collector clamped at V_{c1}). When one or more propositions which enter into a product are false, the current through the product resistor, R_p, is shunted to ground through one or more saturated transistors.

If all the products which constitute a sum are false, no product current is available to turn-on a driver circuit, such as Q_2. The current through R_s is large enough to provide the reverse current and develop sufficient reverse base-emitter voltage, (V_B), via diodes D_1 and D_2, to maintain the driver circuit (Q_2) off.

If one or more products which enter into a sum are true, the output resistors of the building blocks (R_1) reverse bias the corresponding product diodes, D_p, and the current in the product resistors, R_p, is routed through the corresponding sum diode, D_s.

The current in R_p is sufficiently greater than the current through R_s to forward bias and turn-on Q_2 within the specified time.

The proposition P in Figure 2 is given by:

$$P = \overline{A_1 \cdot A_2 \cdot A_3 + B_1 \cdot B_2 \cdot B_3 + C_1 \cdot C_2 \cdot C_3} \qquad (1)$$

Considerations which Dictated the Selection of Current Mode Building Blocks

Diode Logic as Opposed to Transistor Logic. A prime objective of the 315 Program was optimization for cost within the system and subsystem specifications which were established. It was clear at the inception of the program that the then prevailing ratio of transistor cost to diode cost of approximately ten, for the devices which were applicable to the circuit speeds contemplated, militated against a pure transistor logic (TRL) system. Moreover, the obvious speed and tolerance limitation of TRL made that mechanization unattractive for the 315 System. It was felt that the maximum performance per unit cost could be achieved with a diode logic system with transistors performing the amplification and inversion functions. Surprisingly enough, although both transistor and diode prices have decreased drastically, the ratio of transistor to diode cost for comparable speed devices has remained the same.

Current Mode Versus Voltage Mode Diode Logic. Voltage mode diode logic is characterized by signal levels which are large in comparison to the voltage drops across conducting diodes or forward biased base-emitter drops of transistors. Voltage mode systems are typically required to operate at signal swings of from 6V to 8V for purposes of efficiency.

The logical system shown in Figure 2 is usually classified as a current mode system in that information is propogated in the form of a relatively constant current. Thus, rather than a voltage level being high or low, a current is either present or absent.

In a current mode system a current source establishes the product current at the desired value, and the logical voltage swing is determined by the voltage drops across the logic and referencing diodes. Logical voltage swings in current mode systems need only be large enough to ensure the forward and reverse biasing of input transistors of the building blocks.

The advantages of a current mode system over a voltage mode system are discussed below.

The current required in the product resistor of any product-sum diode logic system may be written as follows:

$$^{**}I_p > I_{on} + I_{off} + {}^{***}\frac{\Delta V}{R_S} + \left[(m-1) + (b-1)\right] I_{SL} + I_R + 2\left[(m-1) + (b-1)\right] I_{SC} + 2\, I_S + nI_{PC} \qquad (2)$$

** FUNCTION OF THE BUILDING BLOCKS
*** FUNCTION OF THE LOGIC

I_{on} ==> Current required to turn-on the driven circuit and hold it on.

I_{off} ==> Current required to turn-off the driven circuit and hold it off.

$\frac{\Delta V}{R_S}$ ==> Current required to drive the sum resistor through the logical swing.

m ==> Number of diodes constituting a sum.

n ==> Number of diodes constituting a product.

b ==> Number of sums.

I_{SL} ==> Leakage current of the sum diode.

I_{SC} ==> The transient current which must be supplied per sum diode to charge (discharge) the sum diode transistion capacities through the logical swing in the time specified.

I_{PC} ==> The transient current which must be supplied per product diode to charge (discharge) the diode transition capacities through the logical swing in the time specified.

I_R ==> The additional transient current required to recover the minority carrier stored charge of the sum diodes.

I_s ==> The transient current which must be supplied to charge (discharge) the stray and input capacity at the driver through the logical swing in the time specified.

NOTE

> The recover of sum diodes and the charging (discharging) of capacities are series processes; hence, the current
> $2\left[(m-1) + (b-1)\right] I_{SC} + nI_{PC} + 2\ I_S + I_R$
> must first recover the sum diode and then charge the capacity present.

The current which must be supplied by the product resistor, along with the logical voltage swing, output capacity, number of products being driven, etc., determines the load on the driver, and ultimately the gain of the circuitry. The last seven terms of Equation 2 represent parasitic losses in the logic and should be minimized in order to maximize fan-out and fan-in.

a. Diode Leakage Currents. In current mode logic, reverse voltages seen by logical diodes are low, hence, diodes exhibit reverse currents which approximate the saturation current, permitting low voltage leakage current specifications and greater gain. A curve of leakage current versus reverse voltage for the 315 sum diode is shown in Figure 3.

b. Parasitic Capacities. In order to accomplish the logical swing, the stray and diode capacities must be charged and discharged. Since the total charge which must be supplied by the product resistor is:

$$Q_c = n \int_0^V C_{TP}\, dv + 2 \left[(m-1) + (b-1)\right] \int_0^V C_{TS}\ dv + 2\ C_S\ \Delta V \qquad (3)$$

where:

ΔV ==> The logical voltage swing

C_S ==> The total stray and input capacity at the driver input (card, wire, connector, etc.)

C_{TP},
C_{TS} ==> The transition capacities of product and sum diodes respectively.

C_{TP},
C_{TS} ==> are of the form:

$$C_T \approx A\ (V_R)^{-1/q}$$

where:

V_R ==> The reverse voltage across the diode
A, q are constants

hence,

$$Q_c = \Delta V \left[2C_S + \frac{nA_p}{\frac{q_p - 1}{q_p}} \cdot \frac{1}{V^{\frac{1}{q_p}}} + 2\left[(m-1) + (b-1)\right] \frac{A_s}{\frac{q_s - 1}{q_s}} \cdot \frac{1}{\Delta V^{\frac{1}{q_s}}} \right] \qquad (4)$$

Thus, the charge required increases somewhat slower than linearly. A curve of Q_c versus ΔV using the parameters of the 315 product and sum diodes is shown in Figure 4.

Assuming relatively constant current sources, $Q_c = I_D\ .\ t_c$:
Where I_D is equal to the sum of the last three terms in Equation 2, and t_c is the charge or discharge time.

Thus, for given gate speed (t_c), that portion of the gate current designated to charge capacity increases as per Figure 4, and causes corresponding decreases in the circuit gain.

c. Efficiency of Logic Gates. The product current lost due to logic swing is represented by the term $\frac{\Delta V}{R_S}$ in Equation 2. (Where V_c is assumed a constant.) However:

$$\frac{\Delta V}{R_S} = \frac{\Delta V}{V_S} \cdot \frac{V_s}{R_S} \qquad (5)$$

Where: V_s is the return voltage of the sum resistor.

Since the sum current $\frac{V_s}{R_s}$ is determined by current requirements of the driven current and the stray and diode capacities of the gate, the term $\Delta V/R_S$ may be reduced only by making the ratio $\frac{\Delta V}{V_S}$ small; i.e., by making the logic swing a small fraction of the return voltage of the sum resistor. A similar argument is applicable to the ratio $\frac{\Delta V}{R_p}$ which should be made small to decrease the load presented to the drivers. Hence, lower logical swings permit lower supply voltages for a given gate efficiency, or higher efficiencies for a given supply voltage. The efficiency of the CMDL gate, and selection of logic resistors and return voltages is discussed in more detail in Appendix I.

d. Diode Recovery. The charge stored in the sum diode in any logical system is:

$$Q_S = \tau_S I_p \qquad (6)$$

where:

τ_S ==> The charge storage time constant of the diode.

I_p ==> The current through the sum diode, e.g., the current in the product resistor.

Assume that all k-1 products in the sum which are initially true go false, and that the k^{th} product becomes true (instantaneously); the charge store in the k-1 diodes is:

$$Q_S = (k-1) \ \tau_S \ I_p \qquad (7)$$

the maximum charge which can be recovered from a diode conducting current I_p (assuming no active region width modulation) is(5)

$$Q_R = \tau_R \ I_p \qquad (8)$$

where:

τ_R ==> The maximum charge recovered per unit forward current.

The charge recovered from the k-1 sum diodes is $(k-1) \cdot \tau_R \cdot I_p$. This charge either is routed into the building block and tends to turn off the building block, or charges up the parasitic and stray capacity at the driven circuit's input: Both effects are deleterious since the k^{th} product is becoming true and has to remove from the building block or parasitic capacities the charge deposited by the k-1 recovery sum diodes.

Since this parasitic charge is proportional to the magnitude of the product current, so is the recovery current which must be provided to dissipate the charge in a prescribed interval. Hence, the minimization of product currents due to a, b, c, above allows relaxation of diode charge control specifications and higher operating speed of the logic gate.

e. Power. The lower supply voltage which are possible as well as the decreased gate currents (due to a decrease in that portion of the gate current which must drive leakage and capacitance for equivalent circuit speeds) result in considerably lower power requirements from the d-c logic supplies as well as lower power dissipation within the computer.

f. Noise. One objection invariably raised against low level logic systems is their susceptibility to "noise"; i.e., the lower the logical voltage swing, the greater percent of the logical swing is the induced noise voltage. The principle sources of noise can be classified briefly as follows:

1. Conductive Noise - usually due to the fact that several circuits share a common ground bus.

2. Capacitive Noise - due to the capacitive coupling between two supposedly independent circuits.

3. Electromagnetic and Inductive Noise - due to the mutual inductance between supposedly independent circuits or the self-inductance of the driving circuit.

4. Signal Distribution Noise - ringing on signal leads caused by the sending and receiving end of the lines not being terminated.

Assuming a given system wiring configuration, the amplitudes of the conductive and electromagnetic noise is proportional to the magnitudes of the currents being switched, and the capacitive crosstalk current and amplitude of ringing are proportional to the magnitudes of the voltage being switched.

Hence, capacitive noise is not a serious problem in a current mode system since the ratio of crosstalk current to signal current can be made quite high; ringing due to unterminated source and load impedance presents the same type of problem as in large voltage swing systems since the ratio of ringing amplitude to logic voltage swing is a constant dependent only on the characteristics of the transmission line and the magnitudes of the source and load impedances.

Conductive and inductive noise, however, does present more of a problem in low level circuitry(6), and extreme precautions are required to achieve low impedance transmission lines via careful power and ground distribution systems shielding of signal lines via twisted pairs, minimization of wire lengths, etc.

g. Cost. Current mode systems appear on the surface to have two significant cost advantages over equivalent voltage mode systems. These are: The lower forward currents and reverse voltages which the diodes see permits the use of less expensive diodes since (for equivalent circuit efficiencies and speeds) neither forward conductance nor leakage current requirements are stringent. However, it must be admitted that the high degree of confusion that characterizes pricing policies of diode manufacturers has precluded a quantitative determination of what this savings is.

The lower current levels and lower supply voltages which obtain (for equivalent speeds and efficiencies) (a, b, c, d, above) result in lower volt-amp requirements on the power supplies. Since power supply costs tend to vary linearly with volt-amp requirements , cost savings, are predictable.

The configuration of current mode logic selected ("negative and", "positive or") was based on the availability and projected continued cost reduction of a low cost germanium PNP switch, e.g., the MADT.

Basic Considerations Dictating Circuit Specifications, Choice of Components, and Circuit Design

System Environment

The 315 System is designed to operate in a commercial environment where the inlet air to any equipment is maintained between 64°F and 78°F, and the relative humidity is maintained between 40% and 60%, and the maximum temperature rise is less then 12°C within the units.

315 System Timing Diagram

The basic timing for the 315 System (processor and memory) is shown in Figure 5. The cycle time is six microseconds, approximately 1/3 of which is devoted to reading from the core memory, 1/3 to writing into the memory, 1/3 to logic. The (2.1) microsecond logic time is consumed as follows:

2 clock rises and flip-flop regenerations (A, D);

13 stages of logic and double drivers (B);

time to set the flip-flop clock gate (C).

Preliminary design efforts indicated that the clock rise time and flip-flop regeneration times could be held within .25 microseconds, and clock gate set time to . 35 microseconds. This dictates that the total time for propogating a signal through 13 stages of double drivers is 1.25 microseconds allowing 90 nanosecond (including stacking factor) per stage of two level logic and double drivers.

The 6 microsecond cycle time resulted in a very modest requirement on the flip-flop repetition rate.

Although the logical swing was to be minimized because of considerations a. through g. in the previous section, an even more severe requirement on the logical swing resulted from the method in which the control was mechanized. Figure 6 shows a brief schematic of the PCS (Program Count Sum) mechanization.

Each core in the matrix (1/2 mil 4-79 permalloy) corresponds to a given program count; cores are operated in linear select fashion. Switches at the top and bottom of the matrix are decoded and a voltage equal to 14V is impressed across the winding of the selected core. Single secondary windings are threaded through the various cores to form the desired control terms.

It was desired to use the output of the core matrix directly without any interstage amplification. The minimum secondary voltage from the selected core had to be equal to the logic swing and had to persist for the two microsecond duration of the logic. In addition, the secondary voltage of the selected core must develop, prior to the logic time, sufficient voltage greater than the logic swing to discharge the inductance of the rest of the matrix. The selected core is reset during the remainder of the cycle.

The drive current required from the switches as well as the reverse voltage which they must sustain is proportional to the switch core flux; reduction of core flux could only be accomplished by reducing the logic time A + B + C or reducing the logical voltage. In the final design of the PCS, the switches which are similar to 2N599 must provide 350 ma and sustain approximately 20V. If the core had been much larger it would have been necessary to use far more expensive (silicon) switches.

The load requirements on the logical circuits were studied by preparing histograms of input and output wire length, input and load capacity, and the number of gates being driven for each of the three principle building blocks (Double Driver, Power Driver, Flip-Flop). Wire length and DC loads (Figures 7 and 8) were determined from the preliminary logical design and card cage layout in the critical portions of the processor (adder, etc.). Input and output capacities were computed as described in Appendix II.

In order to optimize the speed of the adder, special adder cards were designed such that the input logic to a given double driver was located on the same board as the driver.

Requirement for Single Ended Input to Flip-Flop

Since the 315 is a parallel computer, many of the commands involve information transfers from one set of storage elements to another. In order to simplify the logic and reduce the number of logical diodes, a single ended flip-flop was employed. The input stage of the flip-flop and inductor clock gate desired (discussed below) was most conveniently implemented with a differential amplifier. The use of a differential amplifier resulted in the logic swing being 2.1V rather than the 1.5V which would have obtained had the flip-flop input stage been similar to that of the double driver or power driver.

Power Supplies

The power supplies specified for the 315 System all employ transistor regulators and are all specified to have total deviation from nominal less than ± 2%.

NOTE

> Total deviation includes variation to static line and load changes, dynamic line and load changes, ripple, stability, thermal variation and setting resolution.

All supplies employ short circuit protection circuitry and overvoltage protection circuitry to prevent development of voltages which might cause catastrophic failures in the event of a supply malfunction. The positive hold-off voltage (± 15V) is sensed for an undervoltage condition, and if such a condition materializes, all supplies are brought to ground.

Characteristics of Building Blocks and Components

The characteristics of the principle components which implement the logic networks and building blocks are the logic sum and product diodes which are similar to the 1N774 and 1N776 and the amplifiers which are similar to the 2N1499 and 2N1500 with controlled specifications. The characteristics which are common to all the building blocks are summarized below:

All logic gates terminate in a base network of a PNP switch (Figure 2). The input transistors of logic building blocks other than the inverter are nonsaturated.

The relative merits of saturated and non-saturated operation have been discussed elsewhere(7). It has been shown that the single disadvantage of saturated operation as compared to non-saturated operation is lowered switch-off efficiency; saturated circuits, however, exhibit better stability of the upper logical level, lower power dissipation, generate less noise, have better noise rejection capability, and are able to supply load current "on demand". The advantages of operating the output transistors of the building blocks in the saturated mode greatly exceed the single disadvantage of lowered switch-off efficiency; hence, the output transistors of all building blocks are saturated when "on". However, the above disadvantages of a non-saturated operation are not very significant with respect to stages isolated from the load; hence, the input stages of the double driver and power driver are not allowed to saturate. The ratio of the non-saturated to saturated switch-off efficiency of a current driven switch is:

$$\frac{(E_s \text{ off})_{\text{non sat}}}{(E_s \text{ off})_{\text{sat}}} = \frac{\frac{I_c}{\omega_t} + K_s \left[I_b - \frac{I_c}{\beta} \right] C_c \Delta V_c}{\frac{I_c}{\omega_t} + C_c \Delta V_c} \qquad (9)$$

Where I_b must be designed to satisfy the d-c "on", transient turn on, and load switching requirement. The maximum value of the ratio (adder double driver) was 2.4.

Design for Load Switching

Computer logic circuits are called upon to support varying loads depending upon the configuration of the logic; for example, assume a given driver, A, is assisted by N-1 other circuits in holding M gates false. Hence, A need only supply $\frac{M}{N}$ loads; however, at some time in the execution of a command, N-1 drivers go true (output transistors shut off) and driver A is called upon to support all M loads. The load switching operation described above is clearly different than switching a transistor from "off" to "on". What is required in the above operation is that a transistor previously "on" (saturated) remains "on" (saturated) as its load is increased. Although the phenomena which occur under load switching have been previously discussed,(8, 9,10) and parameters which describe the phenomena have been defined, the significance of this mode of operation in computer circuits and the design conditions which apply are generally not appreciated.

Two deleterious effects may occur if the logic circuitry is not designed for load switching.

a) If the output transistor of a flip-flop, which gates a regenerative element, (e.g., blocking oscillator) is called upon to switch a load that it cannot support, the output transistor comes out of saturation and the regenerative circuit may be triggered.

b) Failure to design for the load switching may cause turmoil in computer timing; for example, assume that during the add operation, one of the flip-flops assumed to be "on" comes out of saturation as a result of load switching; since the process is not valid again until the offending flip-flop saturates, the duration of the add operation is extended by the time required for the charge in the output transistor in question to build back up to the level that can support the final collector current.

For a transistor to remain saturated as its collector current is switched from a value I_{c1} to I_{c2}, it must have stored in its active region sufficient charge to support the final collector current; if sufficient charge is not available, the transistor comes out of saturation and the charge increases according to the active region time constant. The charge stored in a transistor supporting a collector load current, I_{c1}, and being driven by a base current, I_b, is:

$$Q_s = \frac{I_{c1}}{\omega_t\ (I_{c1})} + K_s \left[I_b - \frac{I_{c1}}{\beta} \right] \qquad (10)$$

The charge required to support the collector current I_{c2} is:

$$Q_{reqd} = \frac{I_{c2}}{\omega_t\ (I_{c2})} \qquad (11)$$

If the transistor is to remain in saturation as its collector current is instantaneously increased from I_{c1} to I_{c2}, then $Q_s \geq Q_{reqd}$.

Hence, I_b must be designed such that:

$$I_b \geq \frac{I_{c1}}{\beta} + \frac{1}{K_s} \left[\frac{I_{c2}}{\omega_t(I_{c2})} - \frac{I_{c1}}{\omega_t(I_{c1})} \right] \qquad (12)$$

If I_{c1} is taken as zero:

$$I_b > \frac{I_{c2}}{K_s \omega_t} \qquad (13)$$

The composite parameter $K_s\ \omega_t$ is defined as β_s, i.e., the allowable switched collector current per unit base current, the collector-base junction remaining forward biased. Note that β_s may be defined and measured independently. Alloy transistors of the 2N599, 2N661, 2N636, 2N404 variety yield excellent agreement of measured values of β_s and $K_s \cdot \omega_t$; unfortunately, the agreement is not very good for diffused base or mesa devices(8).

If sufficient standby base current is not provided and load switching occurs, the transistor comes out of saturation. Since it is initially able to support only the current $\omega_t\ Q_s$. The collector current, then builds up according to the relation:

$$i_c = \omega_t\ Q_s + \left[\beta\ I_b - \omega_t\ Q_s \right] \left[1 - e^{- \frac{t}{\omega_t} \beta} \right] \qquad (14)$$

Lower Clamped Outputs

Since the logic circuitry is current mode, precise definition of the lower logical level is really not necessary. However, the desire to optimize circuit speed dictated that the logical swing be maintained at its minimum required value and that a stable current source be provided at the output of the building blocks to discharge stray collector and product diode capacities. These two considerations resulted in clamped outputs for the building blocks.

Upper Clamps Inputs

The base networks of input transistors of all building blocks are upper clamped at approximately ground. Thus, the base emitter junctions of input transistors are protected against the permanent loss of the negative logic voltage (-15V); moreover, the signal swings at the input are restricted, which is desirable for speed considerations.

Mutually Exclusive Sums

The number of units of turn-on input current (1.5MA nominal) which are provided to the building block are equal to the number of products in the sum which are true. Since the turn-off current to the circuit is constant independent of the number of "true products" a speed degradation due to increased storage time may result when two or more products are "true". The flip-flop is least affected by multiple inputs, the power and double driver more so, and the inverter is most affected. This speed degradation is eliminated in the critical areas of the computer by writing the input logic so that it consists only of mutually exclusive terms.

Design Procedure and Philosophy

All logic circuits in the 315 Computer were synthesized analytically using a "worst case" design philosophy and employing design concepts similar to those described in (11)*. Designs were verified in the laboratory by employing marginal components. Although it was recognized that a "worst case" philosophy is wasteful in component tolerances and power, and does not lend itself to quantitative determination of reliability, this was the only philosophy and procedure which could be implemented at the time.

The objectives of the analytical design effort were threefold; namely,

to insure that the design engineer responsible had the most intimate and thorough knowledge of the circuit and component requirements;

to establish a one-to-one correspondence between component parameters and circuit characteristics;

to access and effect the cost tradeoffs of circuit specifications against component parameter specifications.

The charge control model of the transistor (10) was employed in the design of all switching circuits. The charge control model has a number of advantages over other semiconductor models (distributed, lumped, small signal) which renders it a powerful tool for the circuit designer; e.g., the parameters of the model may be correlated with recognized device parameters, may be evaluated directly using simple measurement procedures, is amendable to standard analysis techniques (LaPlace transform, etc.), and is sufficiently general so that the model

* (This does not imply that the 315 Circuits were designed via linear programming since they were not.)

topology is not a function of the particular physical and geometrical properties of the device under study.

Component Philosophy

The purchasing specifications define all significant parameters and detriments to the extent that the one-to-one correspondence between device parameters and circuit performance is insured. The purchasing specification is supplemented with a Design Standards Manual (D. S. M.), which provides a more extensive description of the device parameters which are necessary for design, but which are impractical to define in a purchasing specification. For example, transistors are described in terms of the minimum and maximum, or typical dependence of β, I_{EO}, I_{CO}, K_S, and V_{eb} versus temperature; β, f_t, and I_c for various values of V_{EC}; C_{TE}, C_{TC} versus V_{BE}, V_{BC}, etc. The measured characteristic curves of the various parameters are extrapolated through specific points to obtain the minimum and maximum parameter curves. Design Standards Manual information is confirmed periodically by sampling production stock.

Table I describes the various derating and tolerancing procedures used for the components. Total design tolerances include irreversible and reversible changes caused by temperature, load and storage, life, humidity, manufacturing, and equipment environment. The integrity of the D. S. M. for various devices (transistors and diodes especially) are controlled by strict vendor approval procedures.

Optimization Procedure

Initial design effort on the logic gates, double driver, and flip-flop resulted in the selection of -15 and +15 volts for the product and sum resistor network (see Appendix I), and the specification of the product resistor at 10K ± 1%. The sum resistor, R_S, however, is associated both electrically and physically with the building block. The selection could be attributed to the requirements of a particular building block. The speed of each particular building block was optimized by designing the sum resistor to minimize the skew between the turn-on and turn-off response of the building block.

Circuit Interconnection

The 315 System is fabricated using printed

TABLE I
Component Derating and Tolerancing Procedures

DEVICE	PARAMETER	DEFINITION	DESIGN VALUE USED
Transistor	$T_{j\ max}$	Max junction temperature	80% initial specified value
	V_{CBO}	Max collector base voltage, emitter current = 0	80% initial specified value
	$V_{CES\ max}$	Max collector emitter volt emitter-base	80% initial specified value
	min	Min common emitter DC given	80% initial specified value
	$I_{CO\ min}$	Max collector-base leakage current	Twice initial specified value
	$I_{EO\ max}$	Max base-emitter leakage current	Twice initial specified value
	Design values used for all transient parameters (K_S, W_T, C_{TE}, C_{TC}, etc.) were initial maximum specified values.		
Diodes	$T_{j\ max}$	Max junction temperature	80% mfg. specified value
	$V_{f\ max}$	Max fwd. voltage at specified current	Worst case extrapolated arrived through maximum control limits
	$I_{L\ max}$	Max leakage current at specified reverse voltage	Twice initial specified voltage
	Design values used for all transient parameters (C_T, R_O, τ_L, τ_R, etc.) were initial maximum values.		
± 5% Composition Resistors	P_{max}	Max power dissipation	50% mfg. rating at design temperature
	Λ	Resistor tolerance	8% includes the tolerance of initial soldering, moisture, shelf and load life at 50% derating.
± 1% Deposited Carbon Resistors	P_{max}	Max for dissipation	50% mfg. rating at design temperature
	Λ	Capacity tolerance	3%: includes the contributing of initial tolerance, effects of moisture, soldering, shelf and load life, etc.
± 5% Mica Capacitors	δ	tolerance	8%
+100% -10% Electrolytic Capacitors	E_{max}	Max voltage	80% mfg. rating at design temperature
	δ	Capacity tolerance	200% -20%: includes the effect of initial tolerance, effects of moisture, soldering, etc.
	$I_{L\ max}$	Max leakage current	Twice mfg. specified value at max temperature
± 5% Inductors	I_{max}	Max d-c current	50% rating at max temp.
	λ	Inductor tolerance	7%: effect of initial tolerance, effects of moisture, soldering etc.

circuit boards and wire-wrap connectors. Figure 9 shows a 315 Flip-Flop board and wire-wrap connector. Hook-up wire is twenty-six gage with Teflon insulation (selected to minimize stray capacity). Interconnection of circuitry is typically single wire point-to-point except in those instances where it is necessary to damp signal ringing rapidly and greatly reduce crosstalk. In these situations, two-wire lines, in which the return lead is twisted about the signal lead and grounded at both the driving and driven ends, are used. The twisted pair so employed serves as a partial electrostatic shield as well as providing a lower impedance communication line.

The 315 is a synchronous computer and employs a flip-flop binary energy storage clock gate and a bi-stable which is well isolated from the logic. Moreover, all regenerative circuits which operate directly from the logic (blocking oscillators, one shots, clocks) are gated, and inputs to these circuits are always carried over twisted pairs to minimize crosstalk. Hence, noise and ringing which occur between clock times do not result in computer malfunction.

Both the delay and distortion suffered by a signal propagating across a section of point-to-point hook-up wire, and spurious signals resulting from crosstalk, significantly influence computer timing. A section of point-to-point hook-up wire driven by a semiconductor switching circuit and terminated in an RC load is illustrated in Figure 10.

The input of Circuit B (Figure 10) is located at the end of the line. Both circuits (A and B) are designed such that:

$$v_o \leq V_L \text{ when } v_i \geq V_U \tag{15}$$

$$v_o \leq V_U \text{ when } v_i \geq V_L \tag{16}$$

where:

V_L = lower logic voltage

V_U = upper logic voltage

The response of the loaded line, which is an exponentially decaying sinusoid, is illustrated in Figure 11 along with the static worst case upper logic level, V_u.

The following two problems are apparent from Figure 11:

a. As illustrated, the termination of the line is quiescently at the upper logic level only after t_p. Time t_s is the time which the input of Circuit B requires to stabilize. Hence, theoretically, the input of Circuit B cannot be sampled (for a dependable precise value) for a time interval, t_s, following t_r; thus t_s represents time lost from the timing cycle and is just as deleterious as circuit delay.

b. Circuits such as blocking oscillators, clocks, etc., begin to regenerate when the signal voltages passes the voltage V_r, and require the signal voltage to be greater than a voltage $V_r < V_u$ over a specific time interval until the regeneration can be completed. In these cases the maximum undershoot which occurs must be controlled.

The effects of wiring configurations on signals were measured in the laboratory by simulating worst case 315 Data Processor wiring. The applied pulse was 2.6V in amplitude with rise and fall times of 20 to 30 millimicroseconds and a load consisted of a 10K ohm resistor to a -15V supply and 7μμfd to ground.

The results of the measurements are summarized in Table II.

TABLE II
Wire Measurements

Wire	loads	
	5	10
	(20% settling time/ft.)	
Single - loose bundled	23.5×10^{-9}	27×10^{-9}
Single - harnessed	-	25×10^{-9}
Twisted - loose bundled	17×10^{-9}	22×10^{-9}
Twisted - harnessed	20×10^{-9}	19×10^{-9}
RG-122A Unterminated	12×10^{-9}	15×10^{-9}

The wire configuration required to realize a given settling time could be determined if the load, maximum wire length, and required settling time (dictated by system timing) are known.

Logic Circuits

Logic Driver

The Logic Driver (Figure 12) is designed to drive five product loads at 20μμf each when driven by a unit logic current. The configuration is a single stage inverter, which exhibits a propagation time of less than 0.25μsec.

Power Driver

The Power Driver (Figure 13) is designed to drive a maximum of twenty product loads with a total load capacitance of 400μμf when driven by a unit logic current. The configuration employed provides high current at fast turn-on and turn-off times and minimum output noise. The required current gain is provided by Q_2, which is operated in the saturated mode to provide a low impedance path for noise. The required voltage gain is provided by Q_1, which is prevented from saturating to obtain high switch-off efficiency.

Alternate Circuits Considered

The required current could be obtained with a Darlington pair (Figure 14). However, the transistor of a Darlington pair is not operated in the saturated mode, and thus exhibits a lower value of the upper logical level, greater power dissipation, is susceptible to noise, and cannot supply load current on demand.

Two inverters can be cascaded to provide the required current gain; however, in such a configuration the total propagation time, which is the sum of the delays of each inverter, is excessive.

Double Driver

The Double Driver (Figure 15) is designed to provide two complementary outputs, each capable of driving eight product loads with a total load capacity of 96μμf, when driven by a unit logic current. The configuration employs a phase splitter amplifier which provides two opposite phased signals to drive two saturated current drivers. The configuration exhibits a propagation time of less than 0.1μsec.

Alternate Circuits Considered

Two complementary outputs can be obtained by cascading a power driver with a single driver stage (Figure 16). However, in such a configuration the two opposite phased signals differ in input to output delay by the delay of the "true" output stage. Furthermore, the overall delay is a sensitive function of the load on the false output, and the total propagation time is the sum of the delays in each driver stage.

In the diode coupled phase splitter (Figure 17) unequal base drives are provided to the output transistors and hence, considerable skew in the "true" and "false" output waveforms may result.

Flip-Flop

The flip-flop circuit, Figure 18, consists of a two saturated transistor inverters diode cross-coupled to form the basic memory element, two output RC coupled inverters, inductor clock pedestal gates, and a differential amplifier for the single-ended logic input.

The basic memory element formed by the diode coupled saturated inverters has the advantage that its repetition rate is not limited by reactive elements in the feedback loop. The RC coupled output drivers isolate the bistable from the logic d-c load, load switching and logic noise, and machine wiring capacitance. Hence, the load on the memory element is always constant and tolerance variations and clock power are therefore minimized. The RC coupled driver was chosen over the current mode inverter because of its higher switch off efficiency.

The non-saturated differential amplifier provides the local gain from the logic and provides the complementary logic levels for control of the inductor pedestal gate. The non-saturated differential amplifier isolates the clock gate from the logic by a high impedance, making the gate time constant independent of the state of the differential amplifier. The clock gate provides memory during the trigger interval by means of a current which is steered into the inductor associated with the bistable transistor which is to be turned off. This type of clock gate was chosen for its excellent signal to noise rejection ratio; i.e., triggering can not occur until the clock pulse reaches a level which will forward bias the gate isolation diode which requires approximately a 4 volt swing. Hence, noise must be approximately 4 volts to initiate false triggering. The gate does not require a strongly augmented clock pulse in terms of tightly controlled clock pulse parameters of width, shape, amplitude, rise, or fall time.

Constant current set-reset gates are provided independent of the logic and clock gate for certain control applications.

The circuit timing is illustrated in Figures 19 and 20. The pedestal set time is the interval during which the logic input is established and the clock gates are set. The bistable regeneration time is the interval during which the clock samples the memory element until the true and false logic output levels are established.

The logic input level when true turns Q_1 "on" and Q_4 "off". Initially all the current is directed thru R_4 to reference voltage $-V_2$.$^{(-4V)}$ The voltage drop across R_4 is controlled such that the potential $-V_b$ is not exceeded. Current into the clock diode D_6 then increases at the gate time constant determined by L_1 and R_4. Nearly all the current provided by $V_1$$^{(+15V)}$ and R_1 is transmitted through L_1 except for the current lost through R_4.

When the clock pulse appears, the level shift at node (b) forward biases D_9 and steers the inductor current into the flip-flop. If, during the clocking period, Q_1 remains "on", a constant current pulse is applied to the flip-flop. If Q_1 turns "off" during the clocking mode the current in L_1 decreases with the gate time constant. The gate time constant is designed to provide the necessary memory, against race and clock pulse skew.

If Q_4 turns on when the clock pulse is present all current through Q_4 passes into R_7 and to the reference voltage $-V_2$. The voltage at node (c) does not exceed $-V_b$, and therefore, diode D_{19} is reverse biased, and false triggering can not occur.

The basic delay times of concern here are the input delay t_1 and t_a, the transfer of current from Q_4 to Q_1, and the charge and discharge interval of the inductor L_1, t_{ch} and t_{dch}.

The positive clock pulse steers the current from inductor L_1 into the base of Q_2 turning Q_2 off. The current through R_{15} is steered to the base network of Q_3 turning it "on". Q_3 then supplies the current through the diode network to

R_{11} resulting in the reverse biasing of Q_2.

The collectors of Q_2 and Q_3 drive the RC coupled drivers. The clock must be present through the saturation and fall time of Q_2, the base delay of Q_3, clamp breaking, and rise time of Q_3.

The flip-flop design was optimized to minimize the regeneration interval and clock power. R_{12} is the control parameter for circuit optimization since it determines both the memory element load and the base drive available to the RC drivers.

The large current transferred from the gate through the isolation diode D_9 necessitates that D_9 be a low charge storage diode to prevent retriggering of F/F at the fall of the clock. D_9's recovery current is provided by the minority carrier charge storage of silicon diodes D_{10} and D_{12}, stay capacity, transistor capacities.

The functional specifications of the flip-flop are as follows:

Input:	Logic; standard 315 logic gate with a maximum fan-in from logic of 10.
Clock:	8 volt clock from -4 reference: Minimum clock pulse of 200 nanoseconds; Maximum clock skew of 50 nanoseconds.
Output:	Provides standard logic levels. Complementary, with each driving 12 standard gates under maximum load switching conditions, machine capacitance of 100 pf.
Delay:	Pedestal set delay 350 nanoseconds maximum, clock to logic output 200 nanoseconds maximum. Repetition rate 1 mc.

Clock Gate Systems

Evaluation

During the generation of the basic building blocks consideration was given to the compatibility of the gating and storage elements with the synchronized clock scheme. Before the clock system and triggering gate was chosen for the C-315 System an evaluation was made to determine the fundamental characteristics of a sychronized clock system. It should be pointed out that a computer system that requires a very precisely controlled clock system has historically been an engineering nightmare.

Clock Gate Characteristics

The evaluation resulted in consideration of the following six fundamental characteristics of a synchronized clock system: clock trigger gate, clock pulse shape, signal to noise rejection ratio, peak drive power, ratio of stand-by power to drive power (efficiencies), and distribution and routing.

The clock gate configuration and its impedance level are a foremost consideration since the gate determines the details of the clock drive circuitry and routing. The ideal clock system is one in which the six parameters are bounded well within the design capability and requirements of the particular system.

Clock Gate Types

Clock systems employ either energy storage or energy transfer gates. The energy storage gates are the classical capacitor store, and more recently the charge storage diode.(4) Energy transfer gates are the classical d-c level shift gates of the threshold and pedestal type, and more recently the diode current steering gates of current mode logic.(12)

The energy transfer gate is susceptible to logic race and clock pulse skewing since it does not provide memory during the trigger interval. However, since there is no memory element to charge-up or discharge, the gate speed can be quite high. Reliable operation can only be assumed by providing sufficient delays in the logic to prevent ambiguous gate conditions during triggering.

The energy storage gate does have memory during the trigger interval, and can be made insensitive to the changing state of the gate driver and clock pulse skewing. However, sufficient time must be allowed between clocks (or from logic time) to charge-up and discharge the memory element.

Theoretically, a delay line can be used to eliminate race problems, either as a charged element in the clock gate or as a logic delay element in the logic. The necessity of providing proper source and load terminations to the delay line usually limits its application to the clock gate.

Implementation of a system using energy transfer gates requires that the logic skews and minimum and maximum delays be precisely defined. Since delays are difficult to predict during the early design phases of the system, over design is employed with the resulting loss of desired speed. In general, the energy storage gate charge and discharge intervals are well defined, thereby minimizing over design and providing a more reliable system, even if somewhat slower than that mechanized with an enery transfer gate.

The choice of gate and clock polarity depends upon the trigger mode; i.e., whether the triggered element is to be turned on or off. The choice of turn-off requires more clock power, however, it establishes better control over the flip-flop regeneration. With the clock gate providing the reverse turn-off charge, the memory element gain is maximized since the gain is based entirely on "turn-on" conditions and d-c bias.

The nominal speed, operating mode, reliability requirements, cost range of the C-315 System, and the desire not to double-end all logic gate and circuit delays resulted in the selection of the energy storage gate. The pedestal gate energy technique employing the capacitor and inductor were very closely compared.

Capacitor Gate

The Capacitor Gate (Figure 21a) does not require d-c current drive which, therefore, eliminates switching large currents through clock line inductances. The initial rise of the clock pulse initiates the triggering process. The clock voltage swing is limited to the minimum logic swing to prevent false triggering. The clock isolation diode must be back biased to prevent noise on the clock line from triggering the memory element, which forces a lower clock voltage (more susceptible to noise), a larger logic swing (more logic delay), or re-referencing of the bistable to provide adequat e noise rejection. Since the clock line is coupled directly into the base of the triggered element, once the reverse bias of D_1 (usually 1 to 1.5 volts) is exceeded, ringing and noise pick-up on the clock line must be eliminated by careful routing and termination techniques. The capacitor gate requires a large peak drive current from the clock, and a very fast rise time to establish a reasonable gate efficiency.

The charge required from the capacitor gate during the clock pulse must satisfy the following requirements:

$$Q_r \geq Q_{off} + I_{dc}T_S + I_{dis}\, T_r \quad \text{max} \qquad (17)$$

Q_{off} = Charge required to turn the "on" transistor off and charge any stray or transition capacities in the flip-flop base network.

I_{dc} = Total d-c current which maintains the bistable transistor "on".

I_{dis} = Average discharge current in the trigger network.

T_r = Rise time of the clock pulse.

T_s = Flip-flop regeneration time.

Since the charge required must be delivered by the pedestal capacitor C_p when the clock pulse V_c ($V_c < \Delta V$ logic swing) is a minimum.

$$Q_r = (C_pV_c)_{min} \qquad (18)$$

Hence, the pedestal capacitor must satisfy the requirement:

$$C_{p\ min} > \frac{Q_{off} + T_SI_{dc} + T_rI_{dis} \quad \text{max}}{V_{c\ min}} \qquad (19)$$

The current I_L required to charge the pedestal capacitor is approximately (assuming current source steps and equal charge and discharge times):

$$I_L \geq \frac{2Q_r}{T_c} = \frac{2\left[\frac{Q_{off}}{T_c} + I_{dc}\frac{T_s}{T_c}\right]}{\left[1 - \frac{T_r}{T_c}\right]} \qquad (20)$$

Where T_c is the time interval between the time at which the logic becomes quiescent to the initiation of the clock pulse.

The peak current I_p required from the clock is approximately:

$$I_p = Q_{off} + I_{dc}T_s \cdot \frac{T_c}{T_r\,(T_c - T_r)} \qquad (21)$$

Thus, the current which must be provided by the logic (or buffering amplifier) and the peak current provided by the clock are very sensitive functions of the system frequency;

$$\frac{1}{T_r + T_s + T_c + T_L}$$

where T_L is the logic time and, may become prohibitively large as T_c is reduced. Hence, the pedestal capacitor finds its principle application in low frequency computers (<200 KC) where considerable advantage can be taken of the large available T_c to keep I_L small, such that the current required to charge the pedestal capacitor can be provided by the logic.

Inductor Gate

The Inductor Gate has an excellent signal to noise rejection ratio, since the noise must be nearly as large as the clock reference voltage before triggering is initiated; however, triggering is not initiated until the clock pulse reaches $-V_b$. The Inductor Gate results in a large d-c current in the clock line during standby and necessitates a clock voltage swing larger than the reference voltage in order to discharge the clock line inductance. The shape of the clock pulse in terms of pulse amplitude, pulse width, or fall time are not critical, and the clock reference is limited only by the power limitations of the gate driver.

The charge which must be delivered by the inductor during the clock pulse interval is given by:

$$Q_{reqd} = Q_{off} + I_{dc}T_s \qquad (22)$$

Hence, to a first order approximation, the inductor pedestal time constant, τ_L, and quiescent current, I_L, must satisfy the relation:

$$I_LT_s\left[1 - \frac{T_r}{\tau_L}\right]\left[1 - \frac{T_s}{2\tau_L}\right]_{min} \geq Q_{off} + I_{dc}T_s \quad \text{max} \qquad (23)$$

Since the clock gate must be designed to discharge the inductor in the interval T_c, which for eighty-six percent discharge requires

$$\tau_L < \frac{T_c}{2},$$

the minimum quiescent inductor current is:

$$I_{L\ min} > \frac{\frac{Q_{off}}{T_s} + I_{dc}}{\left[1 - \frac{2\,T_r}{T_c}\right]\left[1 - \frac{T_s}{T_c}\right]_{max}} \tag{24}$$

The maximum current required from the clock is equal to the quiescent inductor current.

The ratio of the current which must be provided to charge the capacitor and inductor pedestal gates is:

$$\frac{I_{L\ cap}}{I_{L\ ind}} = 2\ \frac{\frac{Q_{off}}{T_c} + I_{dc}\frac{T_s}{T_c}}{\frac{Q_{off}}{T_s} + I_{dc}} \cdot \left[1 - \frac{T_s}{T_c}\right] \tag{25}$$

where $T_r << T_c$

Multiplying numerator and denominator by $\frac{T_s}{T_c}$ and cancelling we obtain:

$$\frac{I_{L\ cap}}{I_{L\ ind}} = \frac{2T_s}{T_c}\ (1 - \frac{T_s}{T_c}) \tag{26}$$

The ratio $\frac{I_{L\ cap}}{I_{L\ ind}}$ is bounded between 0 and 1/2 for $\frac{T_s}{T_c}$ between 0 and 1; the optimum condition for the inductor gate obtains when $\frac{T_s}{T_c} = 1/2$. Thus, the flip-flop gain, which can be achieved with a capacitor gate, will always be at least twice as great for a given $\frac{T_s}{T_c}$ as that which can be achieved with an inductor gate. The principle justification for employing the inductor gate is its greater noise rejection and the lower peak current required from the clock.

C-315 Clock Gate

The C-315 Clock Gate is a pedestal gate of the inductor energy storage type (Figure 21). The ratio of $\frac{T_s}{T_c}$ for the C-315 System is close to the optimum and is 0.57. The gate operation is described in the discussion on the flip-flop circuit and is characterized with design conditions and equations in Appendix IV.

The clock system in Figure 22 is mechanized with a modified Butler Oscillator-Shaper from which control decision logic generates the time initiated (t_i). The basic control time intervals are generated from multi-tap delay lines. The clock pulses, before entering the logic, are reshaped by the low power clock shaper amplifier to minimize pulse stretching and maintain a uniform clock pulse entering the transmission lines. The specifications of the main clock, main clock amplifier, low power clock, and auxiliary clock amplifier are characterized for their functional parameters in Figure 22. The auxiliary clock amplifier drives the logic at logic time (t_L). It is necessary to control the clock line length and circuit skew such that the design of the pedestal gate will be adequate under the worst case race conditions.

The race is most serious for the condition where the earliest clock triggers the fastest flip-flop which then determines the logic input to the slowest flip-slop accepting the latest clock. The maximum clock voltage skew of the auxiliary clock is 20 musec, with a maximum of 50 musec rise time. The maximum clock current skew is 50 musec, with a worst case rise time of 30 musec.

The wire used for the clock distribution is thin-wall Teflon, twisted pair, 26 gage, unterminated. The ringing, due to the short line length of unterminated twisted pairs, is above the upper trigger level and does not interfere with the clocking. Until the clock diodes recover, the clock line is low impedance and is terminated by the clock gates. The discharging of the distribution capacitance and inductance over the entire clock voltage swing minimizes the effect of ringing.

Appendix I

Efficiency of Logic Gate

Gate efficiency is defined as the ratio of the sum of the on and off currents to the required gate drive current of the building block.

$$E = \frac{I_{b1} + I_{b2}}{I_{DL}}$$

The efficiency relates the building block delay, as defined by I_{b1} and I_{b2}, to that required per unit gate drive to obtain the specific delay. The reverse current, I_{b2}, is obtained from the sum resistor in the false logical level. The on drive current, I_{b1}, is derived from the general gate configuration shown in Figure I-1 along with the gate drive current, I_{DL}.

$$I_{b1} = I_{RP} + p\ I_{PL} - (I_{b2} + I_L) - (m-1)\ I_{SL}$$

$$I_{DL} = \ell\ (I_{RP} + I_{SL}) + \ell (p-1)\ I_{PL}$$

Equation I_{DL} can be written in terms of the logic gate currents and reference voltages for the general gate configuration.

$$I_{DL} = \frac{(V_2 - V_D)\left[I_{b1} + I_{SLT} + \left(\frac{V_1 + V_{eb}}{V_1}\right) I_{b2}\right]}{V_2 - (V_{eb} + V_B)}$$

With substitution of the limit coefficients for the design tolerances on components and voltages, the drive current I_{DL} can be calculated vs. reference voltage for different values of desired building block drive currents I_{b1} and I_{b2}. Figure I-2 shows the significant improvement in the necessary drive as the limit coefficient in the product resistor decreased from 5% to 1% and the reference voltage increased from 10 to 15V. The evaluation assumed equal on and off drive currents.

The reference voltage of 15 volts was chosen with the logic resistor of 10K ± 1%. The worst case and best case efficiency for equal on and off drive were calculated with the result that $50\% \leq E \leq 80\%$.

Appendix II

Capacitance of Logic Gate

Since the logic gates may be physically located on a separate card from the drive circuits, power transfer from a logic driver to a logic driver includes the following capacitances:

a. flip-flop card capacitance,

b. flip-flop connector capacitance (output),

c. wire capacitance,

d. gate connector capacitance (input),

e. gate card capacitance,

f. diode capacitance,

g. gate connector capacitance (output),

h. wire capacitance,

i. driver connector capacitance (input),

j. driver card capacitance,

k. transistor input capacitance.

Card and connector capacitances were measured on a capacitance bridge. Diode capacitance was measured with a diode test set. Wire capacitance was calculated assuming the wire to be parallel to itself in 75% of the machine, and parallel to a ground plane at one diameter for the remaining 25%. Wire capacitance calculations were performed with wire insulations of teflon, PVC, and FEP-100 with the manufacturer's specifications. The following design values of capacitance resulted.

H24000-01 (sum diode)	1.0 μμf
H24001-01 (product diode)	1.5 μμf
wire-wrap connector (double wrap)	3.0 μμf
wire capacitance (teflon)	12.8 μμf/ft
wire capacitance (PVC)	26.0 μμf/ft
wire capacitance (FEP -100)	15.7 μμf/ft
card capacitance	1.0 uuf/ inch[1] average

NOTE[1] 0.032 inch conductor width with 0.032 spacing (5 to 6 uuf per card).

The input and output capacities of the various logic circuits are computed as follows:

A. Input capacitance

$$C_{1n} = n \times C_T^P + k\ (C_c + C_k) + L\ C_w + \left[(m-1) + (b-1)\right]\ C_T^s$$

n = number of product diodes

k = number of wire wrap connectors

L = length of wire in feet

m = number of sum diodes

b = number of sums

C_T^P = average maximum junction capacitance of product diode

C_T^s = average maximum junction capacitance of sum diode

C_c = average card capacitance from printed wire

C_k = average terminal capacitance

C_w = wire capacitance (a function of wire insulation)

B. Output Capacitance

$$C_{out} = k\ (C_c + C_k) + L\ C_W + \ell\ (n - 1) C_T^P + C_T^c$$

definitions are the same as for input capacitance

C_T^c = average maximum junction capacitance of clamp diode

ℓ = number of gates circuit is driving

The following desing values were determined for the input and output of the various building blocks.

Circuit	Input from Gate	Output to Gate
Flip-Flop	50μμf	100μμf
Adder Double Driver	30μμf	100μμf
Logic Double Driver	50μμf	100μμf
Power Driver	30μμf	400μμf

Appendix III

Design Equations and Procedure for the 315 Inverter Circuit.

The following includes the design procedure and equations used in the design of the diode coupled inverter circuit shown in Figure III-1 and the timing diagram, Figure III-2. The inverter circuit is shown here since it shows the essential equations and approach used on all the basic logic circuits.

TURN ON CONDITIONS

I. $\left(t_R + t_{fb} + t_{CL}\right)_{max} \leq T_1 - T_r$

II. $t_{r\,max} \leq T_r$

TURN OFF CONDITIONS

III. $\left(t_{rb} + t_F\right)_{max} \leq T_k - T_f$

IV. $\left(t_{RC}\right)_{max} \leq T_f$

DC OFF CONDITION

V. $I_{R1\ off\ min} \geq I_{co\ max} + I_{LD3\ max} + I_{Bias\ max}$

CLAMP CONDITION

VI. $V_{c\ off\ max} \leq V_{L\ min}$

DC ON CONDITION

VII $\dfrac{I_{c\ max}}{I_{B\ on\ min}} \leq \beta_{min}$

LOAD TRANSIENT CONDITION

VIII. $\dfrac{I_{load\ max}}{I_{B\ on\ min}} \leq \beta_{S\ min}$

Formulation of Circuit Equations

To save space we will write the Circuit Equations with the Worst Case Design Philosophy already incorporated.

$$\text{I.}\quad t_{fb\ max} = \frac{\left[m C_T^p + (n-1)\, C_T^s + C_l + C_{Sl}\right] (V_{b\ off} - V_{b\ on})_{max}}{I_{B\ on\ min}}$$

$$t_{CL\ max} = \frac{\bar{\beta}_{min}}{\bar{\omega}_{T\ min}}\ \mathrm{Ln}\left\{\frac{1}{1 - \dfrac{I_{CL\ max}}{\bar{\beta}_{min}\, I_{B\ on\ min}}}\right\}$$

$$t_{R\ max} = \left[\frac{1}{\bar{\omega}_{T\ min}} + R_{L\ min}\left(\bar{C}_{C\ max} + \frac{C_{L\ max}}{\bar{\beta}_{min}}\right]\right.$$

$$\bar{\beta}_{min}\ \mathrm{Ln}\left\{\frac{1 - \dfrac{I_{CL\ max}}{\bar{\beta}_{min}\, I_{B\ on\ min}}}{1 - \dfrac{I_{C\ max}}{\bar{\beta}_{min}\, I_{B\ on\ min}}}\right\}$$

$$I_{B\ on\ min} = \frac{V_{b\ on} - V_{DS} - V_B - V_{2\ max}}{R_{P\ max}} - \frac{V_{1\ max} - V_{b\ on}}{R_{S\ min}} - (n-1)\ I_{SL}$$

$$I_{CL\ max} = \frac{V_{CL\ max} - V_{D4} - V_{2\ min}}{R_{2\ min}} + \frac{V_{CL\ max} - V_{D4} - V_{DP} - V_{2\ min}}{R_{P\ min}/\ell}$$

$$\text{II.}\quad \mathrm{erf}\sqrt{\frac{t_{r\ max}}{\tau_P}} = \frac{1}{1 + \dfrac{I_{r\ min}}{I_{f\ max}}}$$

$$I_{r\ min} = \frac{I_{R_{P\ min}} - I_{R_{S\ max}}}{n-1}$$

$$I_{f\ max} = \frac{I_{R_{S\ max}} - I_{co\ min}}{n} \quad \text{where } I_{R_{S\ max}} = \frac{V_{1\ max}}{R_{S\ min}}$$

$$\text{and } I_{RP\ min} = \frac{-V_{B\ min} - V_{DS\ max} - V_{2\ max}}{R_{P\ max}}$$

III. $$t_{rb\,\text{max}} = \left[\frac{nC_T^s + C_{Sl} + C_l + C_B}{I_{B\,\text{off min}}}\right](V_{b\,\text{off}} - V_{b\,\text{on}})_{\text{max}}$$

$$t_{F\,\text{max}} = \frac{\left\{\frac{I_{c\,\text{max}}}{\bar{\omega}_{T\,\text{min}}} + K_{s\,\text{max}}\left[I_{b\,\text{on max}} - \frac{I_{c\,\text{max}}}{\beta_{\text{max}}}\right]\right\}\left(1 - \frac{t_F + t_{rb}}{2\,\bar{\tau}_p}\right) + \bar{C}_c\,\Delta V_{c\,\text{max}}}{I_{b\,\text{off min}}}$$

$$I_{C\,\text{max}} = \frac{-V_{2\,\text{min}}}{R_{2\,\text{min}}} - \ell\,\frac{V_{2\,\text{min}} - V_{DP}}{R_{P\,\text{min}}}$$

$$I_{B\,\text{on max}} = \frac{V_{b\,\text{on}} - V_{DS} - V_B - V_{2\,\text{min}}}{R_{P\,\text{min}}} - \frac{V_{1\,\text{min}} - V_{b\,\text{on}}}{R_{S\,\text{max}}}$$

$$I_{B\,\text{off min}} = \frac{V_{1\,\text{min}} - V_{b\,\text{off max}}}{R_{S\,\text{max}}}$$

IV. $$t_{RC\,\text{max}} = \frac{C_{\text{max}}\;\Delta V_{c\,\text{max}}}{I_{C\,\text{min}}} = \frac{\left[C_{SO} + \bar{C}_{c\,\text{max}} + \ell\,C_T^p + C_{D4}\right]\;\Delta V_{c\,\text{max}}}{\frac{-V_{2\,\text{max}} - \frac{\Delta V_{\text{max}}}{2}}{R_{2\,\text{max}}} - I_{coQ1} - \ell\,I_{pL}}$$

V. $$I_{R1\,\text{off min}} = \frac{V_{1\,\text{min}} - V_{b\,\text{off max}}}{R_{S\,\text{max}}}$$

VI. $$V_{c\,\text{off max}} = (\ell\,I_{pL} + I_{co\,\text{max}})\,R_{2\,\text{max}} + V_{2\,\text{max}}$$

VII. $I_{B\,\text{on min}}$ defined under I

$I_{c\,\text{max}}$ defined under III

VIII. $$I_{\text{load max}} = \ell\,\frac{-V_{2\,\text{min}} - V_{DP}}{R_{P\,\text{min}}}$$

Formulation of Design Conditions

I. $$\left[\frac{1}{\bar{\omega}_{T\,\text{min}}} + R_{L\,\text{min}}\left(\bar{C}_{c\,\text{max}} + \frac{C_{L\,\text{max}}}{\bar{\beta}_{\text{min}}}\right)\right]\bar{\beta}_{\text{min}}\,\cdot$$

$$\text{Ln}\left\{\frac{1 - \frac{I_{CL\,\text{max}}}{\bar{\beta}_{\text{min}}\,I_{B\,\text{on min}}}}{1 - \frac{I_{C\,\text{max}}}{\bar{\beta}_{\text{min}}\,I_{B\,\text{on min}}}}\right\}$$

$$+\frac{\left[mC_T^p + (n-1)\,C_T^s + C_i + C_{Si}\right](V_{b\,\text{off}} - V_{b\,\text{on}})_{\text{max}}}{I_{B\,\text{on min}}}$$

$$+\frac{\bar{\beta}_{\text{min}}}{\bar{\omega}_{T\,\text{min}}}\,\text{Ln}\left\{\frac{1}{1 - \frac{I_{CL\,\text{max}}}{\bar{\beta}_{\text{min}}\,I_{B\,\text{on min}}}}\right\} \leq T_l - T_r$$

II. let $$\text{erf}\sqrt{\frac{T_r}{\tau_p}} = g$$

$$g \geq \frac{1}{1 + \frac{I_{r\,\text{min}}}{I_{f\,\text{max}}}}$$

$$\frac{I_{r\,\text{min}}}{I_{f\,\text{max}}} \geq \frac{1-g}{g}$$

$$\frac{n}{(n-1)}\;\frac{I_{R_{P\,\text{min}}} - I_{R_{S\,\text{max}}}}{\frac{V_{1\,\text{max}}}{R_{S\,\text{min}}} - I_{coQ1\,\text{min}}} \geq \frac{1-g}{g}$$

III. $$\frac{\left[n\,C_T^s + C_{Sl} + C_l + C_B\right](V_{b\,\text{off}} - V_{b\,\text{on}})_{\text{max}}}{I_{B\,\text{off min}}}$$

$$+\frac{\left(1 - \frac{t_F + t_{rb}}{2\,\bar{\tau}_p}\right)\left\{\frac{I_{c\,\text{max}}}{\bar{\omega}_{T\,\text{min}}} + K_{S\,\text{max}}\left[I_{b\,\text{on max}} - \frac{I_{C\,\text{max}}}{\bar{\beta}_{\text{max}}}\right]\right\} + \bar{C}_{c\,\text{max}}\,\Delta V_{c\,\text{max}}}{I_{B\,\text{off min}}}$$

$$\leq T_k - T_f$$

IV. $$\left\{\frac{\left[C_{SO} + \bar{C}_{C\,\text{max}} + \ell\,C_T^p + C_{D4}\right]\;\Delta V_{c\,\text{max}}}{\frac{-V_{2\,\text{max}} - \frac{\Delta V_{\text{max}}}{2}}{R_{2\,\text{max}}} - I_{co\,Q_1} - \ell\,I_{PL}}\right\} \leq T_f$$

V. $$\frac{V_{1\,\text{min}} - V_{b\,\text{off max}}}{R_{S\,\text{max}}} \geq I_{CO\,\text{max}} + I_{LD3\,\text{max}} + I_{\text{BIAS}}$$

VI. $$(\ell\,I_{PL} + I_{CO\,\text{max}})\,R_{2\,\text{max}} + V_{2\,\text{max}} \leq V_{L\,\text{min}}$$

VII.

$$\frac{-\frac{V_{2\,\text{min}}}{R_{2\,\text{min}}} - \ell\,\frac{V_{2\,\text{min}} - V_{DP}}{R_{P\,\text{min}}}}{\frac{V_{b\,\text{on}} - V_{DS} - V_B - V_{2\,\text{max}}}{R_{P\,\text{max}}} - \frac{V_{1\,\text{max}} - V_{b\,\text{on}}}{R_{S\,\text{min}}} - (n-1)\,I_{SL}} \leq \beta_{\text{min}}$$

VIII.

$$\frac{-\ell\,\frac{V_{2\,\text{min}} - V_{DP}}{R_{2\,\text{min}}}}{\frac{V_{b\,\text{on}} - V_{DS} - V_B - V_{2\,\text{max}}}{R_{P\,\text{max}}} - \frac{V_{1\,\text{max}} - V_{b\,\text{on}}}{R_{S\,\text{min}}} - (n-1)\,I_{SL}} \leq \beta_{s\,\text{min}}$$

DEFINITIONS OF TERMS

$\bar{\beta}$	β averaged over the current swing at V_{CL}
β_s	on demand current β
C_B	effective transition capacity of series combination of D_1 and D_2.
$\bar{C}_c$	collector capacity averaged over the collector voltage swing.
C_i	input capacity of base of Q_1
C_L	load capacity
C_{Si}	stray capacity at input
C_{SO}	stray capacity at output
C^P_T	transition capacity of product diode
C^s_T	transition capacity of sum diode
I_{BIAS}	current supplied to logic by R_1 when Q_1 is OFF.
$I_{B\ on}$	Q_1 base current when ON
$I_{CL\ MAX}$	maximum dynamic clamp current
I_{CO}	Q_1 collector saturation current
I_{LD}	saturation current of D_3 when reverse biased
I_{PL}	saturation current of reverse biased product diode
I_{SL}	saturation current of reverse biased sum diode
I_f	forward current per sum diode before input falls
I_r	reverse current per sum diode during recovery
I_{RS}	current through R_S when the product goes true
I_{R1}	current through R_1
ℓ	number of product diodes driven from inverter
m	number of products per sum term
n	number of sum terms
R_L	equivalent load resistance
t_{CL}	time for collector current to break the clamp (D_4)
t_F	time to neutralize the carriers in the base of Q_1
T_f	specified time for load capacities to be discharged
t_{fb}	time for input voltage of Q_1 to discharge to the point where Q_1 begins to turn ON
T_k	specification on the total time for Q_1 to turn-off measured from beginning of rise of input voltage
T_1	specification on the total time for Q_1 to saturate measured from beginning of fall of input voltage.
τ_p	minority carrier lifetime of sum diodes
$\bar{\tau}_p$	average minority carrier lifetime of Q_1 during turn-OFF
t_r	recovery time of (n-1) sum diodes
t_R	time for collector to rise to saturation voltage
T_r	specified recovery time of (n-1) sum diodes
t_{RC}	time for the collector to fall to the clamp level.
$V_{b\ OFF}$	steady state base voltage when Q_1 is OFF.
$V_{b\ ON}$	steady state base voltage when Q_1 is ON.
V_L	logic level at input to on driver stange
$\bar{\omega}_T$	ω_T averaged over the current swing at V_{CL}.

Appendix IV

Inductor Clock Pedestal Gate. Design conditions and equations.

Figure 19 shows the pedestal set interval. Figure 22 shows the gate current and trigger current and the significant intervals that provide for logic race and clock skew. The gate is characterized in three operating intervals; the initial gate set interval, the reset interval, and the discharge interval during clock operation of the gate. The gate should be optimized to minimize the set and reset interval.

There are eight design constraints on the gate properties.

Set Interval

I. $$t_{ch\ max} \leq T_{TC} - (t_1 + t_{a\ on} + t_r)\ max$$

II. $$I_L \geq I_{L\ min}$$

Reset Interval

III. $$t_{dch} \leq T_{TD} - (t_1 + t_{a\ off} + t_f)\ max$$

IV. $$I_L \leq I_{LS\ max}$$

V. $$I_{LS\ max} \leq \frac{Q_s''}{t_{off\ max}}$$

Clock Trigger Interval

VI. $$T_{s\ max} \leq T_{cp}$$

VII. $$Q^1_{s\ min} \geq Q_{s\ max}\left(1 - \frac{t_{off\ max}}{2\tau_{p\ max}}\right) + \bar{C}_{cb}\,\Delta V_{cb}$$

VIII. $I_{L\ min}$ at $T_{s\ max} \geq I_{LX\ max}$

Design Equations

Set interval

$$t_{ch\ max} = \tau_{L\ max} \ln \frac{I_{L\ max}}{I_{L\ min}} \qquad \text{I}$$

$I_{L\ min}$ = minimum current in L_1 for trig-condition

$I_{L\ max}$ = maximum current in L_1 from component limits

$\tau_{L\ max} = \frac{L_{1\ max}}{R_{4\ min}}$ gate time constant

Reset interval

$$t_{dch} = \tau_{L\ max} \ln \frac{I_{L\ max}}{I_{LS\ max}} \qquad \text{II}$$

$I_{LS\ max}$ = maximum amount of current in L_1 and false triggering prevented

$$I_{LS\ max} = \leq \frac{Q''_s}{t_{off\ max}} \qquad \text{III}$$

Where Q''_s is defined as the minimum disturb charge required to move the least saturated transistor to the edge of saturation.

Discharged during clocking

$I_{LX\ max}$ = is the maximum d-c current which must be provided during $(T_s - t_{off})$ to bias the flip-flop off.

$$I_{LX\ max} = I_{b1\ max} + I_{co\ max} + I_{bias} \qquad \text{IV}$$

where: I_{bias} = additional current necessary to raise $V_{b\ on\ max}$ to $V_{b\ off\ min}$

$I_{LX\ max}$ must exist for the full interval of T_s and is related to the gate time constant by the followfollowing:

$$I_{LX\ max} \leq I_{LT\ min}\, e^{-T_s/\tau_{L\ min}} \qquad \text{V}$$

Figure shows the charge diagram for the gate where the effects of clock voltage and current rise time and skew are considered. During clock voltage rise time some charge is lost and during clock current rise time additional current is lost.

$$I_{LT\ min} \leq I_{LO\ min} - \frac{V_{c\ max}}{R_{4\ max}} \frac{\tau_{L\ min}}{T_{RV\ max}} \left(e^{\frac{-T_{RV\ max}}{\tau_{L\ min}}} + \frac{T_{RV\ max}}{\tau_{L\ min}} - 1\right) \qquad \text{VI}$$

$I_{LO\ min}$ is the minimum current in the inductor at the start of the clock pulse rise time.

where:

$$I_{LO\ min} = I_{CQ_1\ min} - \frac{V_{D6\ max}}{R4\ min} \qquad \text{VII}$$

The transfer of charge into the bistable is the following:

$$Q^1_S \max \leq \tau_{L\ min}\, I_{LT\ min} \left(1 - e^{\frac{-t_{off}}{\tau_L}}\right) - I_{LX\ max}(t_{off} - T_{RI}) - \tau_{L\ min} I_{LT\ min} \left(1 - e^{\frac{-T_{RI}}{\tau_{L\ min}}}\right) + 1/2\, T_{R1} \left(I_{LT\ min}\, e^{\frac{-T_{RI}}{\tau_{L\ min}}} - I_{LX\ max}\right) \qquad \text{VIII}$$

$$\text{where: } Q_{S'\ max} = Q_S \left(1 - \frac{t_{off}}{2\tau_p}\right) + \overline{C}_{cb}\, \Delta V_{cb} \qquad \text{IX}$$

T_{TI} = clock current rise time.

Since Q^1_S, t_{ch}, and t_{dch} are functions of τ_L and $I_{L\ max}$ the gate is optimized by satisfying in the limit equations V simultaneously with VIII.

Acknowledgement

The 315 program was the result of a converted effort by many individuals; particular recognition is due to E. Kline and R. Gunderson for making the original decision to use CMDL, clock gate and the flip-flop configuration, and constructing the engineering model, to R. Baker for his effort in obtaining the information required to formulate the building block specifications, to P. Glaser and C. Fujimoto for originating and implementing the packaging concepts, D. Yamasaki for his work in the design of the clock amplifier and distribution system, J. Fort for the evaluation and specifications of the components. Particular recognition is to S. Mitchell for editing and preparing the manuscript.

References

1. S. Bloom, I. Pardo, W. Keating, E. Mayne. "Card Random Access Memory (CRAM): Functions and Use", Proceedings, Eastern Joint Computer Conference, 1961.

2. A. I. Pressman, "Design of Transistorized Circuits for Digital Computers", Rider, 1959, Chapter 7.

3. J. M. Mitchell, S. Ruhman, "The TRICE: A High Speed Incremental Computer", IRE National Convention Record, 1958, Part 4.

4. L. P. Retzinger, "High Speed Circuit Tech-

niques Utilizing Minority Carrier Storage to Enhance Transient Response", Proceedings, Western Joint Computer Conference, May, 1958.

5. G. H. Goldstick, "Characterization of Semiconductor Diodes for Switching Circuit Design", AIEE Winter Meeting, Fegruary, 1962.

6. E. F. Allgeyer, R. L. Dougherty, "Noise in DCTL Circuitry", TRADIC Computer Research Program Summary Engineering Report, Supplement 1, June,1957.

7. G. H. Goldstick, "Comparison of Saturated and Non-Saturated Switching Circuit Techniques", PGEC Transactions, June, 1960.

8. J. J. Sparkes, "A Study of Charge Control Parameters of Transistors", Proceedings of the IRE, October, 1960.

9. R. Beaufoy, "Transistor Switching Circuit Design Using Charge Control Parameters", ATE Journal, October, 1960.

10. J. J. Sparkes, R. Beaufoy, "The Junction Transistor as a Charge Controlled Device", ATE Journal, October, 1957, Volume 13, pp 310-327.

11. G. H. Goldstick, D. G. Mackie, "Design of Computer Circuits Using Linear Programming Techniques", IRE Convention Record, March, 1961.

12. J. Millman, H. Taub, "Pulse and Digital Circuits", McGraw-Hill Book Co., 1956, Chapter 14.

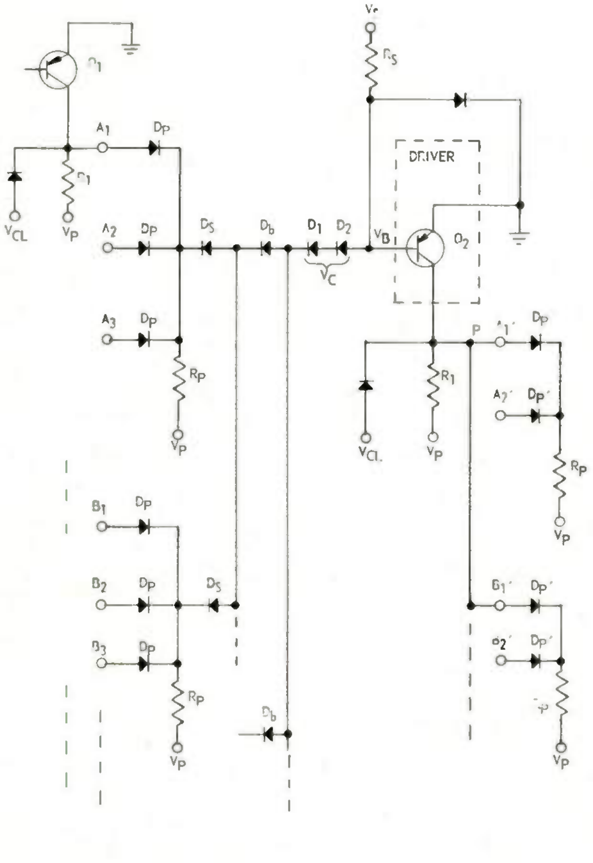


Fig. 2. Basic current mode logic block schematic diagram.

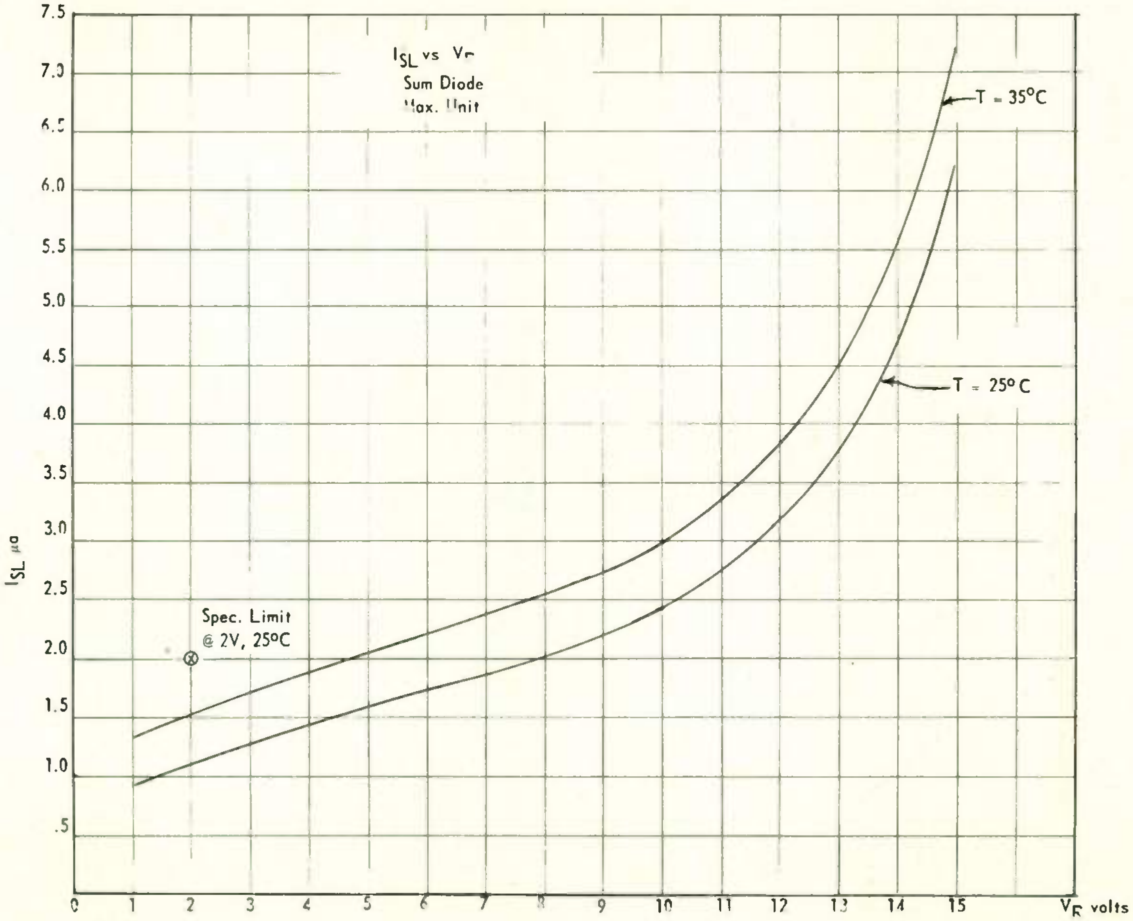


Fig. 3. Plot of sum diode leakage versus reverse voltage.

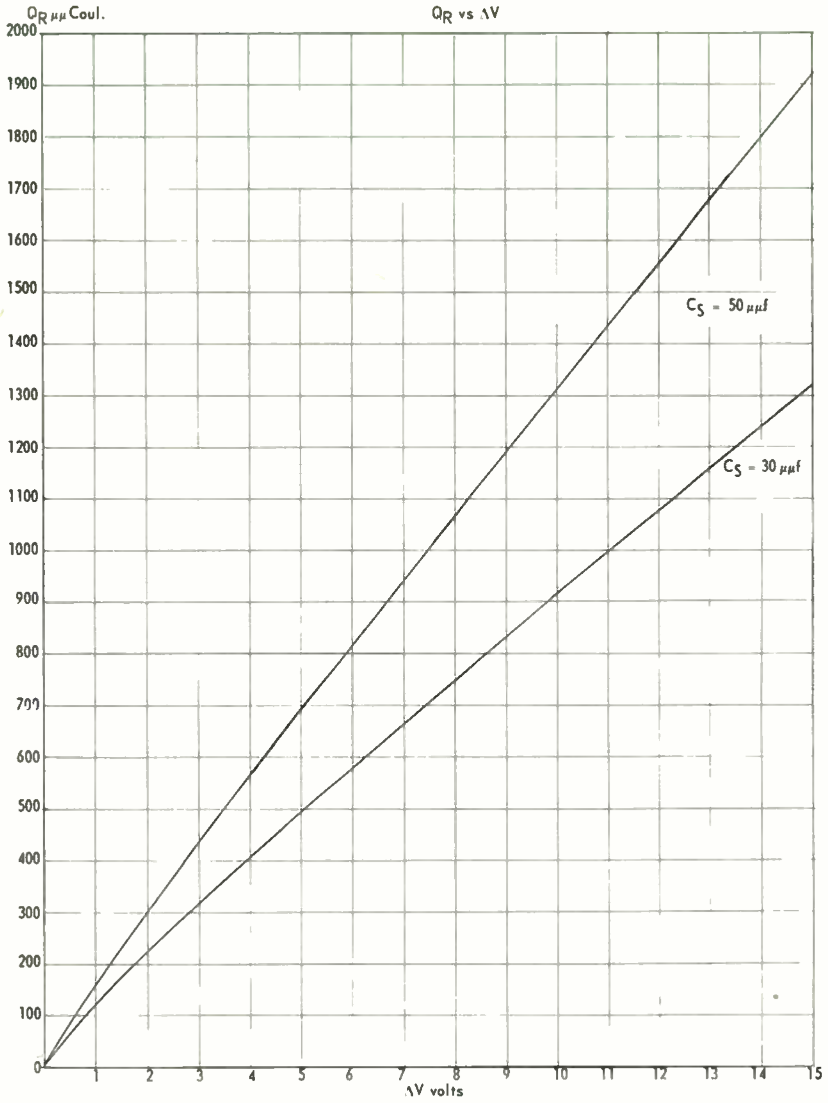


Fig. 4. Plot of Q_C versus ΔV.

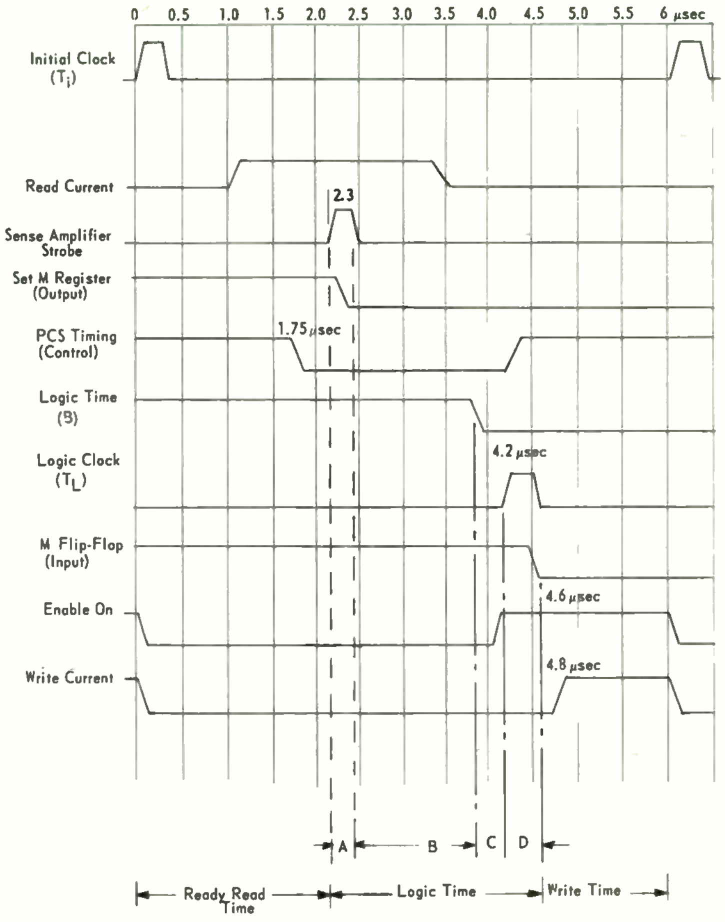


Fig. 5. 315 System timing diagram.

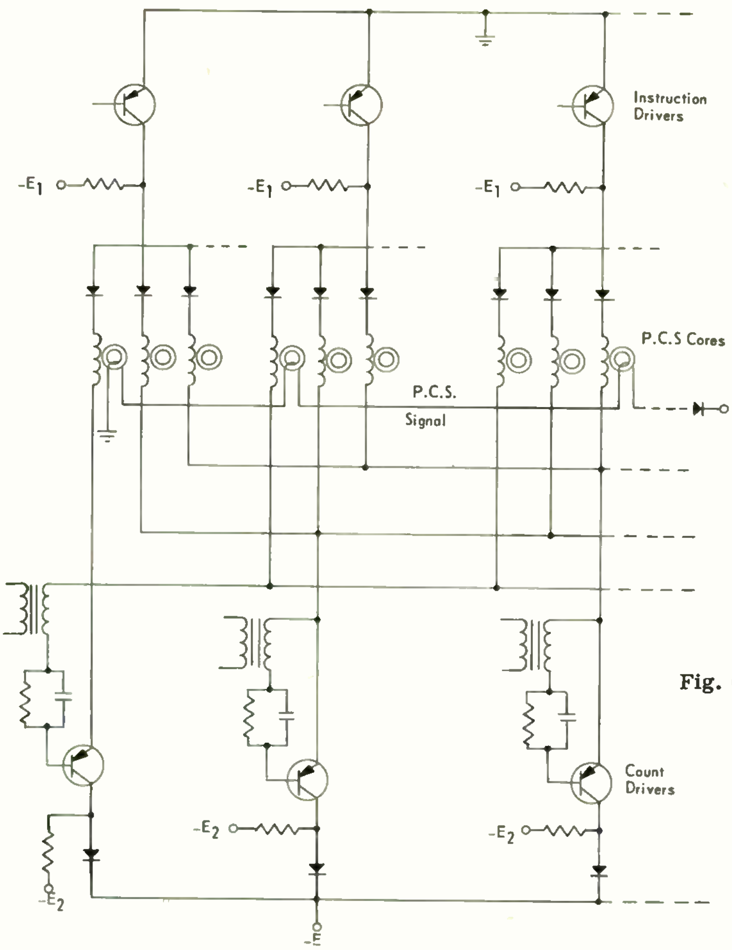


Fig. 6. P.C.S. mechanization.

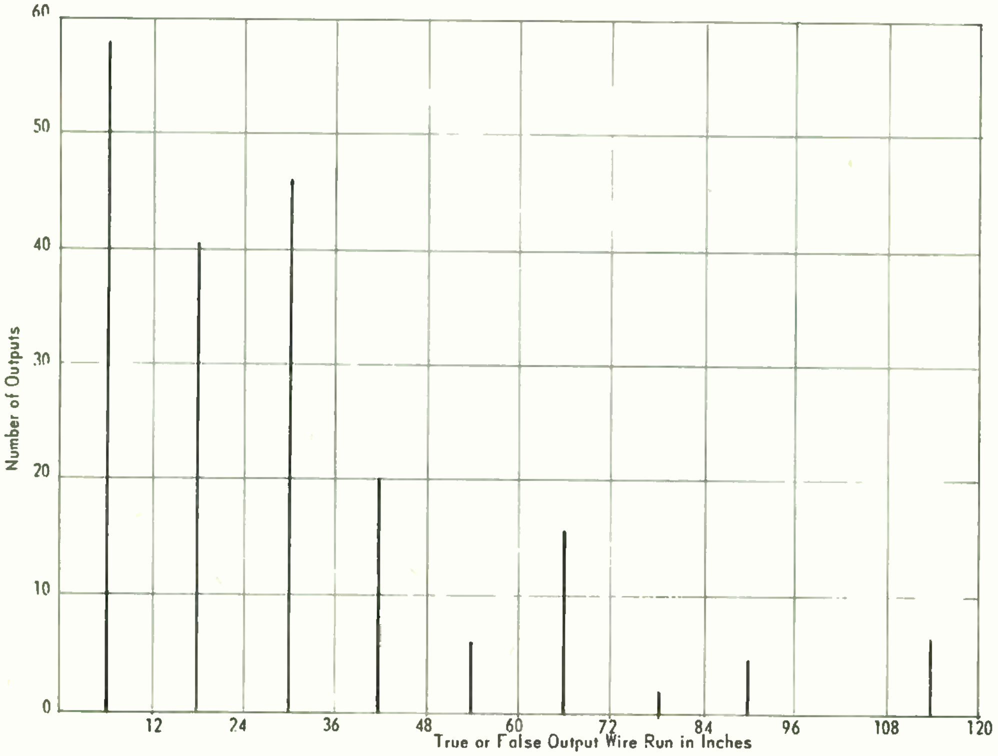


Fig. 7. Histogram of wire lengths.

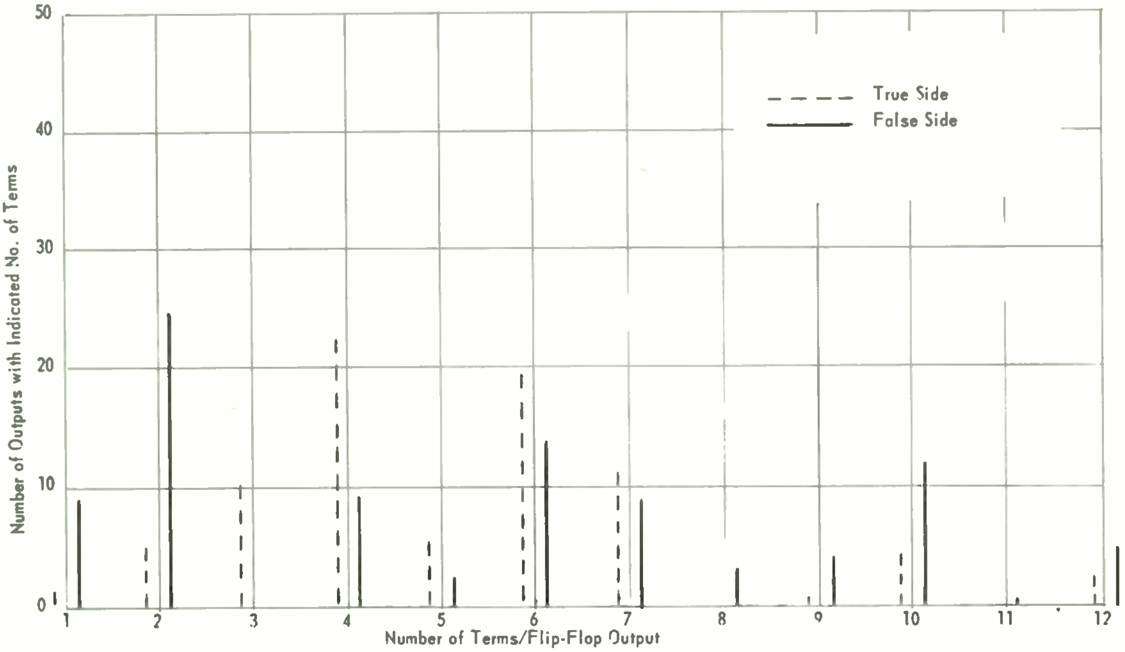


Fig. 8. Histogram of flip-flop loads.

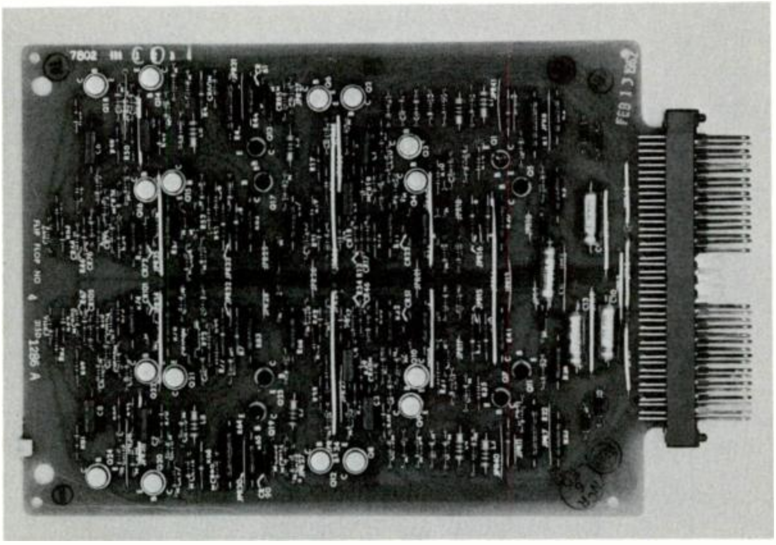


Fig. 9. 315 Flip-flop circuit board.

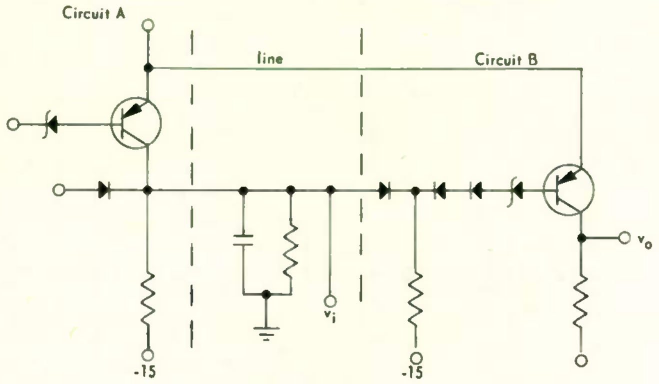
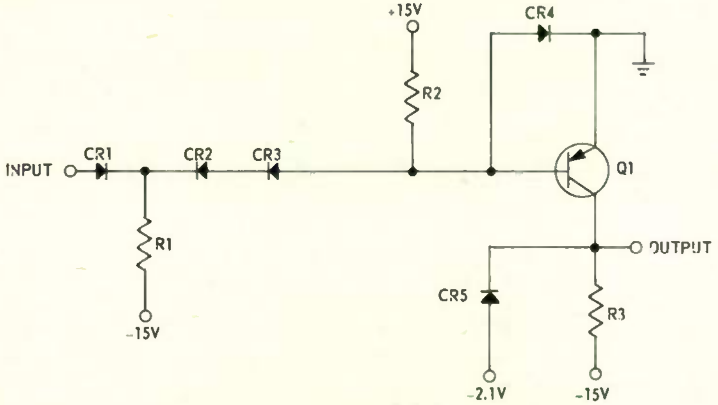


Fig. 10. Line drive and termination schematic diagram.



SPECIFICATIONS

INPUT:	upper logic level:	0 to -0.55V
	lower logic level:	-2.1V to-2.6V
OUTPUT:	load:	5 products
		100μμf max
		7.5 ma nominal
TIMING:	Delay:	0.25μsec max
	rise and fall times:	60m μsec max

Fig. 12. Logic driver schematic diagram.

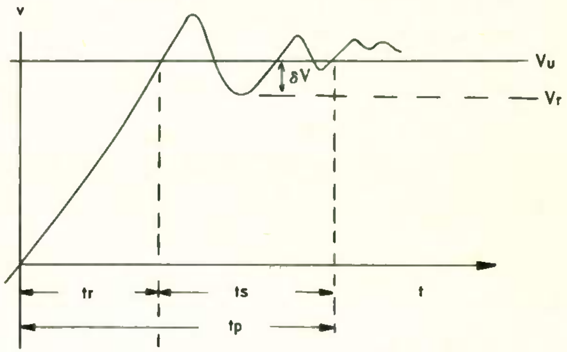
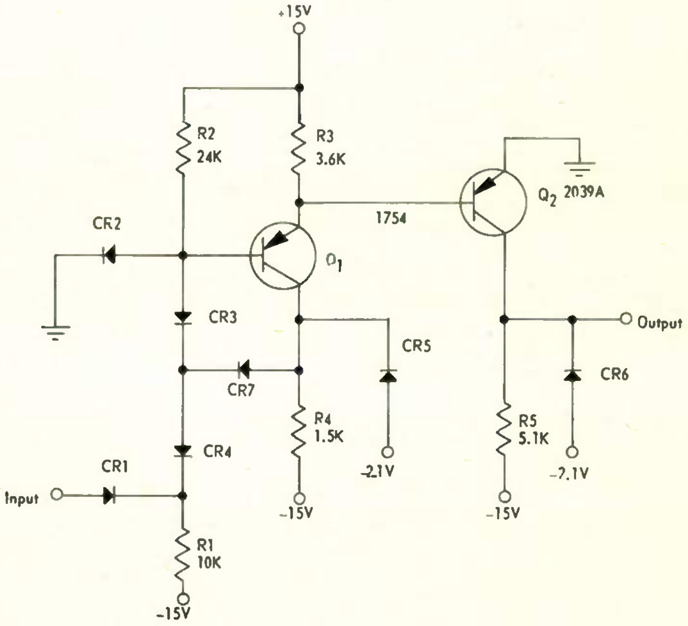


Fig. 11. Loaded line response waveform.



INPUT:	upper logic level:	0 to -0.55V
	lower logic level:	-2.1V to -2.6V
OUTPUT:	20 products	
	400μμf max	
	32 ma max	
TIMING:	delay:	0.2 μsec max
	rise and fall times:	60m μsec max

Fig. 13. Power driver schematic diagram.

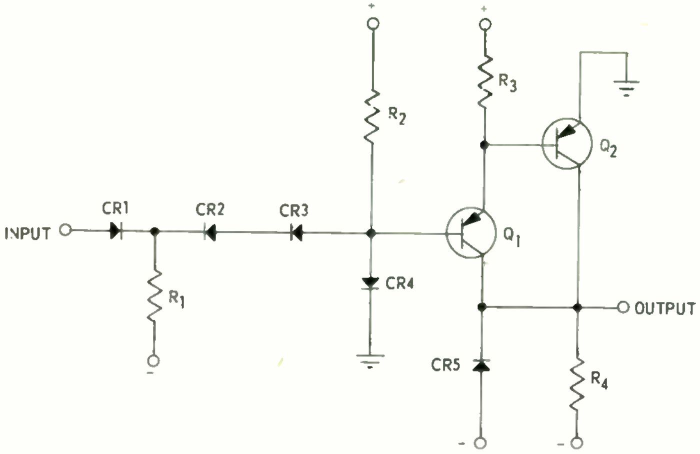


Fig. 14. Power driver employing darlington pair.

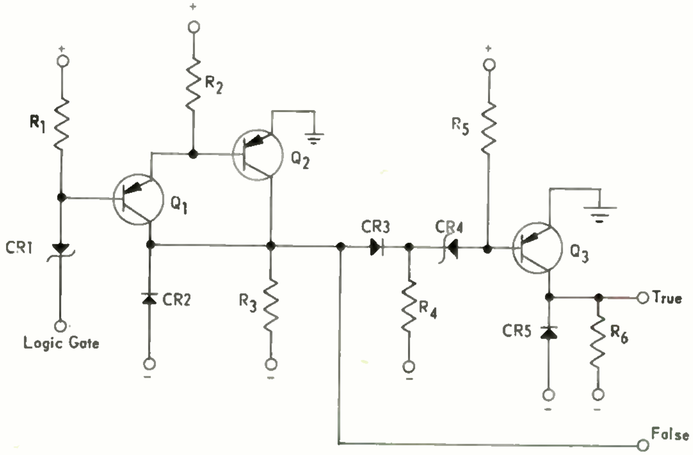
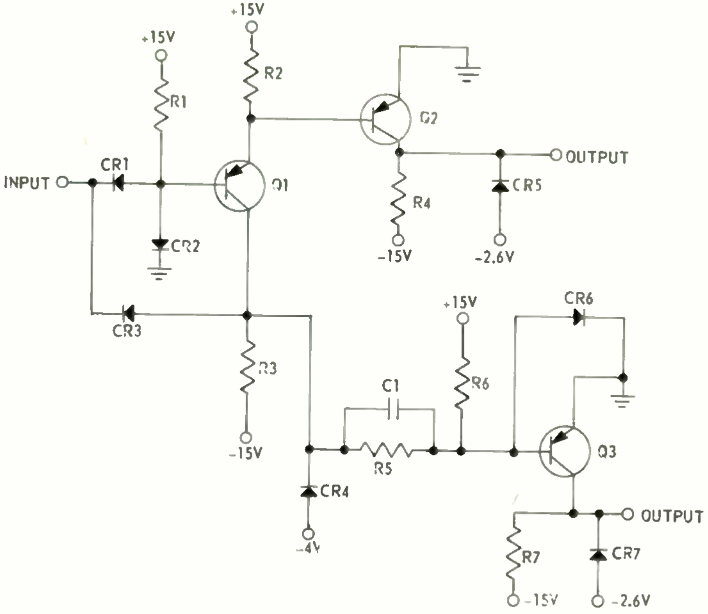


Fig. 16. Double driver employing a cascaded power driver and single driver.



SPECIFICATIONS

INPUT:	upper logic level:	0 to -0.55V
	lower logic level:	-2.1V to -2.6V
OUTPUT:	load:	8 products
		96 µµf max
		12 ma nominal
TIMING:	delay:	0.1µsec max
	rise time:	40m µsec max
	fall time:	50mµsec max

Fig. 15. Double driver schematic diagram.

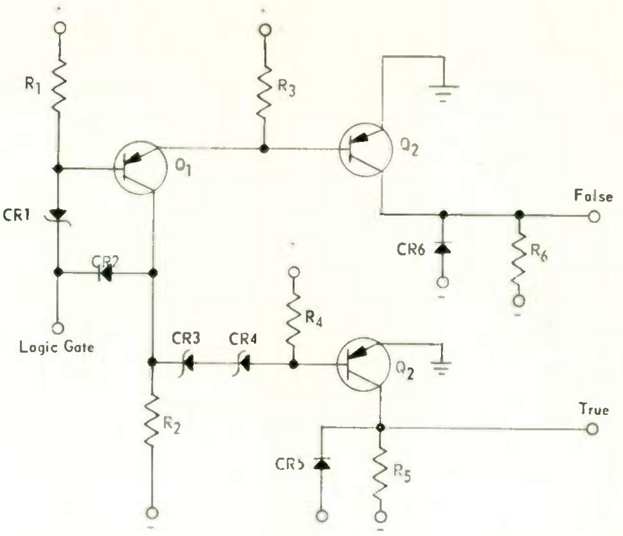


Fig. 17. Double driver employing a phase splitter.

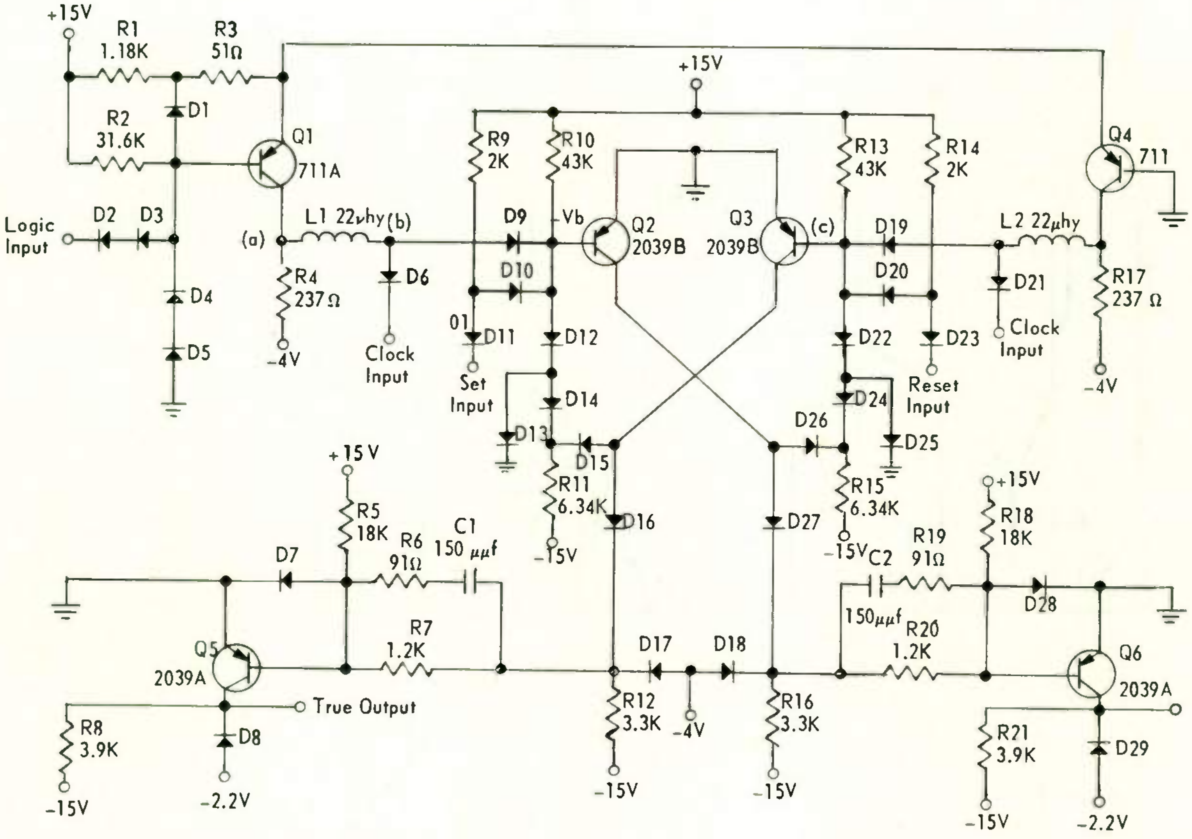


Fig. 18. Flip-flop schematic diagram.

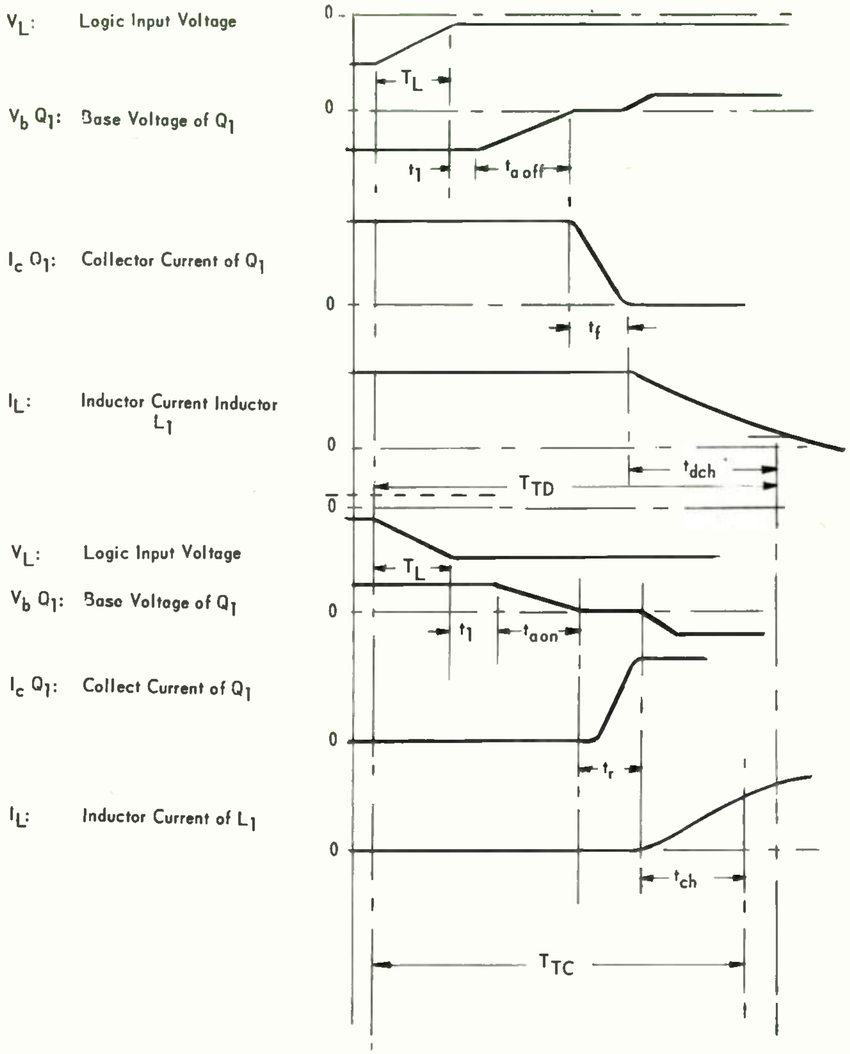

V_L: Logic Input Voltage
V_b Q_1: Base Voltage of Q_1
I_c Q_1: Collector Current of Q_1
I_L: Inductor Current Inductor L_1
0
T_L
t_1
$t_{a\,off}$
t_f
t_{dch}
T_{TD}
V_L: Logic Input Voltage
V_b Q_1: Base Voltage of Q_1
I_c Q_1: Collect Current of Q_1
I_L: Inductor Current of L_1
t_{aon}
t_r
t_{ch}
T_{TC}

Fig. 19. Pedestal set timing.

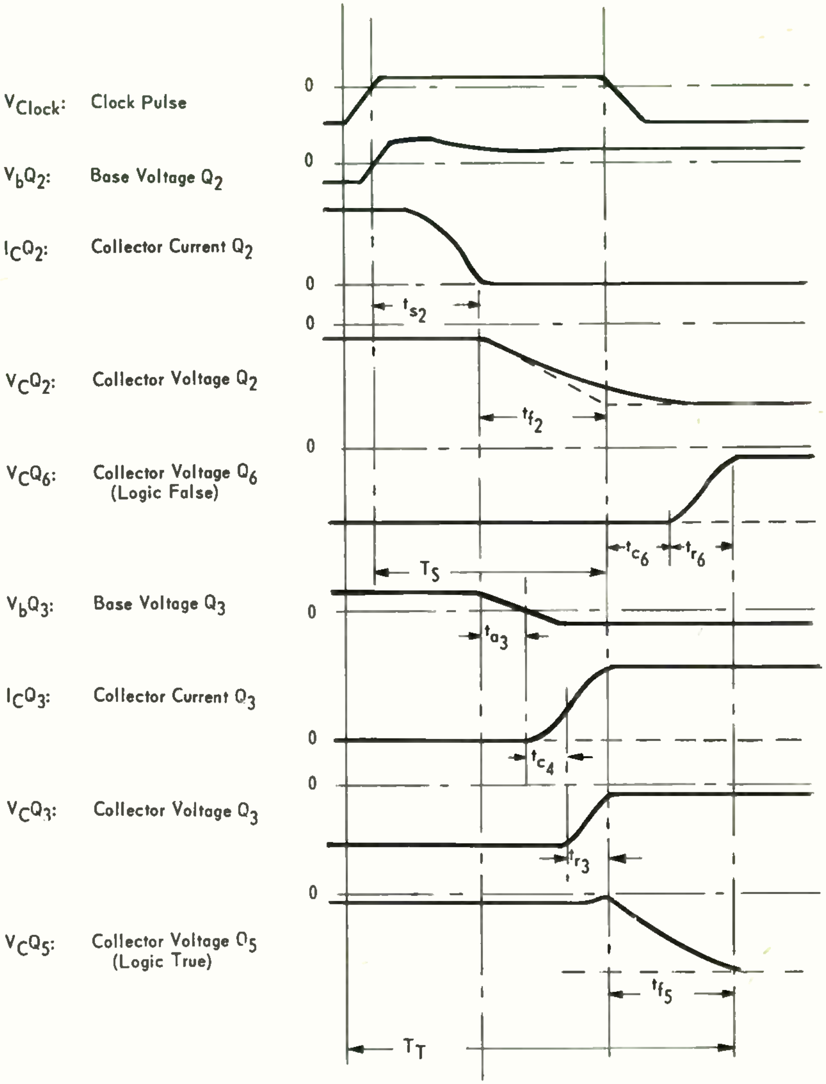
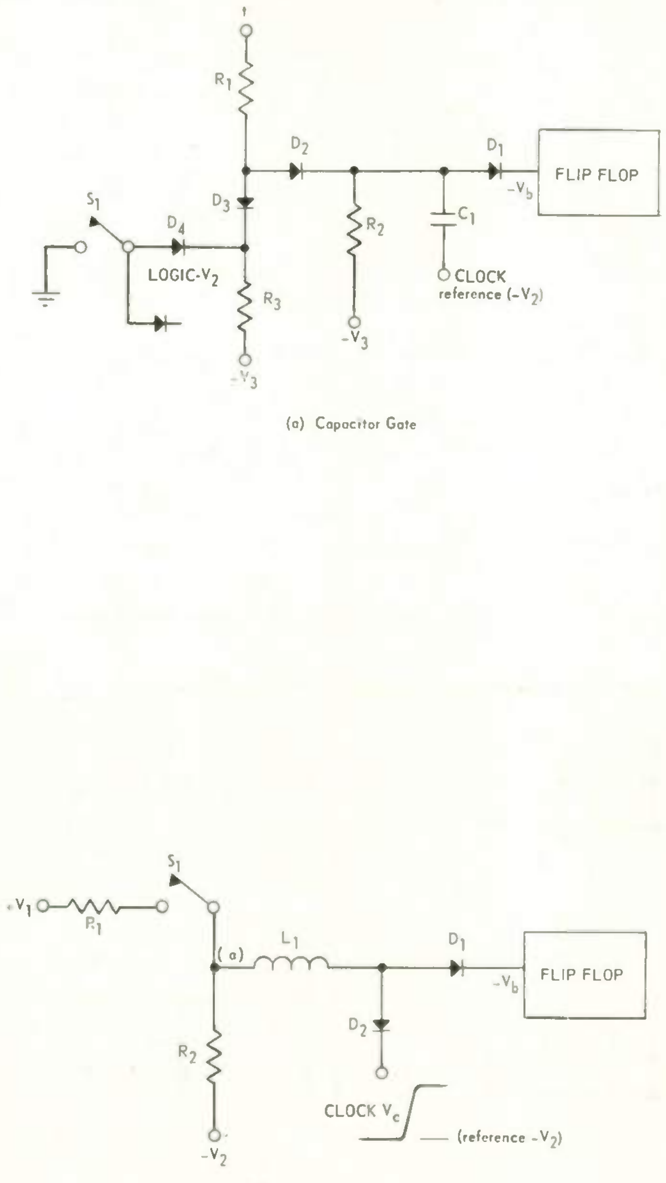

V_{Clock}: Clock Pulse
V_bQ_2: Base Voltage Q_2
I_CQ_2: Collector Current Q_2
0
t_{s_2}
V_CQ_2: Collector Voltage Q_2
t_{f_2}
V_CQ_6: Collector Voltage Q_6 (Logic False)
t_{c_6}
t_{r_6}
T_S
V_bQ_3: Base Voltage Q_3
t_{a_3}
I_CQ_3: Collector Current Q_3
t_{c_4}
V_CQ_3: Collector Voltage Q_3
t_{r3}
V_CQ_5: Collector Voltage Q_5 (Logic True)
t_{f_5}
T_T

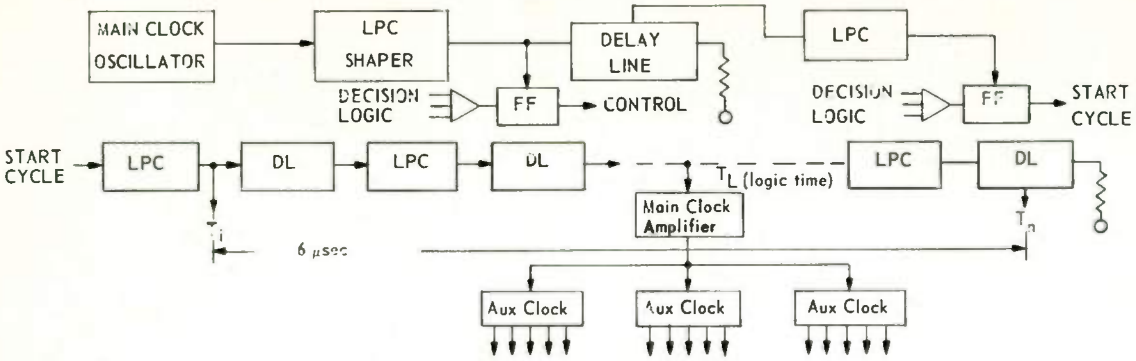
Fig. 20. Flip-flop regeneration timing.



(a) Capacitor Gate

(b) Inductor Pedestal Gate

Fig. 21. Clock gates.



CLOCK SYSTEM C-315 PROCESSOR

CLOCK CIRCUITS	NUMBER OF OUTPUTS	LOAD/ OUTPUT	PULSE AMP	PULSE WIDTH	PULSE RISE TIME	REPETITION RATE	CKT DELAY	LINE LENGTH
Main Clock (Oscillator and Shaper)	1	LPC	4V	Square-Wave	60 msec max	333KC	- - -	- - -
	1		4V	Square-Wave	60 msec max	100KC	- - -	- - -
Main Clock Amplifier (MCA)	3	LPC	18.6V	0.15μsec	40 mμsec	200KC	0.7μsec	3 ft.
Low Power Clock Shaper Amplifier (LPC)	1	1FF 4FF	9.7V 9.7V	0.2μsec 0.2μsec	40 mμsec 40 mμsec	333KC 200KC	0.7μsec 0.7μsec	- - - - - -
		6FF 12FF	9.7V 9.7V	0.2μsec 0.2μsec	40 mμsec 40 mμsec	167KC 100KC	0.7μsec 0.7μsec	5 ft. 5 ft.
Auxiliary Clock Amplifier (ACA)	8	8FF	9.7V	0.1 - 0.2 μsec	40 mμsec	200KC	50μsec	2.7 ft.

Fig. 22. Clock system and circuit performance.

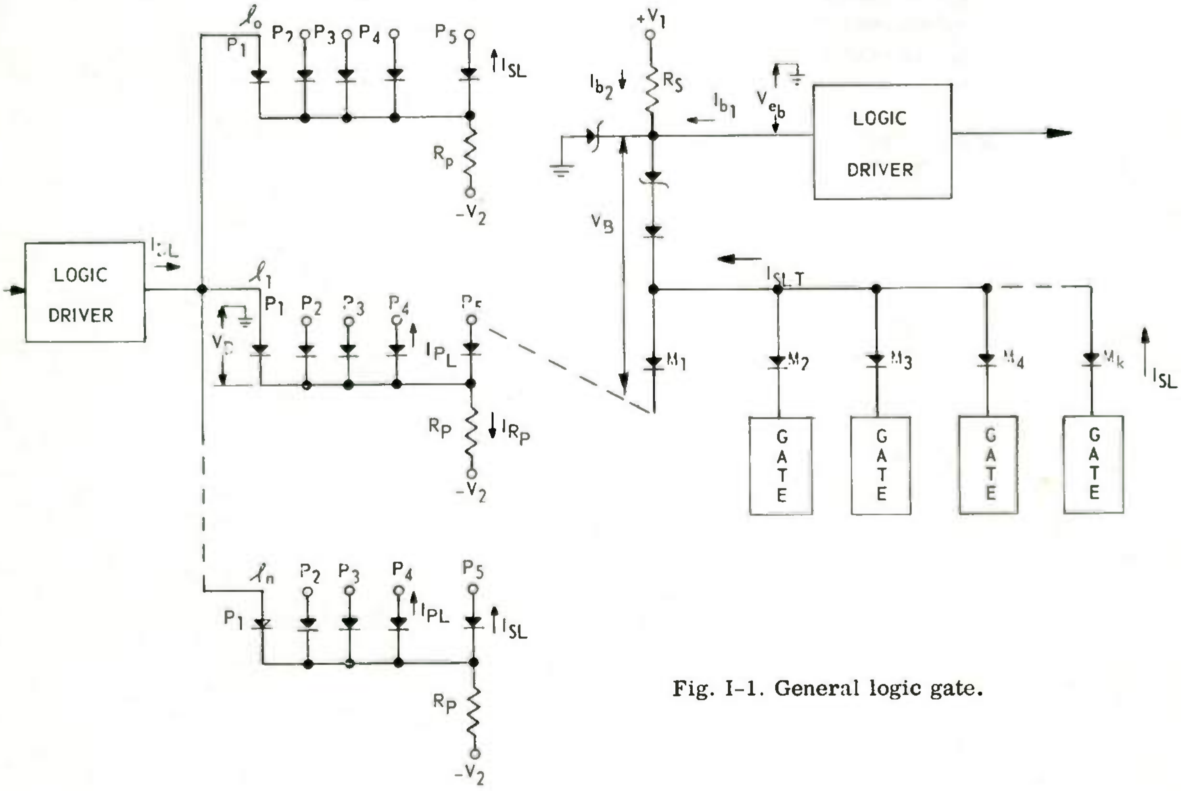


Fig. I-1. General logic gate.

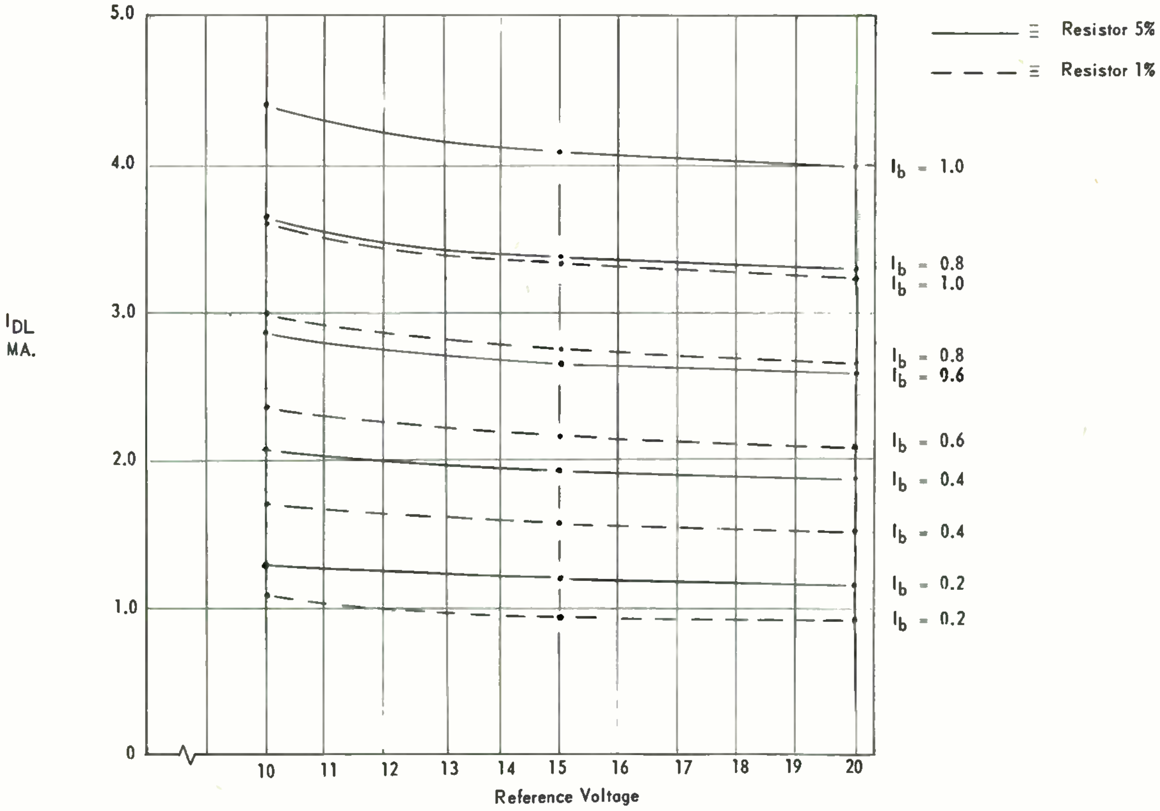


Fig. I-2. Plot of I_{DL} vs V_1 and V_2.

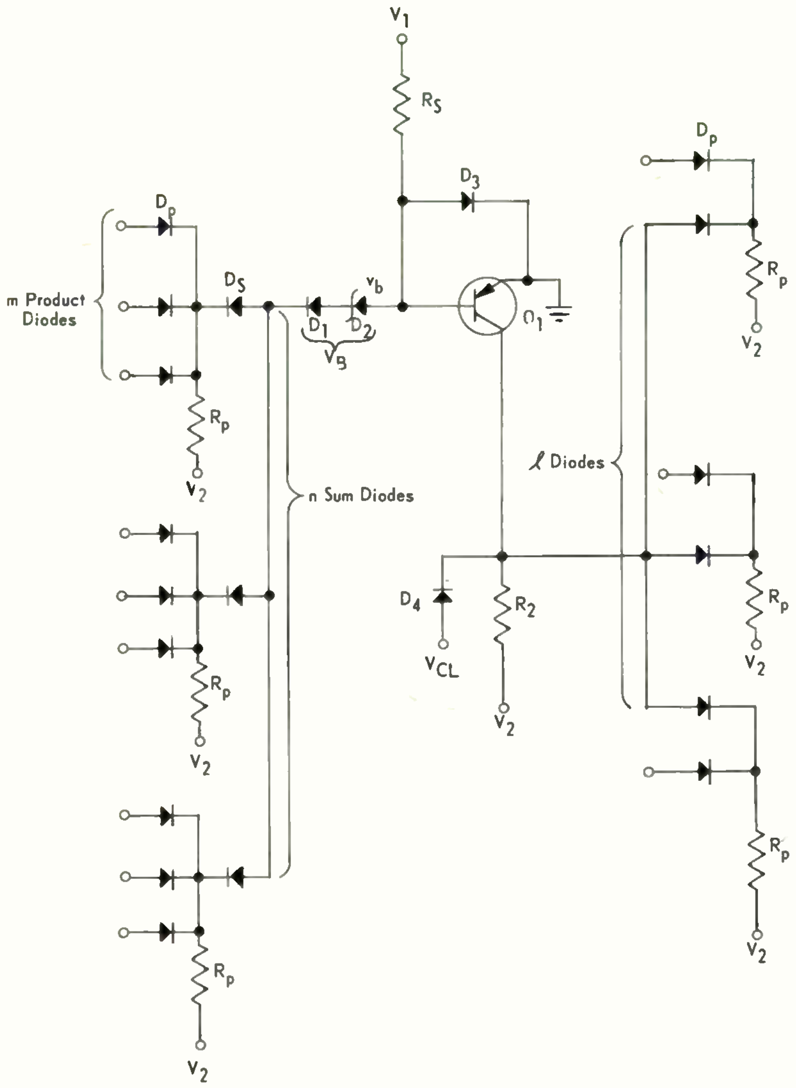


Fig. III-1. Diode coupled inverter schematic diagram.

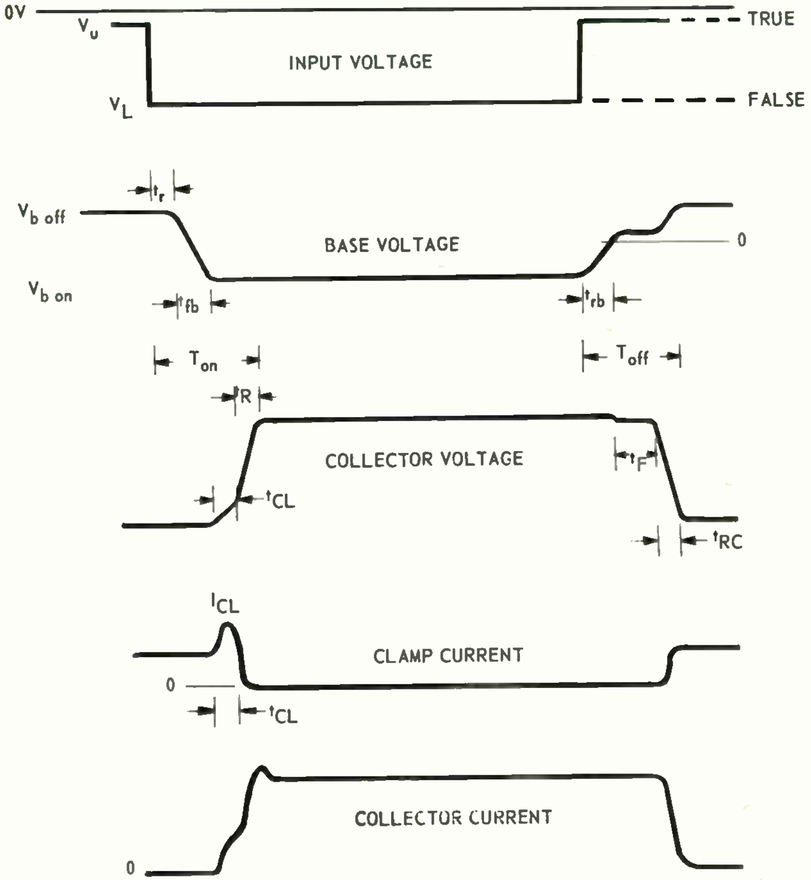


Fig. III-2. Inverter timing diagram.

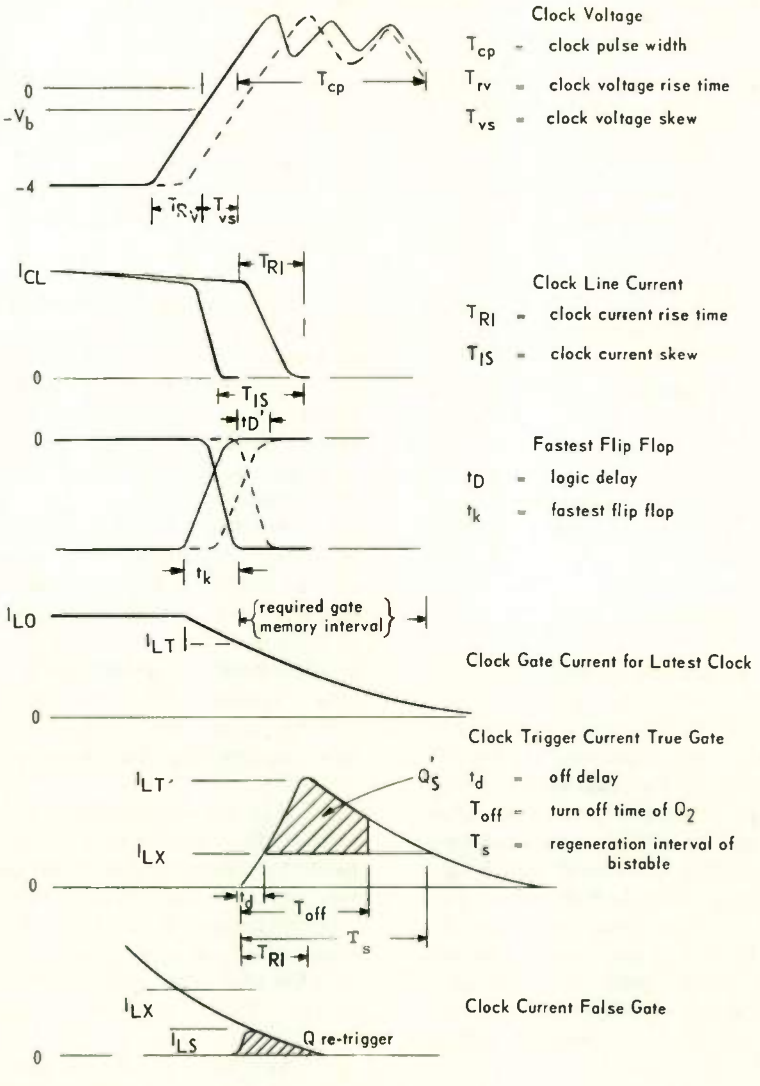

Clock Voltage
T_{cp} - clock pulse width
T_{rv} = clock voltage rise time
T_{vs} = clock voltage skew
0
$-V_b$
-4
T_{cp}
T_{Rv}
T_{vs}
I_{CL}
T_{RI}
Clock Line Current
T_{RI} = clock current rise time
T_{IS} = clock current skew
0
T_{IS}
t_D'
0
Fastest Flip Flop
t_D = logic delay
t_k = fastest flip flop
t_k
I_{LO}
{required gate memory interval}
I_{LT}
Clock Gate Current for Latest Clock
0
Clock Trigger Current True Gate
$I_{LT'}$
Q_S'
t_d = off delay
T_{off} = turn off time of Q_2
T_s = regeneration interval of bistable
I_{LX}
0
t_d
T_{off}
T_s
T_{RI}
Clock Current False Gate
I_{LX}
I_{LS}
Q re-trigger
0

Fig. IV-1. Clock gate charge conditions.

GENERALIZED PULSE RECORDING †

Irving Stein

Ampex Corporation
Research Department
Redwood City, California

Summary: Tape magnetization characteristics resulting from an input switching are analyzed from a general point of view. A previously developed theory[1] is first applied to the simple switching between zero and saturation field levels to determine the recorded magnetization and pulse shape. From this example the characteristics of RZ pulses can be determined. The analysis is then extended to determine the recorded characteristics resulting from switching between saturation levels (NRZ). These two cases of switching are the basis for analyzing the tape magnetization resulting from more general switching. Switching to 'oversaturated' levels is investigated, as is switching between any two input levels. Switching with a non-zero time constant is considered, as is switching between saturation levels with a time delay at zero. Finally, the recorded magnetization and pulses resulting from consecutive saturation switchings are determined.

1. Introduction

In this paper we investigate the recording rather than the reproduction of pulses on a magnetic storage medium such as a tape or drum.We analyze the magnetization recorded on the tape as a result of switching of the input head field (specifically the longitudinal component). Thus, we are not concerned with a particular form of pulse recording such as NRZ* or RZ, but with the tape magnetization resulting from input switching between any two input levels. We will find, however, that it will first be necessary to analyze the simplest kind of recording, i.e., switching from zero to saturation and vice versa (which we designate as ZS and SZ respectively). Then, using a simple hysteresis model for the tape, previously developed[1], we analyze NRZ and pulse recording in general.

In the interest of clarity, we define 'pulse recording' as the recording of a sudden change, or switching, of the level of the input signal field. NRZ (non-return-to-zero), ZS (zero-saturation), and SZ (saturation-zero) are particular kinds of pulse recordings. The first is a change between the saturation levels of opposite polarities, the second a change from zero to saturation, and the third a change from saturation to zero. We do not restrict ourselves to these special cases, however, but include changes between any two levels, including levels 'over' saturation.

Since the usual reproduce head is a 'differentiating' head, i.e., one whose voltage is the time derivative of the magnetic flux it detects from the tape, it is primary that the slope of the magnetization, i.e., the magnetization variation per unit length, be analyzed. Since it is not the magnetization but its space derivative or slope which is the source of the magnetic reproduce head output, it is the characteristics of the latter with which we shall be mainly concerned. The slope will be referred to as the 'recorded pulse', or simply the 'pulse'. If the reproduce head gap is small compared to the write head gap, and if the reproduce head-tape spacing is small, then the actual pulse output of the reproduce head will approximate the recorded pulse. The pulse, therefore, is considered to exist on the tape. The reproduce pulse, i.e., the output of the reproduce head, is generally the recorded pulse broadened.

It is also necessary to define a number of other terms. By 'saturation' field we mean that record head field at the intersection of the gap center line and the tape which is just sufficient to saturate the tape. By 'oversaturation' field we mean a field larger than the saturation field, and thus one which will not only saturate the tape at the gap centerline but over an extended region. This is shown in Figure 1.

As a matter of convention, the motion of the tape will be defined as moving from left to right, or in a positive direction. The 'front of the gap' will be the positive half of the gap; the 'rear of the gap' will be the negative half of the gap.

We concern ourselves here only with the basic characteristics of pulse recording and not with accurate pulse shape determination. To this end, we make a simplifying assumption about the tape hysteresis curves that removes most of the mathematical difficulties and permits the analysis of the basic recording process to

† The author wishes to thank Paul Vadapalas of the Ampex Corp. Research Dept. for checking most of the equations and doing the numerical work for some of the graphs.

* All notation is defined on page .

be made. We assume that $M = kH$, where k, the susceptibility, is a constant; and furthermore that the application of reverse fields also results in a magnetization linear with the field. These curves are shown in Figure 2.

We will find that the calculated pulse shape does not correspond to the observed shape, but that if we take the non-linearity of the magnetization curve into account, we can make our theory correspond to observation. Thus, our simplified approach is highly significant since it allows us to determine the separate effects of the field and the magnetization curve on the distortion, specifically the asymmetry, of the pulse. It is to be further noted that the theory developed does not include the irreversible demagnetization of the shorter wavelength components of the pulse. An understanding of this requires development of particle interaction theory.

2. Saturation (S) Recording

S recording is the simplest kind of recording and is basic not only to RZ (return-to-zero) but also to NRZ and the more general forms of switching. It can be, as defined previously, either a turn-on (ZS) or a turn-off (SZ) input.

We see that for a linear magnetization curve a ZS input simply impresses the head field onto the tape. Thus, a ZS input creates two pulses on the tape, symmetrically placed with respect to the gap centerline. The front pulse is fixed in the tape and moves with the tape; the rear pulse is fixed relative to the gap and remains stationary as the tape moves. This is illustrated in Figure 3. On the other hand, an SZ input creates no pulses but fixes the rear pulse in the tape. Thus, ZS and SZ inputs are not recorded at the same place, the former being recorded in the front of the gap and the latter in the rear of the gap. This is seen in Figure 4. If c is the distance between the two pulses at $t = 0$, then at any later time the distance between them is $vt + c$. If an SZ input is followed by a ZS input, then the distance between them after time t is $vt - c$. Thus, since ZS and SZ pulses are displaced respectively $\pm c/2$ from the gap centerline, the recorded pulses bear a different relationship to each other than the input pulses, a situation not true of NRZ pulses. The value c is a function of g and other parameters and will be evaluated later. Obviously the pulses can be no closer than a distance c. Thus, the packing density of S recording is specifically limited by c, which is a function of the gap size and head-tape spacing. Therefore, RZ recording, which is the recording of two switchings -- a ZS followed by an SZ -- has a width limited by the gap size and spacing. On the other hand, the location of an RZ pulse is not affected by spacing.

In this connection, it is interesting to note that dropouts for a saturated tape less than a gapwidth in extent will not interfere with the recording of pulses; i.e., if an SZ input is almost immediately followed by a ZS input, the latter will, to a large extent, erase the former. On the other hand, a drop-in, i.e., where a ZS input is followed immediately by an SZ input, will record two pulses on the tape spaced at least a gapwidth apart.

Since the magnetization space variation on the tape in SZ recording is the same as the head field variation (for the tape hysteresis curve), a knowledge of the head field is required. If we define $g \equiv \frac{y}{\ell/2}$ and $s \equiv \frac{x}{\ell/2}$, then to a very good approximation for $g > 1/4$, the field is given by the formula[2]:

$$H(s,g) = \frac{H_o}{\pi}\left[\tan^{-1}\left(\frac{1+s}{g}\right) + \tan^{-1}\left(\frac{1-s}{g}\right)\right] \quad . \; . \; . \; . \; . \; . \; (1)$$

It is interesting to note that the lines of constant field are circular arcs intersecting the gap vertices.

The magnetization on the tape is given simply as:

$$M(s,g) = k\,H(s,g) \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; (2)$$

while the equation for the pulse on the tape is given as:

$$M'(s,g) \equiv P(s,g)$$

$$\equiv \frac{M_o g}{\pi}\left[\frac{1}{g^2 + (1+s)^2} - \frac{1}{g^2 + (1-s)^2}\right] \; . \; . \; . \; (3a)$$

or

$$P(s,g) = \frac{4M_o}{\pi}\left[\frac{sg}{4g^2 + (g^2+s^2-1)^2}\right] \; . \; . \; . \; . \; . \; . \; . \; . \; (3b)$$

where $M_o = k\,H_o$.

We immediately see that the pulse is asymmetrical, having a much smaller slope for a large s than for a small s. In order to compare the pulse shapes as a function of g, we must consider the field normalized at the gap centerline; i.e.

$$H_n(s,g) = \frac{H(s,g)}{H(o,g)}$$

$$= \frac{1}{2\tan^{-1}\left(\frac{1}{g}\right)}\left[\tan^{-1}\left(\frac{1+s}{g}\right) + \tan^{-1}\left(\frac{1-s}{g}\right)\right] \; . \; . \; (4a)$$

Then the normalized pulse is given as:

$$M'(s,g) \equiv P_n\,(s,g) = \frac{P(s,g)}{H(o,g)}$$

$$= \frac{2}{\tan^{-1}\left(\frac{1}{g}\right)} \left[\frac{sg}{4g^2 + (g^2+s^2-1)^2} \right] \quad \ldots\ldots \quad (4b)$$

where $P_n(s,g)$, as a function of s, is illustrated in Figure 5.

The effect of spacing on field strength is shown in Figure 6, where the normalizing factor

$$H(o,g) = \frac{2H_o}{\pi} \tan^{-1}\left(\frac{1}{g}\right) \quad \ldots\ldots\ldots\ldots\ldots \quad (4c)$$

is plotted as a function of g. The actual pulse, P(s,g), is obtained by multiplying together corresponding points in Figures 5 and 6. We see that spacing causes a very sharp increase in pulse width and sharp decrease in pulse amplitude -- even the normalized pulse amplitude. We also clearly observe from Figure 5 the pulse asymmetry, which is a function here of the field shape only.

Observed pulses in SZ recording have a reverse symmetry, the smaller slope appearing first and the steeper slope appearing in time after the pulse peak. It will be shown in the next section, where we discuss 'oversaturated pulse recording', that this reversal of asymmetry is due to the non-linearity of the magnetization curve. It is interesting to note that if the tape were a perfect square-loop material, then the recorded magnetization would be identical in shape with the input current no matter what the size of the gap or the shape of the head. Furthermore, the pulse would always be recorded at the gap centerline (assuming the tape was no closer to the head than $\sim \frac{\ell}{8}$, where the 'double hump' field shape begins to appear).

In order to determine the pulse maximum location, s_o, we set

$$\frac{dP(s,g)}{ds} = 0 \quad \ldots\ldots\ldots\ldots\ldots \quad (5a)$$

This gives:

$$s_o^2 = \frac{1}{3}\left[-(g^2-1) + \sqrt{(g^2-1)^2 + 3(g^2+1)^2}\,\right] \qquad (5b)$$

The location (s_o) of the pulse maximum as a function of relative spacing is plotted in Figure 7. As g approaches infinity, s_o approaches $\frac{g}{\sqrt{3}}$. We notice that for $g \leq 1$, s_o is essentially constant, and then it increases at a rate rapidly approaching $\frac{g}{\sqrt{3}}$. Thus, the location of NRZ pulses resulting from a single switching is a function of spacing, while the location of an RZ pulse resulting from two switchings is not a function of spacing. RZ recording, therefore, is not as susceptible to the effect of dropouts as NRZ recording.

From Figure 5 we notice, even for $g < 1$, a considerable pulse broadening as g increases. Thus, although the peak location hardly changes for $g \leq 1$, the pulse does broaden considerably. Although 'pulse width' can be determined from the given pulse curves, we do not calculate it because of its non-standard definition. Probably the most reasonable definition of 'pulse width' is "the area under the pulse curve divided by $P(s_o,g)$". The formula for this pulse width (W) is:

$$W = \frac{\left[H(\infty,g) - H(o,g)\right]}{P(s_o,g)} \quad \ldots\ldots\ldots \quad (6)$$

The value of the normalized pulse maximum, $P_n(s_o,g)$, can now be easily determined. It is plotted as a function of g in Figure 8. It is noticed that the pulse amplitude decreases very sharply with spacing. For $g >> 1$, it can be shown that:

$$P_n(s_o,g) = \frac{1}{\pi\sqrt{3}}\left(\frac{1}{1+\frac{4g^2}{9}}\right) \longrightarrow \frac{3\sqrt{3}}{4\pi g^2} \quad \ldots\ldots \quad (7a)$$

For $g << 1$, it can be shown that:

$$P_n(s_o,g) = \frac{1}{2\pi g} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (7b)$$

Thus, the relative change is much greater for larger spacing.

3. Oversaturated Recording

If the input oversaturates the tape; i.e., if the field at the gap centerline at a given g is more than sufficient to saturate the tape there, then not only a point but a section of the tape will become saturated. If $\alpha \equiv \left(\frac{H}{Hs}\right)$ is the oversaturation factor, then the half-length of tape saturated (s_α) upon applying a ZS input is determined by the formula:

$$\alpha\left[\tan^{-1}\left(\frac{1+s_\alpha}{g}\right) + \tan^{-1}\left(\frac{1-s_\alpha}{g}\right)\right] = 2\tan^{-1}\left(\frac{1}{g}\right) \quad \ldots \quad (8a)$$

or

$$\alpha \tan^{-1}\left(\frac{2g}{g^2+s_\alpha^2-1}\right) = 2\tan^{-1}\left(\frac{1}{g}\right) \quad \ldots\ldots\ldots\ldots \quad (8b)$$

Then $\pm s_\alpha$ is the dividing point between saturated and unsaturated tape. Oversaturation α increases the pulse magnitude everywhere below saturation by α. Therefore, as long as $|s_\alpha| \leq |s_o|$, the pulse amplitude increases by α and its location remains fixed. At $\alpha = \alpha_o$, i.e., where $s_\alpha = s_o$, the tape is saturated at the pulse amplitude location. For $\alpha > \alpha_o$, the pulse amplitude begins to decrease and its location begins to change. Specifically, it moves away from the gap centerline. Then α_o, as a function of g, is determined from Equation (8), where s_o is determined from Equation (5b) or Figure 7. Figure 9 is a plot of α_o versus g. We see that the greatest increase in current (up to 80%) required to bring the pulse amplitude up to saturation is for $g < 1$. For $g > 1$, the increase drops rapidly to 30%. Even for a non-linear virgin magnetization curve, the increase required for pulse amplitude saturation is significant. It should be noted that increasing the pulse amplitude does not better the resolution, since each point on the pulse is increased proportionately.

The magnitude of the pulse amplitude for $\alpha < \alpha_o$ is given by:

$$P(\alpha, g) = \frac{4M_o}{\pi}\left[\frac{\alpha\ g\ s_o}{4g^2 + (g^2+s_o^{\ 2} - 1)^2}\right] \dots\dots\dots \quad (9)$$

The manner in which $P(\alpha, g)$ changes for $\alpha < \alpha_o$ can be estimated by evaluating the extreme cases, $g << 1$ and $g >> 1$.

For $g << 1$, from Equation (8b):

$$s_o^{\ 2} \cong 1 + \frac{2\alpha g}{\pi}, \ s > 1 \dots\dots\dots\dots \quad (10a)$$

For $g >> 1$:

$$s_o^{\ 2} \cong g^2 (\alpha - 1) \dots\dots\dots\dots\dots\dots \quad (10b)$$

Then for $g << 1$:

$$P(\alpha, g) \cong \frac{M_o}{\pi}\left[\frac{\alpha\ (1 + \frac{\alpha g}{\pi})}{g\ (1 + \frac{\alpha^2}{\pi 2})}\right] \dots\dots\dots \quad (11a)$$

Thus, for α not too large and g very small, $P(\alpha, g)$ is proportional to $\frac{\alpha}{g}$. For α quite large, $P(\alpha, g)$ varies inversely with both α and g.

For $g >> 1$,

$$P(\alpha, g) \cong \frac{4M_o}{\pi}\left[\frac{\alpha\sqrt{\alpha - 1}}{4 + \alpha^2 g^2 - 2\alpha} \longrightarrow \frac{\sqrt{\alpha - 1}}{\alpha\ g^2} \longrightarrow \frac{1}{g^2\sqrt{\alpha}}\right] \dots \quad (11b)$$

Thus, for $g >> 1$,the pulse amplitude decreases inversely with the square of g. Generally, the pulse amplitude decreases approximately inversely with the square root of the oversaturation factor, α.

It was previously mentioned (pg) that the observed asymmetry is the reverse of that calculated in Section 2. This reversal, as has been stated, is due to the non-linearity of the magnetization curve. The actual magnetization curve for γ - Fe_2O_3, as shown in Figure 10a,[3] is seen to approach saturation quite gradually; in fact, saturation is reached at a value approximately twice the value it would reach if the curve were linear. In actuality, the pulse for saturation recording will resemble an oversaturated pulse of $\alpha \sim 2$ (based on our linear model) except that the magnetization in the tape will rise much more slowly in the region over the gap, as can be seen from Figures 10b and 10c. Thus, the slope will be much smaller in this region than predicted by the simple linear model. Furthermore, the small slope at the toe of the magnetization curve will tend to decrease more rapidly than the tape magnetization in the region further away from the gap edge (to the left in Figures 10b and 10c). Thus, the slope of the magnetization in this region is increased, tending to further reverse the asymmetry. Therefore, the non-linearity in both the low and high field portions of the magnetization curve will tend to reverse the asymmetry due to the field shape alone. Clearly, then, the shape of the pulse is a function of the tape materials used, the head structure, and the head-to-tape spacing.

4. NRZ Recording

In order to understand the basis of NRZ recording we must expand the concepts used for S recording. Specifically, we now need the complete model that describes the magnetization resulting from a field reversal, as defined on pg The recording of an NRZ pulse through the use of a magnetic head gap can be viewed as an ordered application of two fields, H_1 and H_2, of opposite polarity with a law of combination:

$$M_n = M_1 - 2M_2 = k\left[H_1 - 2H_2\right], \ H_2 \leq H_1$$

$$M_n = -M_2 = -kH_2, \ H_2 \geq H_1 \dots\dots\dots\dots \quad (12)$$

This model is described in Reference 1 and illustrated in Figure 2. The magnetization process itself is seen in Figure 11.

In other words, we view the switching of the polarity of the input signal as a combined process; namely, as the ordered application of two S inputs, the first SZ and the second ZS of opposite polarity. The fields combine in a simple algebraic manner to give a net magnetization. It is to be noted that even though these two fields are quite distinct timewise, they do overlap spacewise in the region of the gap.

Now, if H(s) is the field shape around the gap, then for an NRZ pulse not above saturation we have:

$$H_1 = H(s)$$
$$H_2 = -H(-s) \dots\dots\dots\dots\dots \quad (13)$$

Now, S recording is such that a turn-off fixes a magnetization change in the tape occurring over the rear part of the gap, while a turn-on fixes a magnetization change in the tape occurring over the front part of the gap. Thus, if the absolute value of the maximum field is designated as

$$H_1 = \frac{2H_o \tan^{-1}}{\pi} \left(\frac{1}{g}\right)$$

then Equations (12) and (13) give us:

$$M_n = k\left[H_1 - 2H(-s)\right] \quad , H_2 < H_1$$

$$M_n = -kH(-s) \qquad , H_2 > H_1 \, . \, . \, . \, . \, . \, . \, . \, . (14)$$

Again we consider fields at sufficient distance from the head so that the head field is a maximum at the gap center, a distance larger than $g \sim 1/4$.

The equations imply:

$$M_n(s) = -kH_1 \qquad s \leq 0 \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, (15)$$

This merely expresses the fact that the tape in the region $s \geq 0$ will pass the gap center line and achieve maximum remanence at H_1. (Of course, we are assuming that no other reversals take place.) The pulse is created, then, over the front half of the gap. It is there that the 'interference' of the two magnetizations takes place and M(s) spatially reverses itself.

We immediately see that the shape and location of an NRZ and a ZS pulse are identical except that the magnitude of the NRZ pulse is twice as large. The magnetization, of course, is quite different. It is interesting to compare the NRZ crossover point with the pulse location. The crossover point, s_c, is defined at:

$$M(s_c) = 0 \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, (16a)$$

or:

$$H(s_c) = \frac{H_o}{\pi} \tan^{-1} \left(\frac{1}{g}\right) \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, (16b)$$

Therefore:

$$\frac{H_o}{\pi} \tan^{-1} \left(\frac{2g}{g^2 + s_c^{\,2} - 1}\right) = \frac{H_o}{\pi} \tan^{-1} \left(\frac{1}{g}\right) \, . \, . \, . \, . \, . \, . \, (17)$$

Thus, the locus of crossover points as a function of spacing is given by:

$$s_c^{\,2} = g^2 + 1 \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, . \, (18)$$

This is plotted in Figure 7. It is seen that for $g >> 1$, $s_c \rightarrow g$ and the crossover point increases linearly with spacing. Since for $g >> 1$, $s_o \rightarrow \frac{g}{\sqrt{3}}$, we see that s_c and s_o increase in the above manner with g, but that s_o is always considerably closer to the gap center. For $g < 1$, the crossover point is never within the gap but always at the edge or beyond. A comparison here with the pulse location shows that for small spacing the pulse amplitude and crossover points are almost identical. We see, therefore, that across any thickness of tape the crossover point will vary more rapidly than the pulse amplitude point; thus the data will be recorded with poorer resolution.

It is to be noted that the recording of an NRZ input always takes place over the front half of the gap, in contrast to S recording, which takes place in both the front and rear portions, depending on the direction of the S input. Thus, although NRZ pulses are not recorded at the gap center line, the spacing between them is proportional to the timing of the NRZ inputs; the pulse shifts do not limit the packing density as they do for S inputs.

5. Non-Zero Time Constant for S Input

Suppose now that the S input is not, as it cannot strictly be, a step function but that it takes a time T to build up or decay. Assuming that the decay or build-up is linear, then for the decay, the field around the gap at any time $t \leq T$ is:

$$H(s,t) = \frac{H_o}{\pi} \left[1 - \frac{t}{T}\right] \tan^{-1} \left[\frac{2g}{g^2 + s^2 - 1}\right] \, . \, . \, . \, . \, (19)$$

In the process of turning off, it is only the already created pulse over the rear of the gap that is affected, i.e., where $s \leq 0$. If the field is turned off fast enough, then each point of this pulse, as it moves towards the gap center line, will move not into a stronger but into a weaker field and will thus be unaffected. On the other hand, in the process of turning on, the pulse being created in the front part of the gap will always be affected by the time of build-up.

We wish now to determine the maximum decay time, T_o, for no effect on the pulse. We require that each point of the tape, after $t = o$, not increase its magnetization as it moves towards the gap center. The point requiring the quickest decay is s_o, the pulse location, at time $t = o$. This is so because there is the greatest spatial rate of change in the field at s_o when $t = o$. The magnetization slope (or pulse amplitude) at s_o is determined by substituting the value for $s_o = s_o(g)$ from Equation (5b) into the formula for P(s,g) given by Equation (3). Then:

$$P\left|s_o(g), g\right| = \frac{4M_o}{\pi} \left[\frac{s_o g}{4g^2 + (g^2 + s_o^{\,2} - 1)^2}\right] \, . \, . \, . \, . \, . \, (20)$$

If the tape velocity is v, then the maximum decay time, T_o, for this point (and therefore for all points) is determined from the formula:

$$dH(s,t) = \left[\frac{\partial H(s,t)dt}{\partial t} + \frac{\partial H(s,t)ds}{\partial s}\right]_{\substack{s = s_o \\ t = o}} = 0 \quad (21a)$$

From this we obtain:

$$P\left|s_o(g),g\right| v\, T_o = H\left|s_o(g),g\right| \frac{\ell}{2} \quad \ldots\ldots\ldots (21b)$$

Therefore, if we define a 'normalized' time,

$s_n \equiv \frac{2v\,T_o}{\ell}$, we have:

$$s_n = \frac{4g^2 + (g^2+s_o^2-1)^2}{4s_o\, g} \tan^{-1}\left[\frac{2g}{g^2 + s_o^2 - 1}\right] \quad \ldots . (22)$$

In Figure 12, s_n is plotted as a function of g, the relative spacing.

For g >> 1, we have:

$$s_n \cong \frac{4g^2 + (g^2 + s_o^2 - 1)^2}{2s_o(g^2 + s_o^2 - 1)} \cong g\frac{2}{\sqrt{3}}$$

or

$$T_o \cong \frac{y}{v}\left(\frac{2}{\sqrt{3}}\right) \quad \ldots\ldots\ldots\ldots\ldots (23a)$$

Thus, for v = 15 ips, y = 10^{-4} inches, we have:

$$T_o \cong 7\ \mu\text{sec}$$

In other words, the decay time for g >> 1 can be as much as 7 µsec without the pulse being affected in any way.

For g << 1, we obtain

$$s_n \cong \pi g$$

or

$$T_o \cong \frac{\pi y}{v} \quad \ldots\ldots\ldots\ldots (23b)$$

Thus, for the same parameter values we obtain:

$$T_o = 25\ \mu\text{sec}$$

The much larger value for g << 1, of course, is a result only of the much larger gap width.

Unlike an SZ input, a non-zero build-up time will always affect the shape of the pulse. The pulse shape is determined from the equation for the field shape:

$$H(s,t) = \frac{H_o}{\pi}\left(\frac{vt}{vT}\right) \tan^{-1}\left[\frac{2g}{g^2 + s^2 - 1}\right] \quad \ldots\ldots (24a)$$

If s' is a point fixed in the tape and s is a point fixed relative to the head, then:

$$s' = s + s_1, \text{ where } s_1 \equiv \frac{vt}{\ell/2}$$

therefore:

$$H(s,s') = \frac{H_o(s'-s)}{\pi s} \tan^{-1}\left[\frac{2g}{g^2+s^2-1}\right] \quad \ldots\ldots (24b)$$

The right-hand side of this term is ambiguous unless a further statement is made. If the argument of the inverse tangent is negative, then the angle is in the second quadrant; if positive, then the angle is in the first quadrant.

In order to determine the point of recording, s = s(s'), we set:

$$\frac{dH(s,s')}{ds} = 0 \quad \ldots\ldots\ldots\ldots\ldots (25)$$

then:

$$\frac{4gs(s' - s)}{4g^2 + (g^2+s^2-1)^2} = \tan^{-1}\left[\frac{2g}{g^2+s^2-1}\right] \quad \ldots . (26)$$

From this equation, s = s(s') can be determined and substituted back into Equation (24) to determine H = H(s'). Since the general solution of Equation (26) is not in a closed form, we consider only the limiting cases.

For g << 1, we have:

$$\frac{4gs(s' - s)}{(1 - s^2)^2} = \begin{cases} \pi & s \leq 1 \\ \dfrac{-2g}{1 - s^2} & s \geq 1 \end{cases}$$

or:

$$\begin{aligned} 4gs(s' - s) &= \pi(1 - s^2)^2 & s \leq 1 \\ 2s(s' - s) &= (s^2 - 1) & s \geq 1 \end{aligned} \quad \ldots . (27a)$$

It is to be noted that for a range of values around s = 0, the equation for s ≤ 1 cannot be satisfied, implying that no recording takes place in the region of the gap center line. This, of course, is due only to the assumption of a linear rise in time; in reality the rise starts smoothly from zero. We also notice the interesting fact that the point recorded at t = 0 (and therefore with zero magnetization) is s = 1. At later times, recording takes place at both sides of this point.

For g >> 1, we have:

$$\frac{2s(s' - s)}{4g^2 + (g^2+s^2)^2} \cong \frac{1}{g^2 + s^2} \quad \ldots\ldots\ldots\ldots (27b)$$

Here we see that the point that records at t = 0 is s = ∞. Thus, as we space out from the head, the zero time recording point moves from s = 1 to s = ∞. Also, in sine-wave recording we see that the recording point sweeps back over the gap in a direction opposite to the tape motion. As a general rule, the pulse width for a ZS recording increases by the rise time.

6. NRZ Delay

We now consider an NRZ input having a slight delay at polarity reversal so that it actually consists of two slightly separated S inputs, as shown in Figure 13a. The net magnetization is still, of course, the result of two opposite field applications. However, since there is a delay in the application of the field reversal, the two magnetizations are added with a phase difference. It is seen that the slope of the net magnetization can increase up to 50%, producing a recording of greater sensitivity that can be translated into better resolution. This is done at no increase of pulse width, in contrast to oversaturation recording. On the contrary, the trailing edge of the pulse is attenuated, which tends to decrease the pulse width, as shown in Figure 13b. Thus, by apparently increasing the 'width' of the NRZ input, the resolution can be increased, a rather surprising result. If the normalized distance the tape has moved during the reversal delay is

$$s_1 = \frac{vt}{\ell/2}$$

then we have:

$$M_n = k\left[H_1 - 2H(-s)\right], \quad s_1 \leq s \leq \infty$$

$$M_n = k\left[H(s_1 - s) - 2H(-s)\right], \quad \frac{s_1}{2} \leq s \leq s_1$$

$$M_n = -k\left[H(-s)\right], \quad 0 \leq s \leq \frac{s_1}{2} \quad \ldots\ldots (28)$$

We consider in detail only the special case of maximum pulse amplitude, i.e., where the tape displacement during the delay interval is $2s_o$. The normalized pulse over the 'three' regions is now given as:

$$P_n(s,g) = \frac{2}{\tan^{-1}\left(\frac{1}{g}\right)}\left[\frac{sg}{4g^2 + (g^2+s^2-1)^2}\right]$$

for $2s_o \leq s \leq \infty$,

$$P_n(s,g) = \frac{3}{2\tan^{-1}\left(\frac{1}{g}\right)}\left[\frac{sg}{4g^2 + (g^2+s^2-1)^2}\right]$$

for $s_o \leq s \leq 2s_o$,

$$P_n(s,g) = \frac{1}{\tan^{-1}\left(\frac{1}{g}\right)}\left[\frac{sg}{4g^2 + (g^2+s^2-1)^2}\right]$$

for $0 \leq s \leq s_o$. .... (29)

These distorted pulse curves, normalized with respect to the undistorted curves, are plotted in Figures 14 through 17 for various g's. It is seen that there is a significant decrease in pulse width and increase in asymmetry, especially for the larger g's.

With the methods of analysis applied here to time delay in NRZ recording, insight can now be easily obtained into the effect of a delay upon asymmetric recording.

7. General Switching

General switching is a generalization of S and NRZ recording. It is defined as a step function input between any two levels, as shown in Figure 18. Thus, general switching is a step function between levels H_1 and H_2, where H_1 and H_2 can be of any magnitude over or under saturation of the same or different polarity. The practical advantages of using a particular form of switching, such as NRZ, S, or any other kind, depend upon the application. In order to determine the fundamentals of pulse recording and the advantages and disadvantages of any particular form of switching, however, it is necessary to analyze switching in its most general form.

Switching can always be considered as the ordered application of two fields of the same or different polarity. It should be noted that because of the remanence the ordering is generally significant, and therefore different information is contained in different orderings. This is clearly evident in S recordings, where ZS and SZ pulses are located at different recording points.

The insight gained by considering the switching process as the ordered application of two fields is illustrated by the example of an NRZ recording where one field is at saturation and the other is below saturation. Assume, at first, that the first field is at saturation. During the field change there are then two independent S magnetizations which will interfere with each other. The pulse due to the first S will appear in the rear part of the gap and the pulse due to the second S, since it is derived from an opposite polarity, will appear in the front part of the gap, as shown in Figure 19. The interference will cause a doubling of the magnitude of the second pulse (but will not otherwise change it) and a partial erasure of the first pulse. If the recording were saturation-to-saturation, then the first pulse would be completely erased. Since the second recording is not saturation, part of the first pulse will appear. If the second recording were reduced to zero, then the location of the pulse amplitude due to the second recording would remain fixed, but the amplitude would be reduced to zero. The pulse due to the first recording will appear at the rear of the gap, a distance s_o from the gap centerline, and will remain fixed there as its amplitude grows to maximum. Thus, a single input switching, saturation-to-below saturation, can locate the gap centerline in the same manner as an RZ input does.

In Figure 20 the same process is illustrated with the order of application reversed. It is seen that only one pulse appears.

Even if the two applied fields are of the same polarity, they can be viewed as two independent S recordings of the same polarity interfering with each other. Depending on the ordering and the relationship of the amplitudes, two pulses can appear quite often.

8. Recording of Consecutive NRZ Pulses

In this section we indicate some of the more significant results of the analysis applied to consecutive NRZ inputs. The complete wave shape of interfering pulses is not derived, although the method is clearly indicated.

Assume a second NRZ input at time t after the first NRZ input. In this time t, the tape has moved a relative distance $s_1 = \frac{vt}{\ell/2}$. At the location of the first NRZ pulse, the second NRZ input will produce a field insufficient to affect the magnetization there. Therefore, the location and amplitude of a given NRZ pulse are never affected by subsequent NRZ inputs, no matter how small t is. Furthermore, the location and amplitude of the second NRZ pulse are unaffected by the presence of the first NRZ magnetization. In fact, the only region where the pulse wave shape is at all affected is between the pulses. The following statements can be made about the wave shape between the two pulses:

For $s_1 \leq s_c$ (where s_c is the magnetization crossover point for a single NRZ pulse), the wave shape of the first pulse is unaffected in the region s_2 to ∞ ; the wave shape of the second pulse is unaffected in the region $-\infty$ to s_2, s_2 being the point where the magnetizations from both pulses are the same and can be determined from the equation:

$$(s_2-s_1)^4 (g^2+s^2-1) + (s_2-s_1)^2 \left[(g^2+s_1^2-1)(g^2-g-4) - (g^2-4) \right] - (g^4+g^3-2g^2+1) = 0 \quad \ldots\ldots\ (30)$$

This equation is derived from the mathematical statement of the equality of the magnetizations of the two pulses:

$$\tan^{-1}\left[\frac{2g}{g^2+s^2-1}\right] = 2\left[\tan^{-1}\left(\frac{1}{g}\right) - \tan^{-1}\left(\frac{2g}{g^2+(s_2-s_1)^2-1}\right)\right] \quad \ldots (31)$$

The point s_2 can be seen graphically in Figure 21; however, the wave shape is discontinuous at s_2.

For $s_1 > s_c$, the shape of the first pulse remains unaffected in the same region, i.e., s_2 to ∞, as well as in the region $-\infty$ to the point s_c of the second pulse. However, in the region s_c to s_2, the wave shape is a varying superposition of the separate pulse wave shapes.

Notations

S	-	Saturation recording
SZ	-	Saturation to zero input switching
ZS	-	Zero to saturation input switching
NRZ	-	Saturation to saturation recording
RZ	-	SZ followed or preceded by ZS
x	-	Head coordinate in direction of tape motion
ℓ	-	Gapwidth
g	-	$\frac{y}{\ell/2}$, relative distance from head surface
s	-	$\frac{x}{\ell/2}$, relative distance from gap centerline
s_o	-	Pulse maximum location
s_c	-	Tape magnetization crossover point
s_α	-	Location of dividing point between saturated and unsaturated tape
H	-	Head field
H_o	-	Head field at point (0,0)
H_s	-	Head field required to saturate tape
k	-	Tape susceptibility
M	-	Tape remanent magnetization
P	-	$\frac{dM}{ds}$, the recorded pulse
P_n	-	$\frac{P}{H(0,g)}$, the normalized pulse amplitude
α	-	$\frac{H}{H_s}$, oversaturation factor
v	-	Tape velocity
T	-	Switching decay or build-up time
T_o	-	Maximum switching decay time
s_n	-	$\frac{vT_o}{\ell/2}$, 'normalized' decay time
s_1	-	$\frac{vt}{\ell/2}$
s'	-	$s + s_1$, relative coordinate fixed in tape

References

1. Stein, I., "Analysis of the Recording of Non-Biased Sine Waves", Research Report No. 121, Ampex Corp. Also published externally as "Analysis of the Recording of Sine Waves", IRE Transactions on Audio, Vol AU-9, No. 5, Sep - Oct 1961, pp 146 - 154.

2. Schwantke, G., "Beitrag Zur Darstellung Des Spaltfeldes Beim Magnetton", Acustica, Vol 7, No. 6, 1957.

3. Eldridge, D. F., "Magnetic Recording and Reproduction of Pulses", Research Report No. 112, Ampex Corp. Also published externally under the same title in IRE Transactions on Audio, Vol AU-8, No. 2, Mar - Apr 1960, pp 45 - 47.

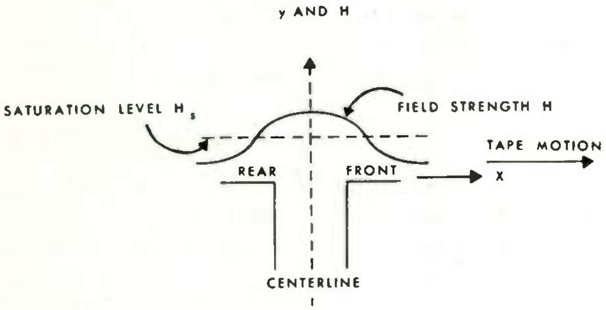
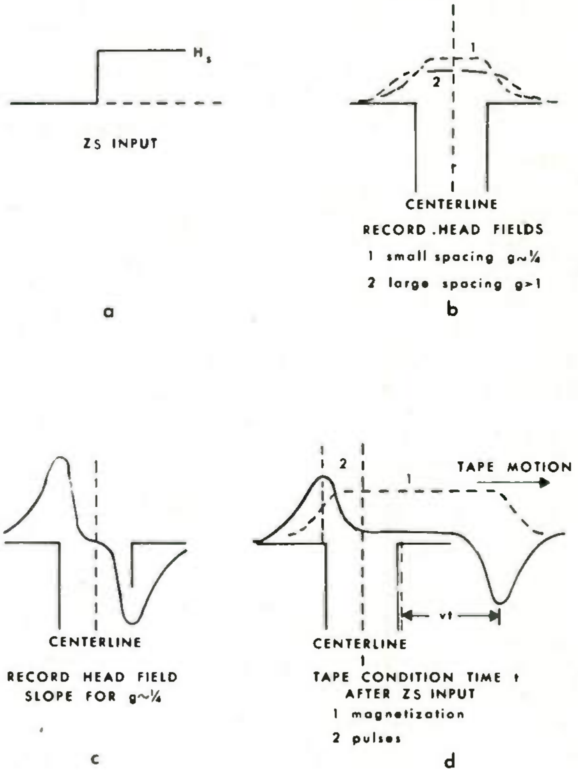


FIGURE 1 HEAD TAPE SYSTEM



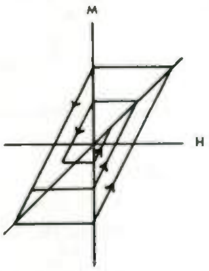


FIGURE 2 SIMPLIFIED HYSTERESIS CURVE

FIGURE 3 ZS RECORDING

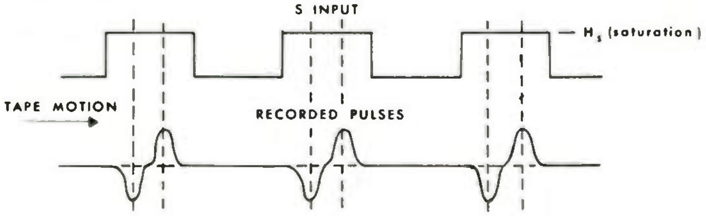


FIGURE 4 LOCATION OF RECORDED PULSES RELATIVE TO S INPUT

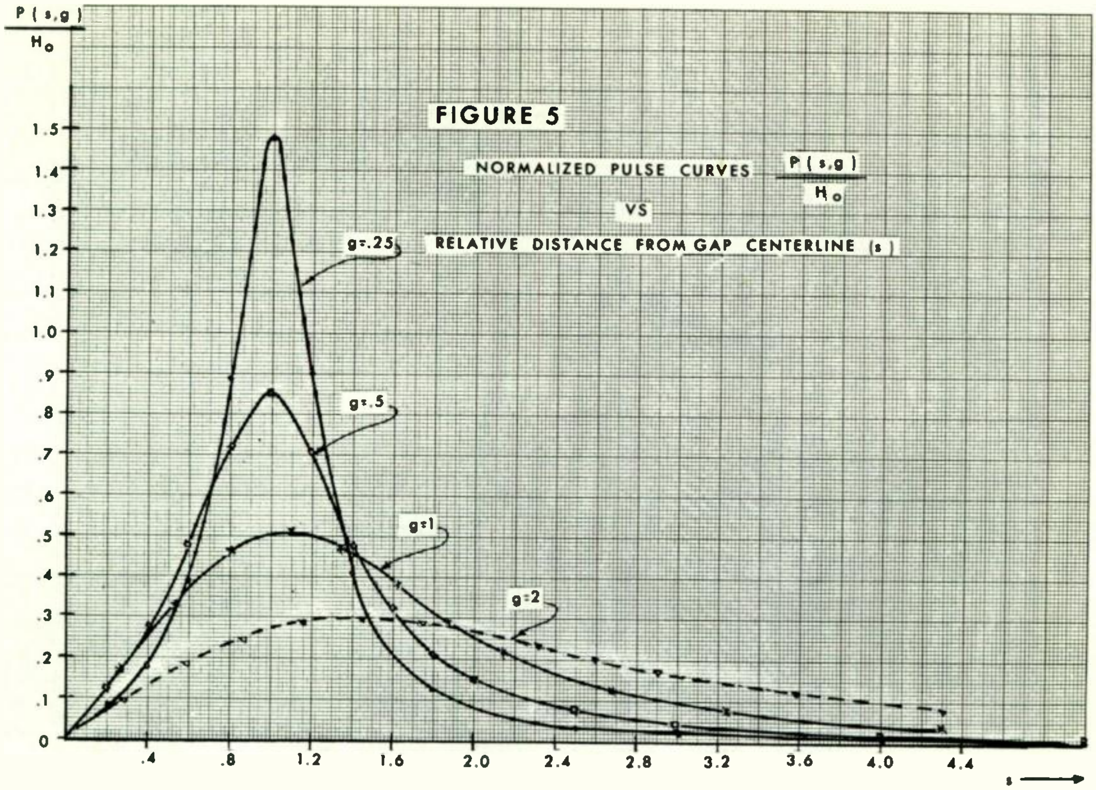


FIGURE 5

NORMALIZED PULSE CURVES $\frac{P(s,g)}{H_o}$

VS

RELATIVE DISTANCE FROM GAP CENTERLINE (s)

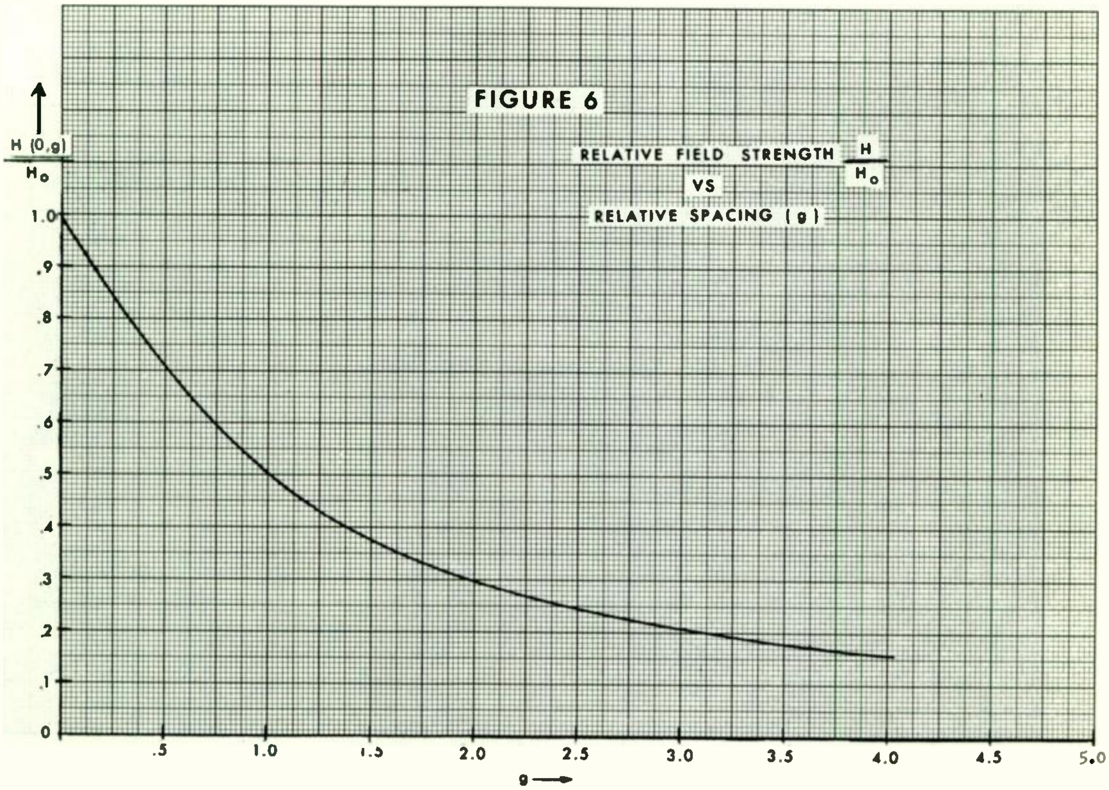


FIGURE 6

RELATIVE FIELD STRENGTH $\frac{H}{H_o}$

VS

RELATIVE SPACING (g)

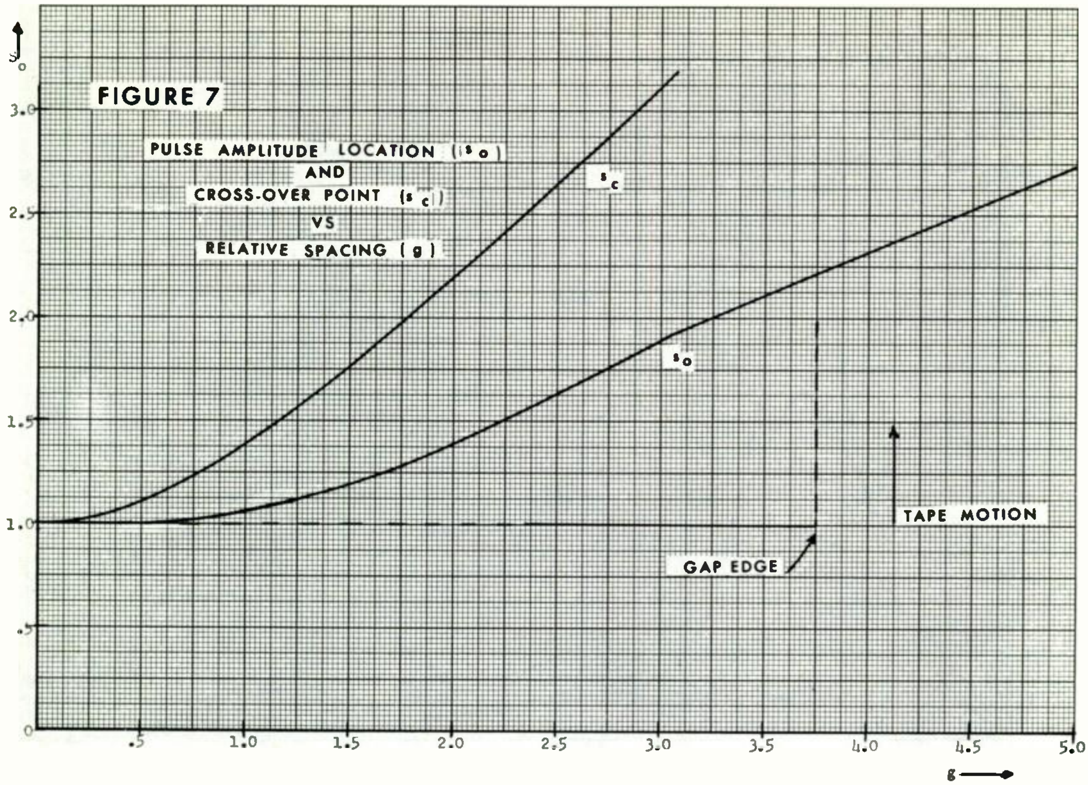


FIGURE 7

PULSE AMPLITUDE LOCATION (s_o) AND CROSS-OVER POINT (s_c) VS RELATIVE SPACING (g)

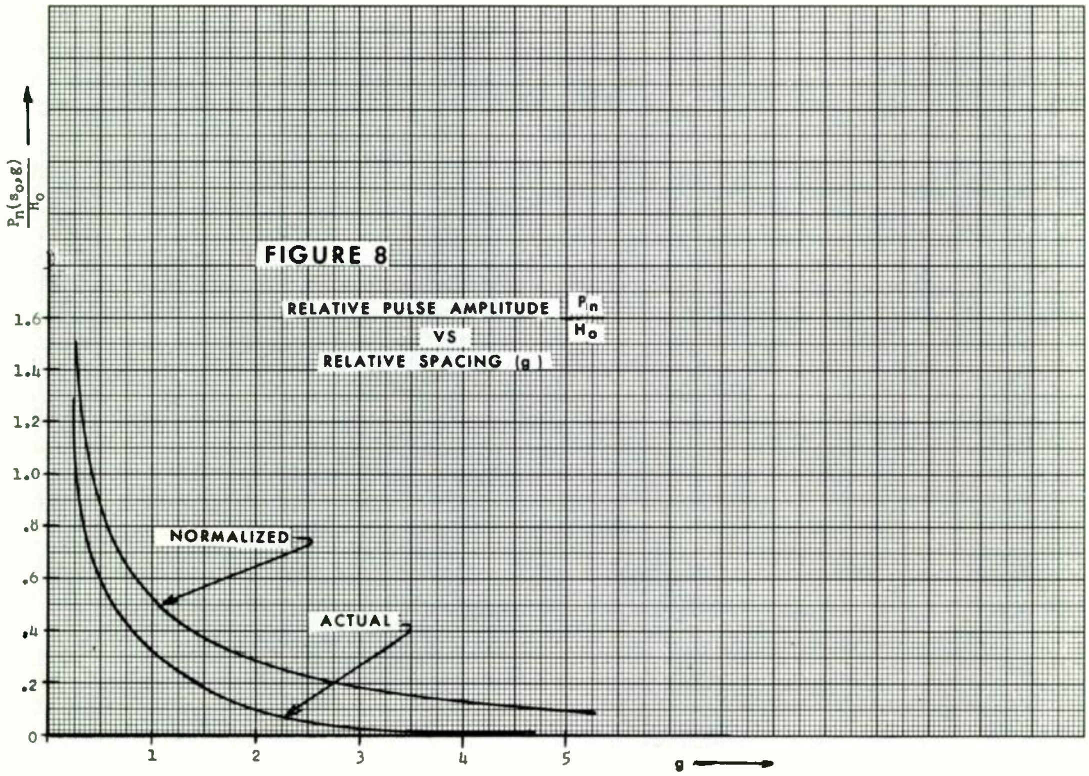


FIGURE 8

RELATIVE PULSE AMPLITUDE $\frac{P_n}{H_o}$ VS RELATIVE SPACING (g)

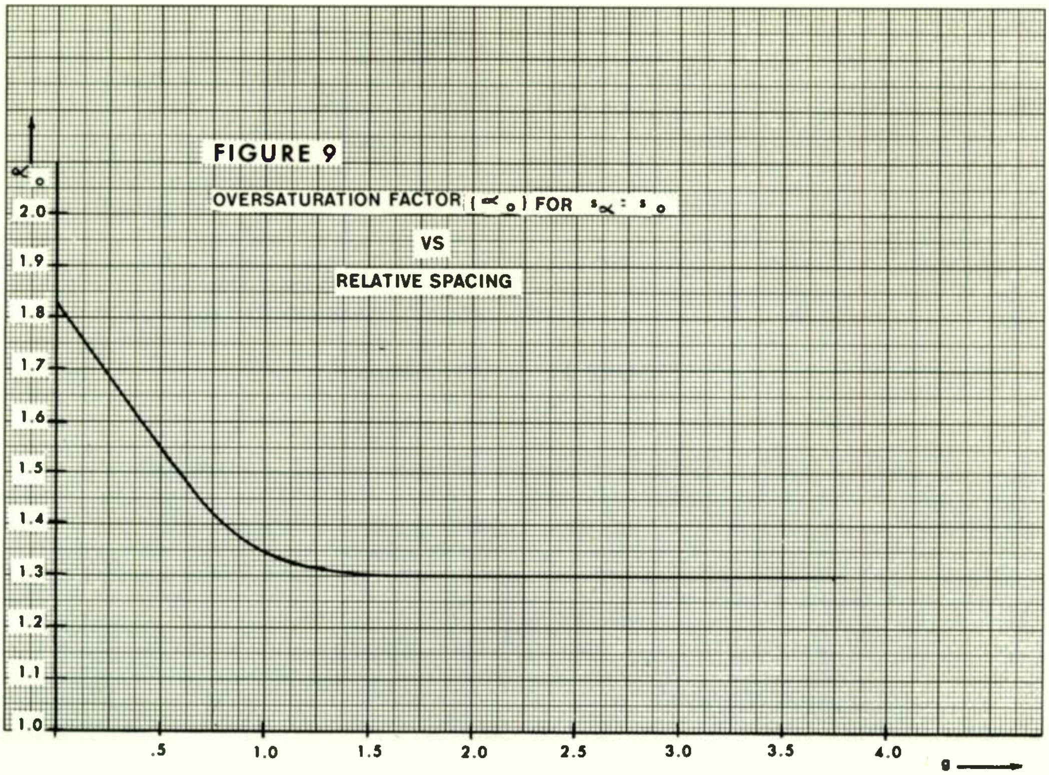
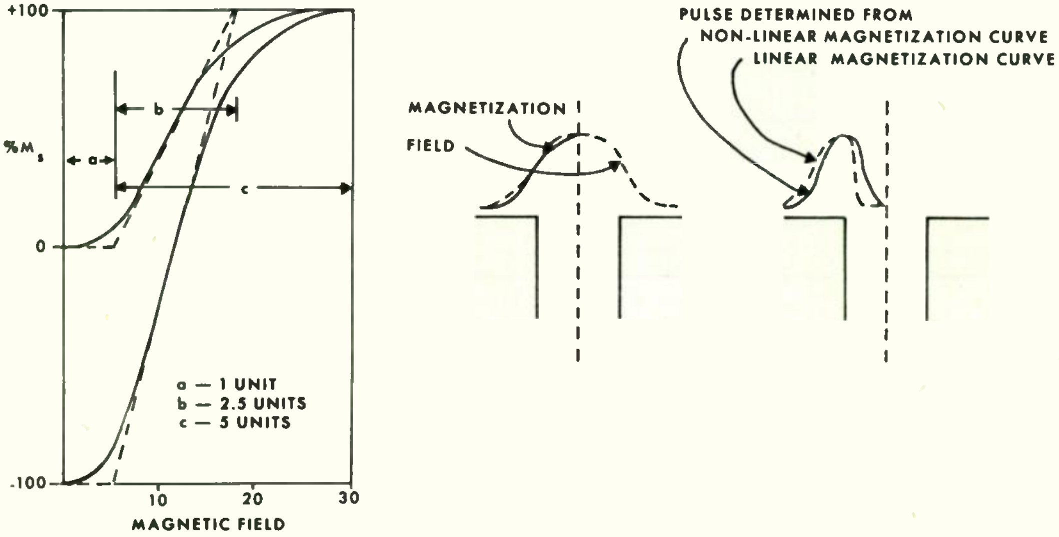


FIGURE 9

OVERSATURATION FACTOR (α_o) FOR $s_\alpha : s_o$

VS

RELATIVE SPACING

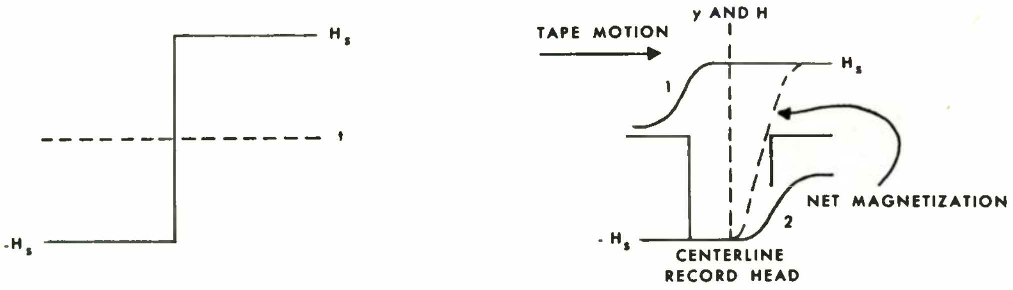


a $\gamma - Fe_2O_3$ CURVES

b TAPE MAGNETIZATION

c PULSE-ON-TAPE EFFECT ON NON-LINEAR MAGNETIZATION CURVE

FIGURE 10



a NRZ INPUT

b TAPE MAGNETIZATION DUE TO NRZ INPUT

FIGURE 11 NRZ RECORDING

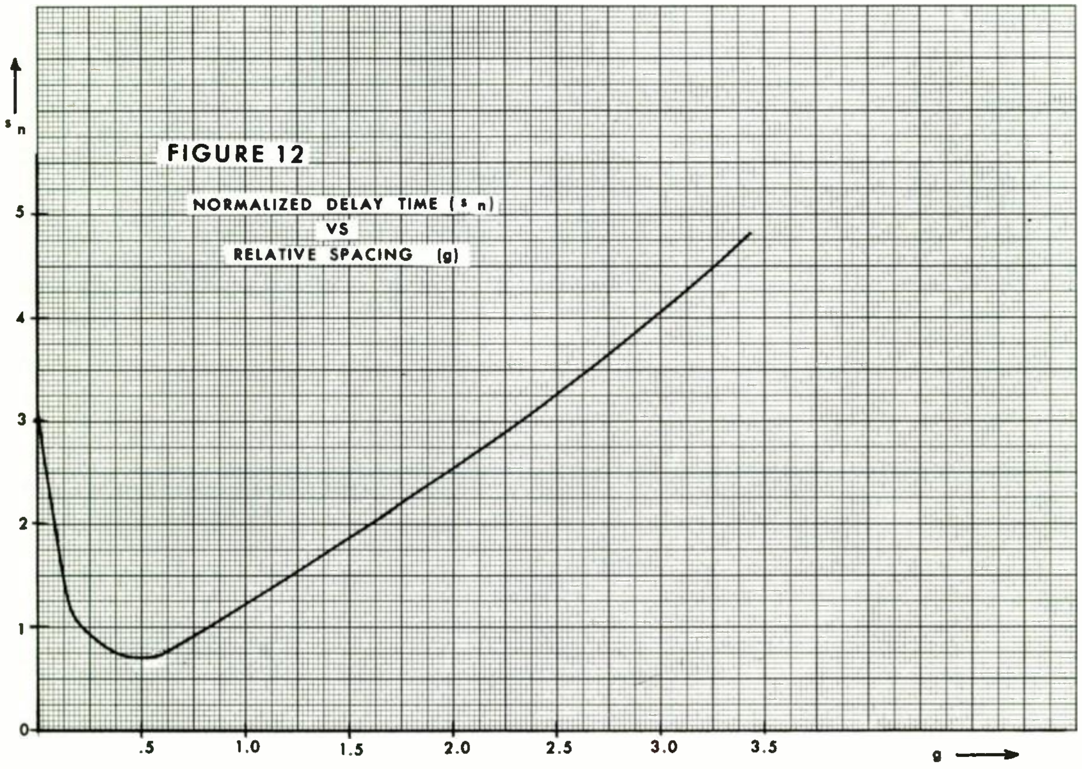
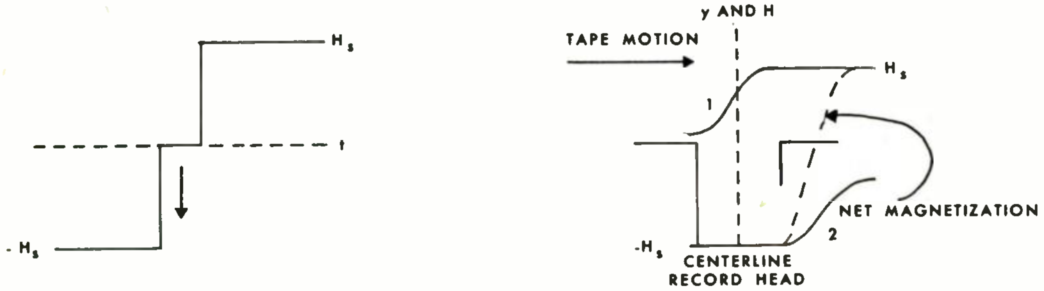


FIGURE 12

NORMALIZED DELAY TIME (sn)
VS
RELATIVE SPACING (g)



a DISTORTED NRZ INPUT

b TAPE MAGNETIZATION DUE TO DISTORTED NRZ INPUT

FIGURE 13 NRZ DELAY

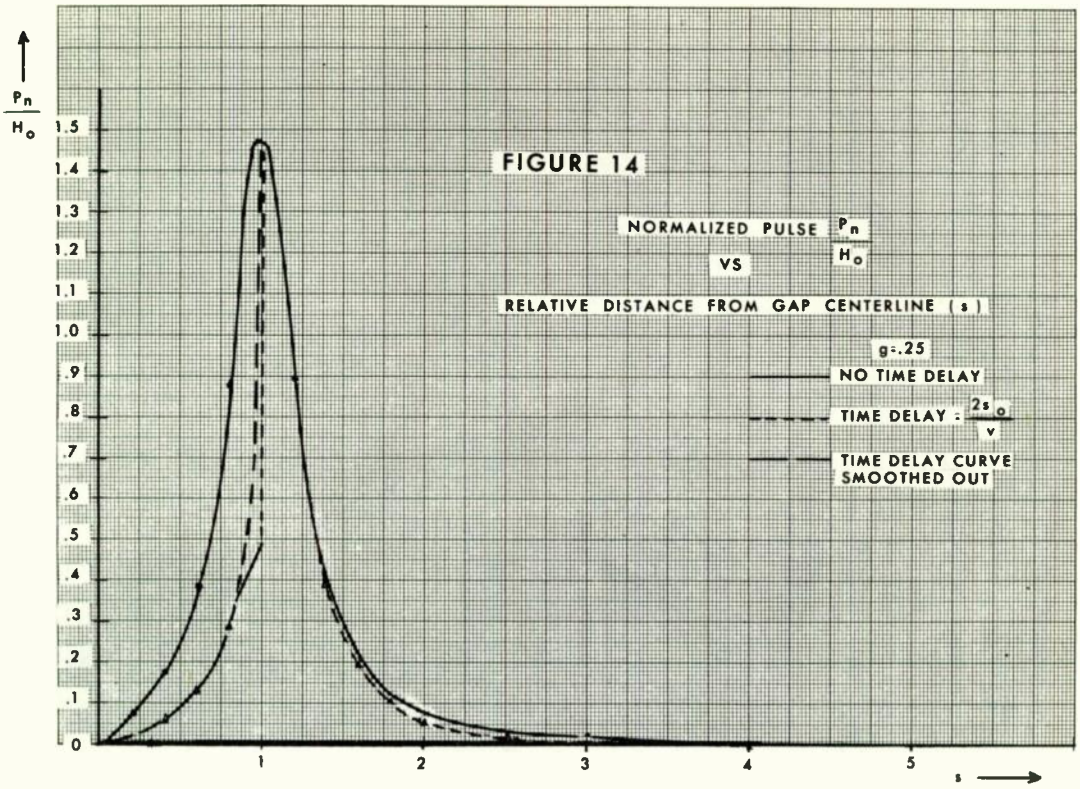


FIGURE 14

NORMALIZED PULSE $\frac{P_n}{H_o}$ VS RELATIVE DISTANCE FROM GAP CENTERLINE (s)

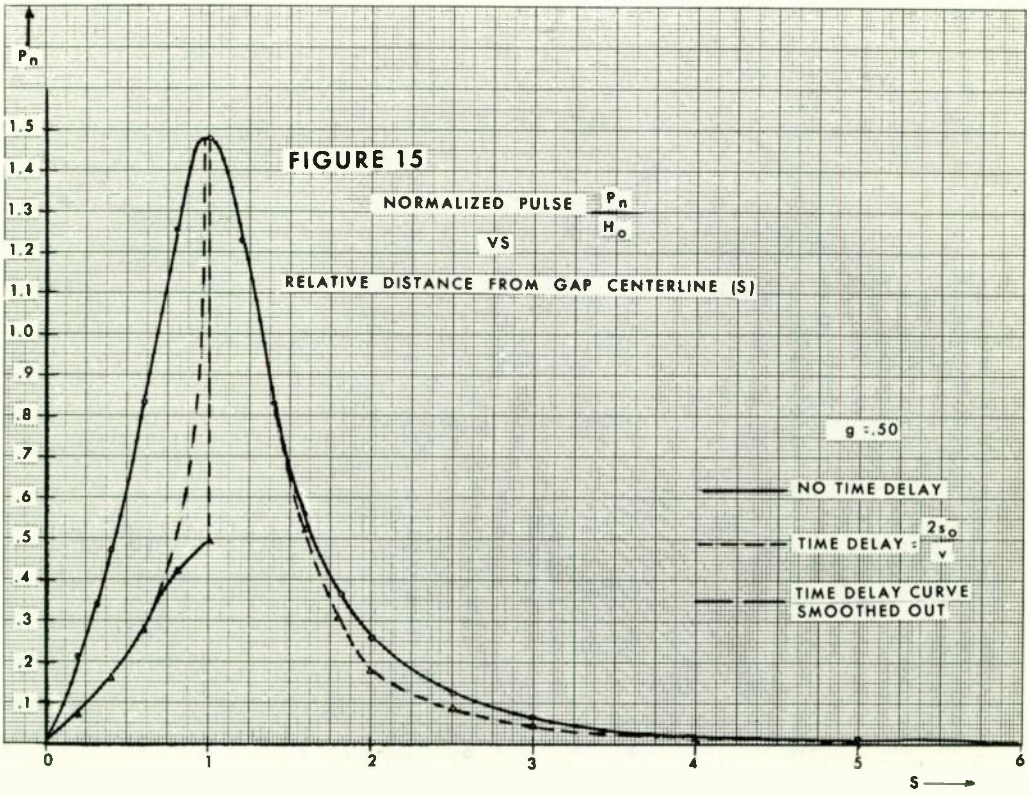


FIGURE 15

NORMALIZED PULSE $\frac{P_n}{H_o}$ VS RELATIVE DISTANCE FROM GAP CENTERLINE (S)

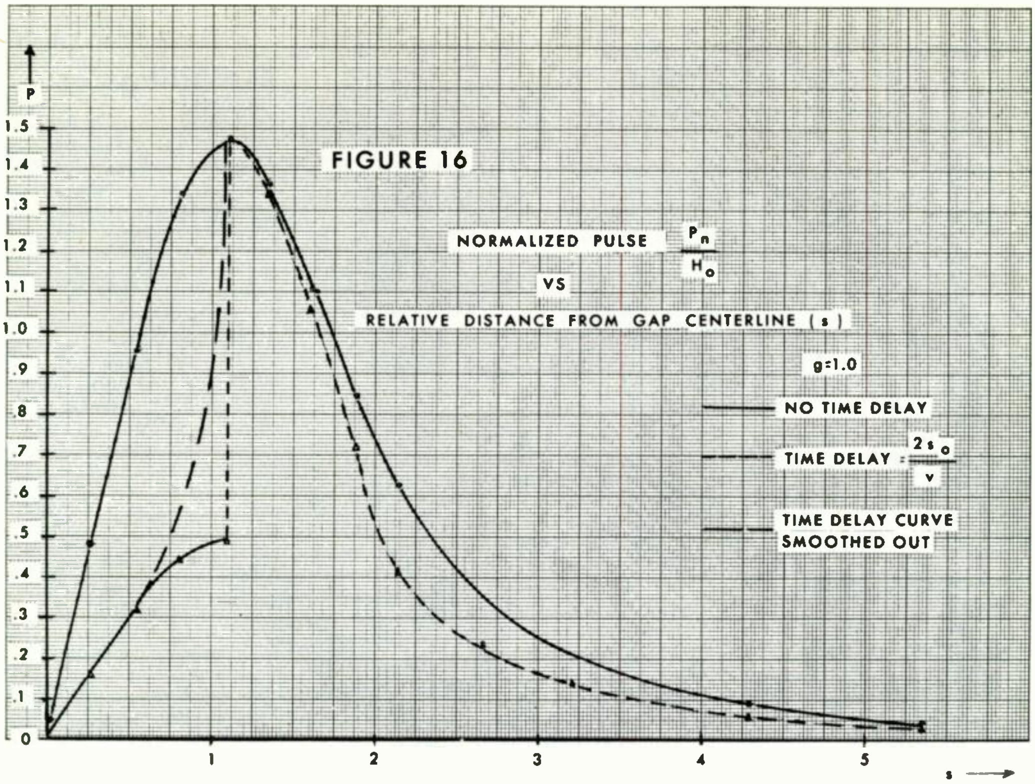


FIGURE 16

NORMALIZED PULSE $\frac{P_n}{H_o}$ VS RELATIVE DISTANCE FROM GAP CENTERLINE (s)

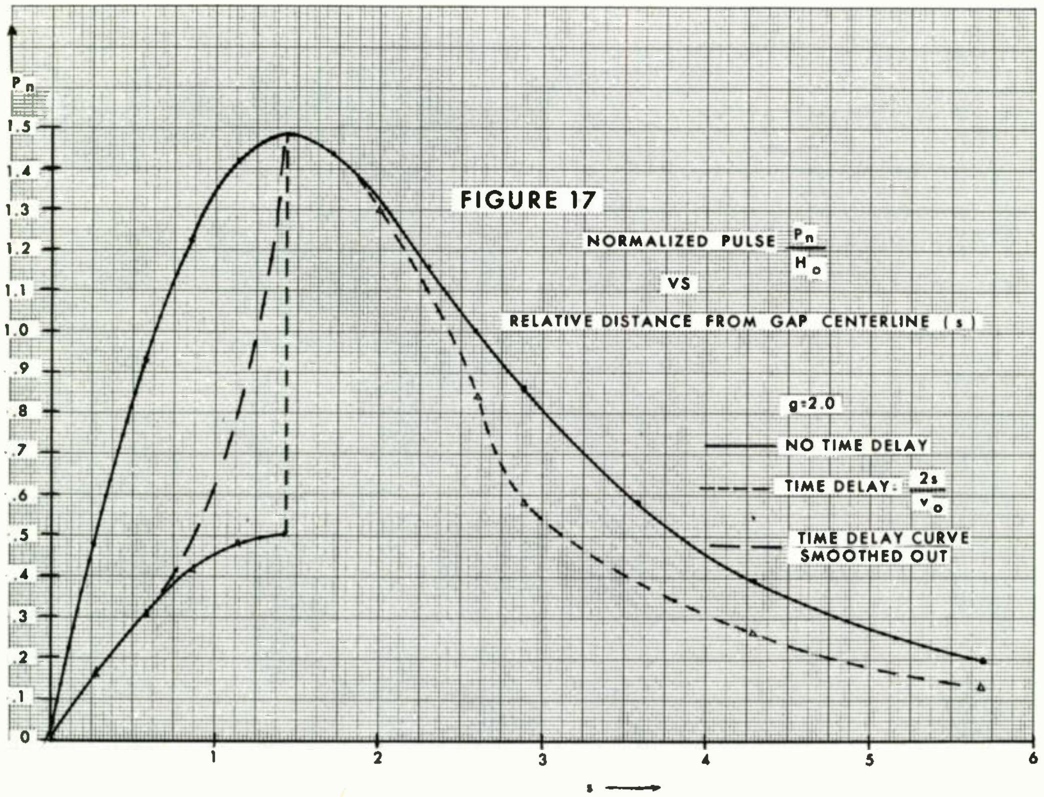


FIGURE 17

NORMALIZED PULSE $\frac{P_n}{H_o}$ VS RELATIVE DISTANCE FROM GAP CENTERLINE (s)

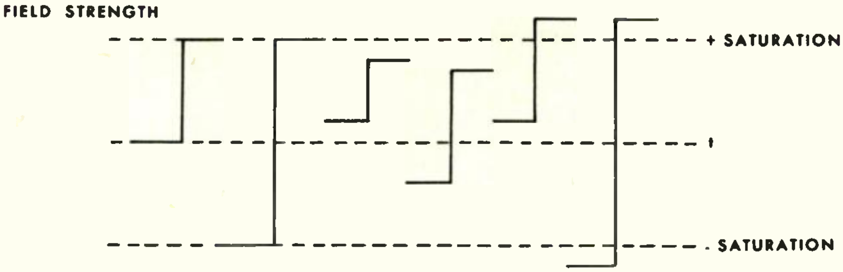


FIGURE 18 SOME DIFFERENT FORMS OF GENERAL SWITCHING

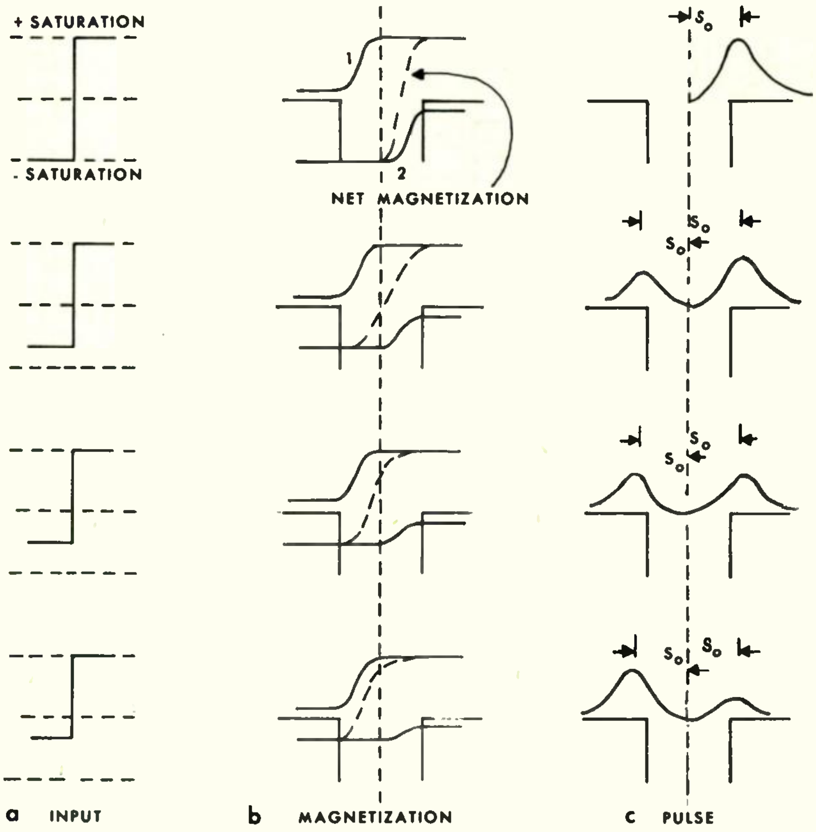


FIGURE 19 VARIABLE N R Z RECORDING

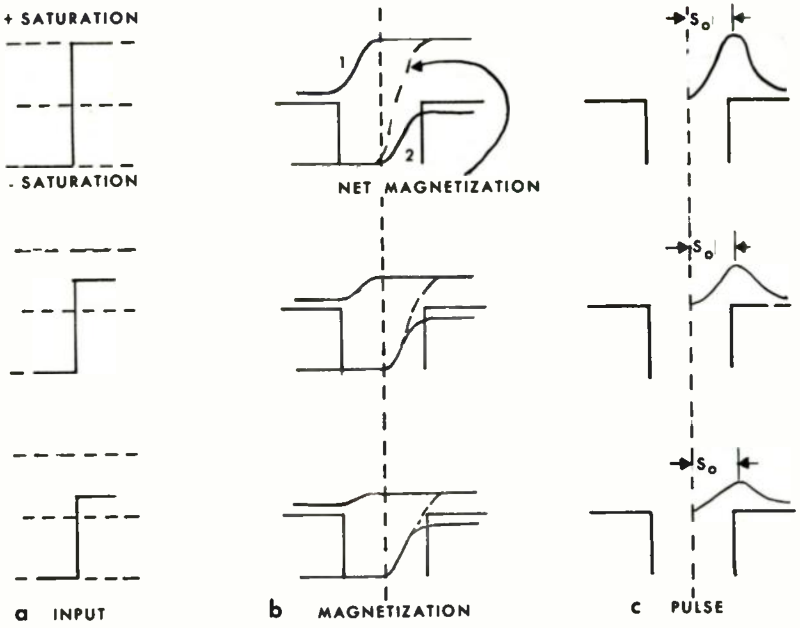


FIGURE 20 VARIABLE N R Z RECORDING

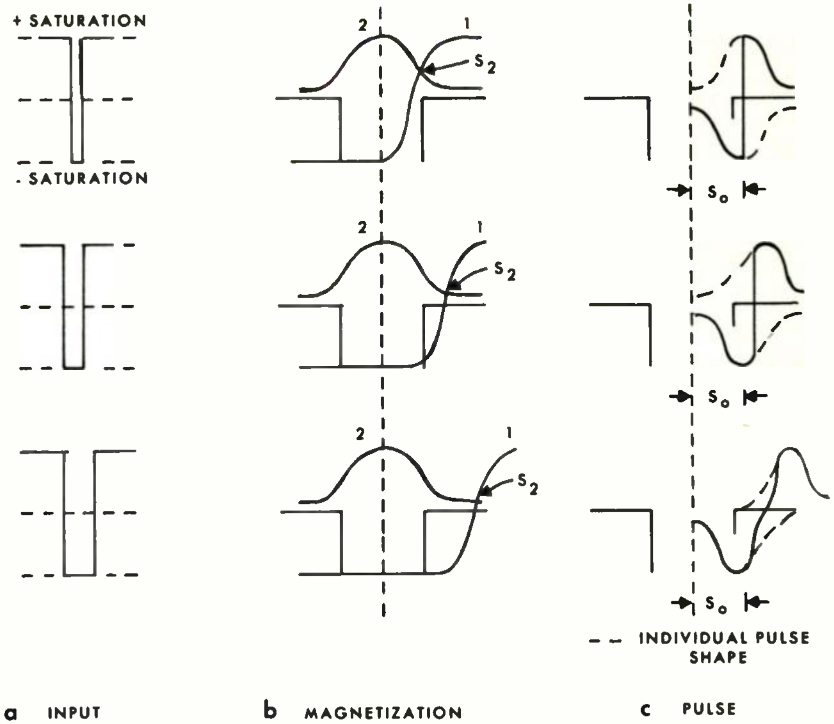


FIGURE 21 CONSECUTIVE PULSE RECORDING

HIGH-DENSITY MAGNETIC HEAD DESIGN FOR NONCONTACT RECORDING

Lester F. Shew
General Products Division
IBM Development Laboratory
San Jose, California

Summary -- The information storage density in digital magnetic recordihg is dependent on both the pulse resolution and the track definition. This paper is concerned with these two factors in the design of magnetic heads for noncontact recording.[1] A concept of changed pole-tip geometry which led to a significant improvement of pulse resolution is introduced. A general expression based on "single-pulse" superposition is derived for various bit densities and data codes. In addition, several recording methods are discussed for achieving near-maximum track density under various head-repositioning error conditions. As shown, high-density heads for noncontact recording have been designed successfully by applying the concept and techniques developed. Good correlation has been realized between analytical and experimental results. Performance characteristics under simulated machine conditions are presented.

INTRODUCTION

One of the most fundamental motivations in magnetic recording as applied to information storage is to achieve increasingly higher storage densities. Since the storage density per unit area in digital recording is the product of the longitudinal bit density and of the transverse track density, the information storage potential is dependent on both the pulse resolution and the track definition. These two factors are of primary concern in the design of high-density magnetic heads. In contact recording, magnetic heads for digital recording have been demonstrated, within the laboratory environment, for a pulse resolution of up to 2,000 BPI (bits per inch) and a track definition of 500 TPI (tracks per inch).[2]

In noncontact recording, techniques are not available for achieving densities higher than 1,000 BPI and 100 TPI,[3] with satisfactory system operation. This paper discusses the approaches to these two major problems. To attain higher pulse resolution, the concept of controlling the writing field distribution is applied. To achieve better track definition, an intertrack shield and a well-chosen recording method for a given head-repositioning error are suggested.

The semi-infinite pole shape is most commonly used in a longitudinal-recording head (ring-type structure). This form of pole face operates quite satisfactorily on tapes, and attains high bit density in contact recording. In noncontact recording, however, its resolution suffers seriously as a result of the increased head-to-medium (H-to-M) separation. That is, the field gradient is steep near the pole face but becomes more gradual as the distance from the pole face increases. The field distribution of the write head and the sensitivity function or weighting function (a function giving the sensitivity of the reading coil to the surface magnetization) of the read head can be controlled by changing the geometry of the pole tip. Heads based on this concept of control through changed pole tip geometry have been designed, and significant improvement in pulse resolution has been achieved.

When the bit density is higher than a figure equal to the reciprocal of the pulse width (transition region), the readback pulse form changes due to partial modification by the adjacent pulse. An equation based on "single-pulse" superposition has been derived for predicting the readback signal for various bit densities and data codes.

Mechanical limitations and tolerances under actual machine conditions make head repositioning a major problem in magnetic recording. When a read head is displaced from a given written track, two undesirable effects normally result. One is the drop in the readback signal amplitude; the other is the noise picked up from residual magnetism outside the recorded track and from crosstalk from adjacent tracks. This paper presents several methods commonly employed to achieve near-maximum track density and to minimize the undesirable effects; it also discusses the design analysis and consideration of these methods.

No attempt is made to discuss the frequency-dependent factors, such as the head frequency response, read amplifier response, write current rise and fall times, etc. The primary concern in this study is the pulse widening caused by the broad field distribution in noncontact recording. Track density limitations that may arise from interhead crosstalk in a multiple-head stack are not considered, but stress is placed instead upon the limitation caused by head repositioning errors.

By applying the theory and techniques developed in connection with the IBM magnetic disk storage development program, practical high-density heads for noncontact recording have been successfully designed. This paper includes test results and their correlations with the theory, and presents performance characteristics of several high-density heads.

DESIGN ANALYSIS AND CONSIDERATIONS

Coding Methods

Many coding methods of recording of information on a magnetic medium have been proposed. The most widely used techniques are the RZ (return-to-zero), NRZ (non-return-to-zero), and PM (phase modulation). Each method has its advantages and disadvantages. Various modifications and different detection techniques have been successfully employed to increase information density. However, the bit density achievable by any method is determined by the basic resolution of the written transition of magnetization reversal and the steepness of the head-sensitivity function. Thus the maximum packing density is a function of the actual width and quality of an isolated readback pulse. The NRZI (non-return-to-zero, IBM code) method of recording was used throughout the study presented in this paper, and an isolated NRZI pulse was employed for analyzing the factors which influence its width and quality in noncontact recording.

Head-to-Medium Separation

Besides meeting the electrical circuit requirements on information storage density, bit rate, signal amplitude, noise level, phase shift, and track density, the magnetic head for a given application must function reliably within the mechanical tolerances imposed by the system. These tolerances include possible variation in head-to-medium separation, change in relative head-medium motion, head repositioning error, physical size, etc.

The major design problems of the heads for high information storage density are good pulse resolution and fine track definition. These problems become more difficult with increasing head-to-medium separation.

Pulse resolution is a function of head-to-medium separation. The greater the distance from the head gap, the more gradual the field gradient becomes. Consequently, if the recording medium is excessively distant from the write head, it is subjected to a field with a much decreased gradient. Fig. 1 shows the field distribution near the surface and at some distance above the pole face. The more gradual this gradient gets, the longer becomes the transition region of magnetization reversal along the recording medium. This effect will result in poor pulse resolution.

Fringing of the write field causes the width of the written track to be somewhat larger than the physical width of the write pole tip. In saturation-type recording, the written track width becomes wider for thicker recording media and for larger head-to-medium separations. However, the amount of increase in the written track width caused by the fringing field is independent of the physical size of the pole tip width; for this reason the increased track width from fringing is a larger percentage of the total track width for the narrower tracks.[2]

Write Head Geometry

The conventional pole tip geometry of the head used for longitudinal recording is usually semi-infinite. This shape of pole face is quite satisfactory for contact recording. For noncontact recording, however, the bit density that can be achieved is limited. The write field of the head is assumed to be composed of longitudinal and perpendicular (horizontal and vertical) components,[4] (Fig. 1). The magnitude of this field at a given point is of importance, as it determines the remanence of magnetization of the particles in the recording medium. However, the trailing field is of utmost concern, because the final effect upon the medium is essentially that exerted by the trailing pole, since the influence of the leading pole will be modified by the trailing field.

When a particle passes the gap, it is magnetized in one polarity. If the write field changes its polarity before this particle has passed completely out of the influence of the reversed field, it is partially magnetized in a reversed polarity to a varying degree dependent on the gradient across the head.[5] This condition results in partially demagnetizing the recorded medium which is still within the field of the head as it changes polarity. Consequently, the pulse resolution is limited in noncontact recording, by a semi-infinite pole face, where the field gradient is very gradual at some distance from the head. To attain high pulse resolution, the write trailing field effect must be eliminated or reduced to a very minimum. This calls for a steep gradient at the trailing edge of the gap. Fig. 1 shows that the field gradient is very poor at some distance above the pole face of a semi-infinite pole tip head.

Accurate presentation of the field pattern is extremely difficult because of the nonlinear relationship and asymmetrical nature of the magnetic cycle. The models given here are highly simplified. However, they do explain qualitatively the effects of the pole-tip geometry in the writing and reading processes.

This paper does not consider the effects of the head gap length, of recording medium characteristics, and of thickness; these parameters have already been well treated by previous papers.[5,7] The primary interest in this study is to investigate the effect of the pole-tip configuration of a ring-type-structure head on the readback pulse resolution in longitudinal recording.

The field distribution of the head can be controlled to a large extent by changing the geometry of the pole tip. Figure 2 shows the pulse width (at the level that is 10% of the peak) as a function of the pole-tip length (L). Data was obtained experimentally with the same head for both writing and reading. That is, the pole-tip length was increased by lapping the head surface between each set of write-read data taken. Three different head-medium relative velocities were taken, but the same head-to-medium separation was maintained. Figure 3 shows the pulse width as a function of the angle θ which the edge of the head pole tip makes with the recording surface. The same head (with its angle θ changed after each step, by machining the pole-tip edges) was used in taking the data throughout the test, and the pole-tip length was maintained at approximately one milli-inch. These curves indicate that, in noncontact recording, the pulse width widens as L increases and θ decreases. The pole-tip geometry approaches closer to semi-infinite as θ becomes smaller. Consequently, decreasing θ has a similar effect to that caused by increasing L. In writing, as L increases and/or θ decreases, the trailing field effect increases, resulting in a wider transition region of magnetization reversal. In reading, as L increases and/or θ decreases, the head sensitivity function becomes more gradual, which results in a wider readback pulse.

Read Head Geometry

Since for most practical purposes the readback pulse form can be assumed to resemble the field distribution function of the head,[6] it is apparent from Fig. 1 that the resolution in the reading process becomes poor when the head-to-medium separation gets large.

The oscillograms in Fig. 4 show the effects of the pole-tip length on both the writing and reading processes. The pulse width at the 10% (of peak) level is: in (a) 3.0 microseconds, in (b) 5.7 microseconds, in (c) 10.5 microseconds and (d) 6.0 microseconds. Both heads L-1 and L-8 are of the same design and have the same gap length. The only difference is in the pole tip length. (Head L-1 = 1.0 milli-inch, and L-8 = 8.4 milli-inches). They were tested on the same recording medium (0.43 milli-inch γ $Fe_2 0_3$), at the same velocity (1000 IPS) and at the same head-to-medium separation (0.25 milli-inch). The sensitivity function of head L-1 should be the same in scanning over either transition region of magnetization reversal: That written by head L-1 and that written by head L-8. The pulse- width difference in (a) and (d) is primarily due to the difference in the widths of these two transition regions. By comparing the pulse widths in (a) and (d), it is apparent that in noncontact recording a wider transition region is written by a head with larger L. The transition region after being written, remains practically the same width regardless of which head is used for readback. Therefore, the pulse width difference shown between (a) and (b) and that shown between (c) and (d) must be due to the difference of the sensitivity functions of the readback heads. These differences indicate that a head with larger L has a broader sensitivity function. This data is in good agreement with the analysis on the effects caused by the pole tip geometry.

Based on the principles of reciprocity[6], the readback process is analyzed, and the voltage e(x) produced in the reading coil is approximated by a convolution integral of the form:

$$e(x_1)=vN \frac{d\phi}{dx_1} = KvN \int_{-\infty}^{+\infty} H(x) \frac{\partial M(x-x_1)}{\partial x_1} dx \qquad (1)$$

where $x_1 = vt$ (t=time), ϕ is the flux in the reading coil, K is a constant, v is the relative head-medium velocity, N is the number of turns in the readback coil, H(x) is the sensitivity function of the readback head, and $M(x-x_1)$ is the distribution of magnetization in the recording medium.

In longitudinal recording by a ring-type-structure head, it is reasonable to assume that the perpendicular component of the magnetization is negligible[7] and that the longitudinal component in the x direction is the only significant component of the magnetization. This means that My = 0 and Mx = M. For an idealized case in saturation recording, M_x (x) can be considered as a step change of magnetization from $-M_s$ to

$+M_s$, as shown in Fig. 5. The induced voltage due to the step change of M(x) at x_1 is

$$e_1 = 2\ KvNM_sH_x\ (x_1) \qquad (2)$$

where $H_x(x)$ is the sensitivity function of the readback head which contributes to the longitudinal component of the magnetization.

According to (2), the readback pulse form, e(x), is proportional to the sensitivity function, $H_x(x)$. Since the field distribution of the head is a measure of its sensitivity function, e(x) can be considered, for most practical cases, to closely resemble the field distribution gradient.

The induced voltage due to the step change of M(x) of opposite polarity at x_1+d can be written as

$$e_2 = -2\ KvNM_sH_x\ (x_1+d) \qquad (3)$$

In NRZI coding, writing two "1's" would result in an approximate rectangular-function distribution of magnetization. If the space interval between the two step functions of opposite signs is d, the rectangular function of magnetization can be shown as in Fig. 5. This rectangular function is composed of a positive step at x_1, followed by a negative step at (x_1+d).

The induced voltage can be found by adding (2) and (3):

$$e(x) = e_1+e_2 = 2KvNM_s\ \left[H_x(x_1)-H_x(x_1+d)\right] \qquad (4)$$

Eq. (4) may be interpreted that the pulse form of the induced voltage at a given position relative to the read gap centerline is proportional to the relative head-medium velocity and the surface magnetization. The pulse form of the induced yoltage also corresponds in time to the sensitivity function, H_x, at x_1 and at x_1+d. That is, in addition to the application of the principle of reciprocity, the principle of superposition is extended to the analysis of the readback pulse.

Each character is represented by one or more transitions of magnetization reversal spaced at various intervals, d, which varies according to the bit density and data code. The readback pulse form and the signal-to-noise ratio varies with the bit density and the data code. Therefore, to achieve high bit density and good signal-to-noise ratio, noise in the baseline and asymmetry on the pulse form should be eliminated or reduced to a very minimum.

The resultant signal and noise can be analyzed and predicted from a single isolated pulse. Fig. 6 (a) shows two isolated pulses (e_s), each of which has a preceding noise spike (e_n). When the interval, d, between the two data pulses gets smaller, the noise spike is superposed onto the data pulse at a spacing which is equal to that between the data pulse and the noise spike, giving a distorted pulse form as shown in Fig. 6(b) and (c). These oscillograms exhibit clearly the phenomenon of pulse superposition and give good qualitative confirmation of the foregoing analysis (Eq. (1) through (4)).

Recording Methods

For a given system, the track density achievable depends largely upon the head-repositioning accuracy, a problem of particular importance in "random-access"-type memories. Tight mechanical tolerances and special techniques, such as automatic servo control, have been developed to minimize the head-repositioning error. The discussion of these techniques is beyond the scope of this paper. Because the head-repositioning accuracy has such important effects on the achievable track density, it cannot be neglected in the design consideration of the magnetic heads. Since mechanical tolerances always exist in any system, techniques in magnetic recording itself must be developed to attain maximum track density for a given head-repositioning error on a given system. A good comprehension of the overall system requirement and a knowledge of the mechanical tolerances imposed on the head are therefore required.

The four parameters of the head output that are of primary interest concerning reliability are the readback pulse amplitude, pulse width, peak shift, and noise level. When the detection-circuit tolerances in respect to these major parameters are known, a nearly optimum head design can be achieved, so that satisfactory and reliable performance is attained. The proper recording method for a given system is a prime consideration in attaining this performance.

Three methods of recording commonly used in digital computers are:

(1) Write and read on same track width.

(2) Erase-wide and read/write-narrow.

(3) Write-wide and read-narrow, without erasing.

Method (1) requires only one magnetic element, and the previously written data is not erased before writing. This method can be economically accomplished with one common pole tip for both writing and reading. It is generally used in tape and drum applications where the off-track head displacement is small. However, merely writing over old information without first erasing can result in several undesirable characteristics. Under certain conditions this method of recording can produce phase shift if the intensity of the write field is insufficient to completely reverse the previous magnetization or to overcome the distortion that may be caused by the previously written data. A write current equal to twice saturation is required to make the peak shift unmeasurable.[7]

For a given head-to-medium separation and on a given recording medium thickness, for most systems optimum resolution is attained at the intensity which is just sufficient to saturate the recording medium. Excessive write current tends to widen the readback pulse, reduce its amplitude and shift its peak position.[5].

Method (2) can be employed to overcome these undesirable characteristics caused by over-saturation. This method requires two separate magnetic elements: One for erasing, and the other (a somewhat narrower pole tip) for writing and reading. When the widths of the pole tips are properly designed, the residual noise and crosstalk are low. Because previously written data is first (ac or dc) erased, the write current used can be the same intensity as that of the saturation current, and the undesirable effects caused by excessive write current can be avoided. This method has a disadvantage, in that readback signal amplitude drops gradually as the head is displaced transversely across the track during reading. However, if an AGC (automatic gain control) is used in the sensing circuit, the resulting small gradual drop in signal amplitude does not seriously affect the sensing operation. The manufacturing cost of the head is low because the erase pole tip design and fabrication are simple.

Fig. 7 shows the track performance characteristic obtained by method (2). The head used was designed for 50 TPI with a total head repositioning error of 6 milli-inches. The data was taken on an iron-oxide (γFe_2O_3) coated disk having a coating thickness of 0.4 milli-inch. The relative head-medium velocity was maintained at 1200 inches per second, while the head-to-medium separation was 0.25 milli-inch. The readback signals of these adjacent tracks are presented in Fig. 8. As shown, partial erasure by the erase element, when the adjacent track was written at a maximum head repositioning error of 6 milli-inches, reduced the track's signal amplitude by approximately 20 per cent. Under this worst condition, however, an examination of the data indicates that noise due to old data and crosstalk is negligible, and that no peak shift is detected.

Two separate magnetic elements are also used in method (3): A wider pole tip for writing (without first erasing), and a narrower pole tip for reading. Method (3) is preferred when the head-repositioning error is large, and when constant readback signal amplitude is desired. However, both the write and read elements must be well designed and fabricated. The residual noise and crosstalk are generally higher in method (3) than in method (2). In addition, the same undesirable characteristics on the readback signal as stated in method (1) are present.

The track performance characteristic obtained by method (3) is illustrated in Fig. 9. The head used was designed for the same track density (50 TPI) and for the same head-repositioning error (6 milli-inches), and was tested under the same operating conditions shown in Fig. 7. The constant readback signal amplitude through a wide range of head displacement is attractive, but is lost when the adjacent track is rewritten at the maximum head repositioning error. (See the solid line of track C, Fig. 9.) In addition, the noise due to old data and crosstalk is somewhat higher than that attainable by method (2).

Two heads were constructed for high track-density study in noncontact recording. One was designed for 200 TPI and the other for 500 TPI. Both heads are of the dual-element type using the erase-wide and read/write-narrow method of recording. For the 200 TPI head, the erase pole-tip width is 5.2 milli-inches and the read/write pole-tip width is 2.8 milli-inches, while those for the 500 TPI head are 1.7 milli-inches and 1.0 milli-inch, respectively. Inter-track shields are used on both heads to minimize crosstalk.

Fig. 10 shows the track performance characteristic of the 200 TPI head with a head repositioning error of 1.0 milli-inch; Figs. 11 and 12 present the amplitude and peak shift characteristics of its readback signal at various bit densities. Data obtained by the 500 TPI head indicates that no appreciable head displacement can be tolerated. The written data at 2.0 milli-inch track centers, however, is good enough for information recovery. The readback signals of three adjacent tracks written and read back by this head at 500 TPI and 2,000 BPI are shown in Fig. 13. The readback signal amplitude and peak shift characteristics of this head are very similar to those shown in Figs. 11 and 12 for the 200 TPI head.

The intertrack shield has been found to be very effective in high track-density recording, for confining the widths of the erase track and of the write track, and for eliminating crosstalk. Its primary function is to provide a sufficiently low reluctance path, for shunting the excessive fringing field in recording, and for shielding the flux emanating from adjacent tracks in reading. The spacing between the shield and the pole tip depends largely on the track density and on the head-repositioning error.

Mechanical precision in fabrication is the present major limitation to the attainment of higher information storage density in magnetic recording. The head-gap length, pole-tip geometry and head-to-medium separation are the chief parameters affecting the achievement of better pulse resolution. Head-repositioning error and manufacturing tolerances are the major factors which limit the accomplishment of higher track density. With better head design and with proper choice of the recording method, the limitations imposed by mechanical tolerances can be reduced. In most cases, the use of a thinner recording medium with a more rectangular B-H characteristic gives a further improvement in pulse resolution.

CONCLUSIONS

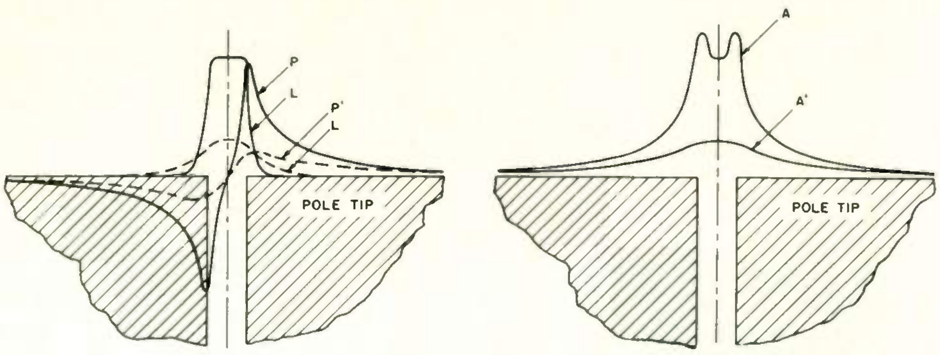
The effects due to head pole tip geometry in noncontact recording have been presented. Analysis and experimental data are in good agreement, indicating that significant improvement in pulse resolution is achieved by making the pole-tip length small and the angle between the pole-tip edge and recording surface large. Both features help to limit the spread of the writing field, and consequently reduce the trailing-field effect. This form of geometry also yields a sharper gradient of the head sensitivity function and therefore resolves a narrower readback pulse.

In addition, several recording methods and the effect of the head-repositioning error on the track density have been presented. Performance characteristics obtained from experimental data show no fundamental magnetic limitations in using very narrow tracks (up to 500 TPI) in noncontact recording.

Further, the value of the concepts and techniques developed has been demonstrated through their successful application in the design of high-density magnetic heads for noncontact recording. With good mechanical precision, eight hundred thousand to one million information bits of digital data per square inch in noncontact recording were obtained in this study.

REFERENCES

1. Noncontact refers to head-to-medium separation of 100 microinches or larger.
2. D. F. Eldridge and A. Baaba, "The Effect of Track Width in Magnetic Recording." IRE Transactions on Audio, pp. 10-15, January-February, 1961.
3. Bit packing density is defined here as the number of pulses per inch where amplitude reduction starts, and track packing density is defined as the number of tracks per inch where crosstalk from adjacent tracks begins.
4. S. J. Begun, "Magnetic Field Distribution of a Ring Recording Head", Audio Engineering, Vol. 32, pages 11-13, December, 1948.
5. J. J. Miyata and R. R. Hartel, "The Recording and Reproduction of Signals on Magnetic Medium Using Saturation-type Recording," IRE Transactions on Electronic Computers, pages 159-169, June, 1959.
6. A. S. Hoagland, "Magnetic Data Recording Theory: Head Design," Communications and Electronics, Vol. 27, pp. 506-512, November, 1956.
7. Donald F. Eldridge, "Magnetic Recording and Reproduction of Pulses," IRE Transactions on Audio, pages 42-57, March-April, 1960.



(A) THE LONGITUDINAL AND PERPENDICULAR COMPONENTS NEAR THE SURFACE OF THE RECORDING HEAD ARE L AND P, AND SOME DISTANCE ABOVE THE POLE FACE ARE L' AND P'.

(B) THE ABSOLUTE MAGNITUDE OF THE FIELD NEAR THE SURFACE OF THE RECORDING HEAD IS A, AND SOME DISTANCE ABOVE THE POLE FACE IS A'.

Fig. 1. Longitudinal and perpendicular components, and absolute magnitude of the field distribution.

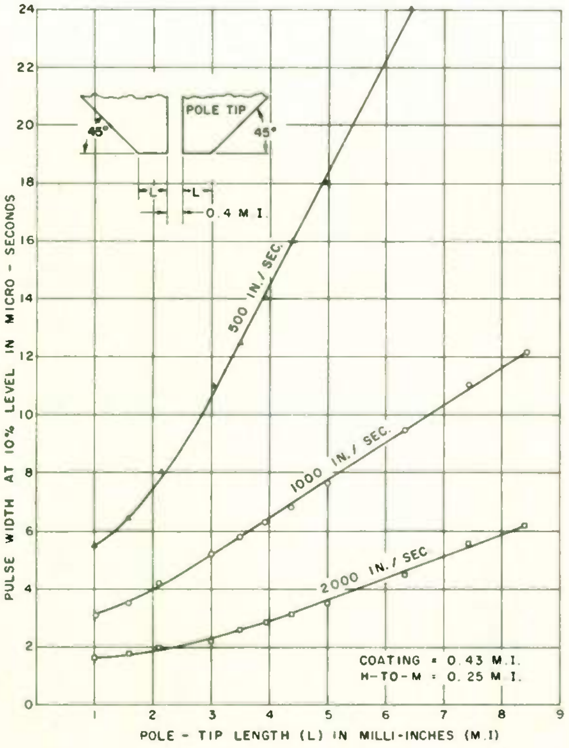


Fig. 2. Effect of head pole-tip length on readback pulse width.

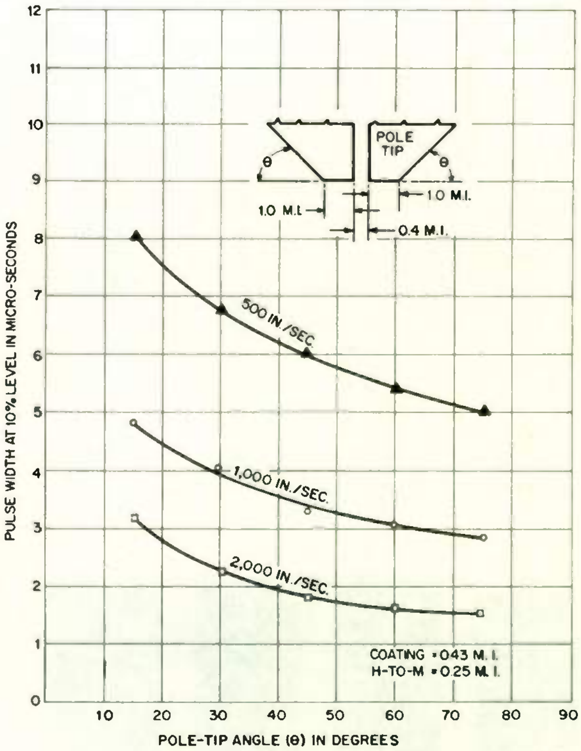
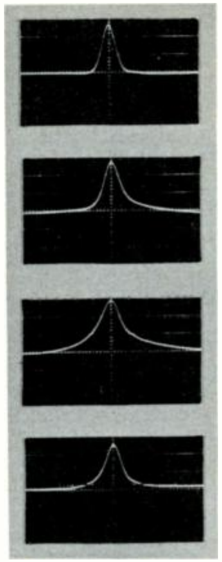


Fig. 3. Effect of the angle between head pole tip-edge and recording surface on readback pulse width.



(A) WRITTEN AND READ BACK BY HEAD L-1

(B) SAME WRITTEN TRANSITION OF (A) BY HEAD L-1, BUT READ BACK BY HEAD L-8

(C) WRITTEN AND READ BACK BY HEAD L-8

(D) SAME WRITTEN TRANSITION OF (C) BY HEAD L-8, BUT READ BACK BY HEAD L-1

POLE TIP LENGTH:
L-1 = 1.0 MILLI - INCH;
L-8 = 8.4 MILLI - INCHES.

Fig. 4. Readback pulses by two heads with different pole tip lengths.

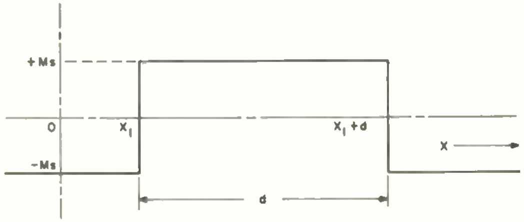
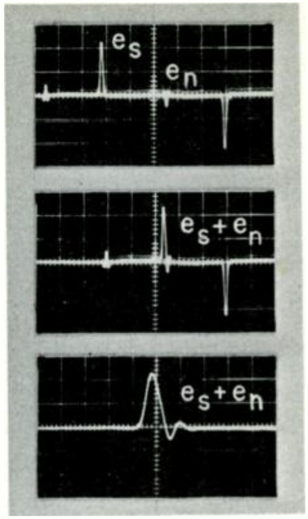
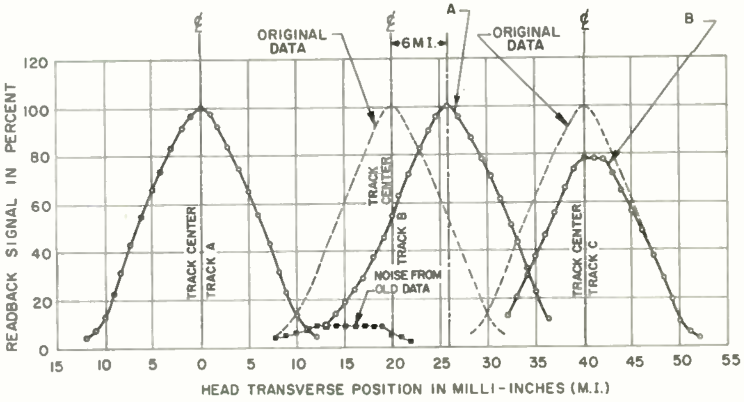


Fig. 5. Two step-function distribution of magnetization from - M_s to + M_s, showing two "1's" (NRZI Recording).



AS THE DATA PULSE GETS TO THE POSITION OF THE SMALL NOISE PULSE, THE TWO ARE SUPERPOSED, RESULTING A DISTORTED PULSE AS SHOWN IN THE LOWER TRACE.

Fig. 6. Readback signal waveforms superposed.



A- REWRITTEN DATA AT 6 M.I. OFF TRACK
B- SIGNAL LEFT AFTER ADJACENT TRACK WAS REWRITTEN AT 6 M.I. OFF TRACK

Fig. 7. Readback signal amplitude versus head transverse position by the erase-wide and read/write-narrow recording method.

NRZI WORD PATTERN

TRACK A 1000011110 AT 0

TRACK B 1111111111 AT 0.020 IN.

TRACK B 1000000010 AT 0.026 IN.
(REWRITTEN)

TRACK C 1100001110 AT 0.040 IN.

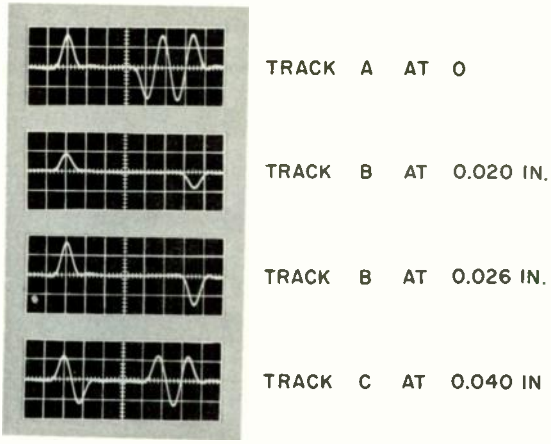


Fig. 8. Readback signals of three adjacent tracks after center tract was rewritten at 6 milli-inches off track.

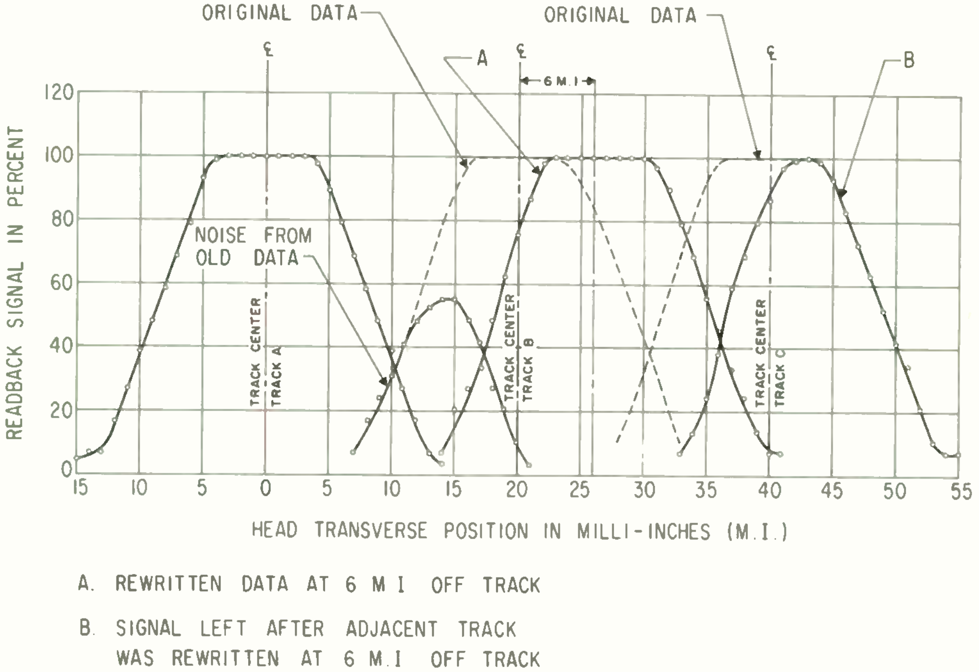


Fig. 9. Readback signal amplitude versus head transverse position, by the write-wide and read-narrow recording method.

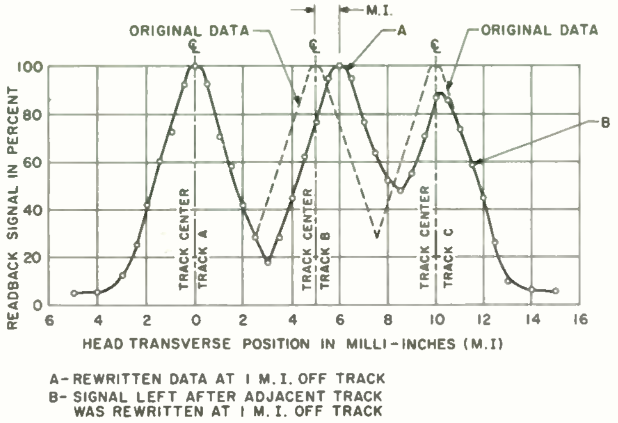


Fig. 10. Readback signal amplitude versus head transverse position (200 TPI).

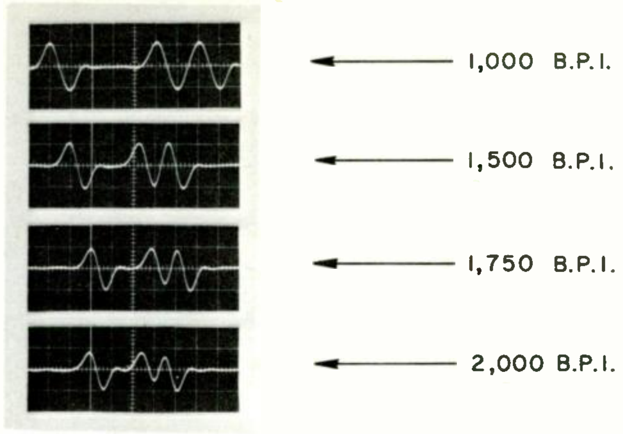


Fig. 12. Readback signal at different densities (200 TPI).

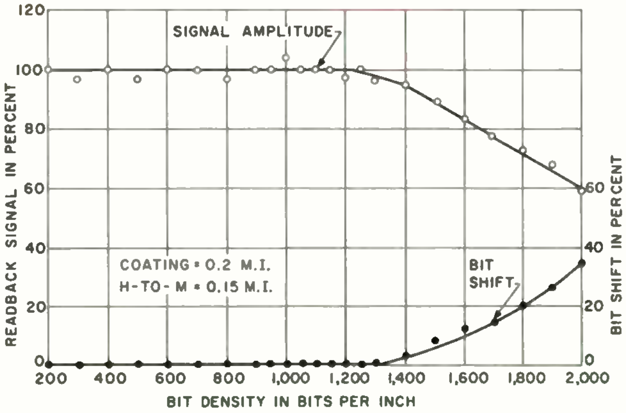


Fig. 11. Readback signal amplitude and bit shift versus bit density (200 TPI).

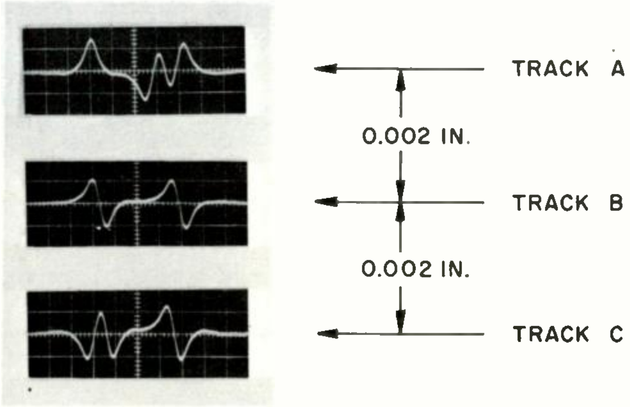


Fig. 13. Readback signals from three adjacent tracks (500 TPI and 2,000 BPI).

A COMPACT 166-KILOBIT FILM MEMORY

R. D. Turnquist, V. E. Christiansen, and C. O. Hogenson
Remington Rand Univac
Division of Sperry Rand Corporation
St. Paul, Minnesota

Summary

The theory, design, and operating characteristics of a 166 thousand-bit thin film memory system are presented in this paper. Although this compact memory system is an integral part of a guidance and control computer designed specifically for an aerospace application, the design is adaptable to other military control computers. This random-access, parallel-readout, word-organized memory system includes a 256-word, 24-bit destructive readout (DRO) memory and a 6656-word, 24-bit, alterable, nondestructive readout (NDRO) memory. Cycle time in the program mode is 3.0 μ sec, and access time is 0.7 μsec.

The unit is designed to operate reliably without temperature compensation while substaining high levels of shock and vibration. The packaging techniques inherent in the use of thin film memory elements have permitted the fabrication of a memory stack, including selection circuitry, weighing 16 pounds and occupying about 0.2 cu. ft.

Modular circuit packaging provides a compact configuration and yet permits effective maintenance. The entire memory system, including all associated circuitry, requires 50 watts of power.

I. Introduction

This paper concerns the development of a computer memory system which maintains a high degree of reliability under conditions of unusually severe environmental stress. Typical uses of a computer incorporating this memory system include control applications in mobile equipment; high Mach-number aircraft; ballistic, orbital, and powered flight vehicles; and near- and deep-space craft, both manned and unmanned.

The memory system was designed for the particular aerospace computer shown in Figure 1. This machine, known as the ADD Computer, is intended for use in a missile guidance system. The computer, including input-output, memory, control, power supply, and arithmetic sections, weighs 85 pounds, occupies 1.1 cubic feet, and consumes 265 watts of power. The computer operates in the binary, parallel mode and employs two's complement fractional arithmetic. The logic design takes maximum advantage of the high-speed memory system. Assigned memory addresses are used in place of conventional arithmetic registers and counters.

II. Memory Design Criteria

A. Design Goals

A memory system to meet the requirements for use in an aerospace environment should have these general characteristics:

SOLID STATE CONSTRUCTION - to eliminate mechanical or rotating parts, thereby minimizing shock and vibration problems and insuring long life.

NONDESTRUCTIVE READOUT - for at least part of the memory, to protect critical programs and constants from chance of alteration due to transient readout errors and to eliminate power needed to regenerate the information.

ELECTRICAL ALTERATION - to allow rapid change of memory contents by remote means if so desired.

RANDOM ACCESS - to enhance speed capability and to simplify programming.

In addition to the system requirements, other pertinent factors such as speed, power, weight, and high-temperature operation must be considered.

B. Design Approach

A design centering around the use of magnetic-film memory elements satisfies the above requirements in a more optimum way than any other memory technique. A film memory is characteristically solid state, and allows random-access, parallel readout. The ADD computer was designed around a memory system which provides two separate types of storage: a permanent one for programs and constants and the other for scratch-pad or transient storage.

The design of the permanent storage section was accomplished by using a special device developed at UNIVAC called the BICORE film element.[1] This element provides reliable nondestructive readout (NDRO) with a wide latitude of read current and also provides for electrical write-in. The scratch-pad design utilizes a film memory element similar to that reported by Raffel.[2] The flat geometry of film arrays lends itself to a packaging approach which is highly resistant to shock and vibration.

A circuit packaging technique was developed which permits high-density packaging, provides good heat conduction, and allows for operation under high levels of shock and vibration. The exclusive use of silicon semiconductors permits reliable circuit operation over a wide temperature range.

III. Memory Operating Principles

A. NDRO Operation

The BICORE film element (Figure 2A) is used as the nondestructive readout (NDRO) or permanent storage element. This element, consisting of a high coercivity storage film and a low coercivity readout film, is fabricated of several layers, yet physically appears as a single film. The operating principle of the BICORE element relies upon the strong external field of the storage film to control the state of the readout film. Writing is accomplished by applying a field of sufficient strength to switch the storage film to a given state. Upon removal of the write field, the readout film will switch to the opposite state. A read field, which is approximately 25 per cent of the write field, will switch the readout film without disturbing the storage film. Upon removal of the read field, the storage film again assumes control of the readout film and reswitches it to its original direction.

This section requires only word-line and digit-line conductors oriented 90 degrees to the easy axis of the BICORE element (Figure 2B). The digit line is used to sense the film output when reading and to carry digit current when writing. Figure 2C illustrates the waveforms existing on the word and bit lines for read and write modes of operation.

The BICORE elements are deposited on a 6-mil glass substrate measuring 1.8 by 2.4 inches. They are 35 mils in diameter and have a total thickness of 6000 angstroms. Each substrate has 768 BICORE elements deposited in a 24 by 32 array.

B. DRO Operation

The destructive readout (DRO) memory element for the scratch-pad section is a thin Permalloy film similar to that reported by Raffel[2] but with the unique addition of a second layer of cobalt-iron film. The scratch-pad section is operated in the DRO mode, so each read reference is automatically followed by a restore operation during the latter part of the cycle. The word-line parallels the film easy axis, whereas the bit and the sense lines are perpendicular to the easy axis (Figure 3A).

The transverse field produced by the word current causes the film magnetization to rotate 90 degrees from the easy axis. The longitudinal bias field from the cobalt-iron film insures that the storage film magnetization will rotate consistently to the same state upon removal of the transverse field. This state is defined as the "zero." To orient the magnetization of the storage film in the "one" state, the direction of the longitudinal field is reversed by the application of the bit current before the transverse field is removed. The sense line is placed adjacent to the film to sense the flux change in the longitudinal direction.

Figure 3B illustrates the word, bit, and sense line waveforms for a read/restore cycle. Although a unidirectional word current pulse could be employed, the bipolar pulse requires less power since it takes advantage of the reswitch properties of the square loop core used in the selection system.

The 3600-angstrom DRO films are 35 by 70 mil rectangles deposited on a 6-mil glass substrate. Each substrate has 384 films arranged in a 12 by 32 array.

IV. Memory System

A. Organization

A logical diagram of the memory system is shown in Figure 4. The memory system requires only three control links with the computer: (1) an initiate, (2) a line which selects the NDRO or DRO section of the memory system, and (3) a line which selects a read or a write mode for the two sections. The computer may select either an NDRO read or a DRO read-restore operation at a 3-μsec cycle time. Access time of each memory is 0.7 sec.

The NDRO section of the memory system requires 52 memory planes to obtain a capacity of 6656 twenty-four bit words. For economy the memory stack is actually organized into 3328 forty-eight bit, double-length words. When the computer requests a word, one bit of the address is used to determine which half of the 48-bit word is to be transmitted to the computer. Two sets of 24 sense amplifiers share the same data register (register no. 1). During loading operations both data registers are used under computer control to load the 48-bit word. The write cycle is slow (25 milliseconds per word) since the memory is loaded from a paper tape reader, and power, weight, and space are reduced by using a slow cycle. A remote unit furnishes power for the write operation, thus simplifying the computer power supply unit and insuring that writing cannot take place inadvertently.

The DRO section of the memory stack requires four memory planes to obtain a capacity of 256 twenty-four bit words which are stored in the memory in 24-bit length. The DRO predrivers provide computer control over the read-restore path so that bits in a word may be shifted or restored to different bit locations. This feature enables the DRO section to function as a shift register under computer control.

Because it was not required to reference the two sections of the memory simultaneously, many of the circuits are shared. Address selection circuits are common to both sections, with the lowest order 256 addresses used for DRO and the remaining

6656 addresses being NDRO. The sense amplifiers serve both memory sections.

B. Address Selection

The selection of a word in either memory section is accomplished under control of a 13-bit address register by selecting one of 64 word lines of both the NDRO section and the DRO section and then selecting one of 52 NDRO memory planes or one of four DRO memory planes.

Figure 5 depicts the address selection scheme and illustrates how both sections share the 64 X-selection switches. Sixty-four tape-wound Permalloy switch cores and 64 high-conductance microdiodes are mounted on a component board on each of the 56 memory planes. Switch cores and diodes are arranged electrically in a 64 by 52 matrix for the NDRO section and a 64 by 4 matrix for the DRO section. The Y-selection, or plane selection, is accomplished by using smaller matrices, an 8 by 7 for the NDRO section and a 4 by 1 for the DRO section.

To select a word, one of the 64 X-selection switches is turned on, and a short time later one of the 56 Y-select lines is pulsed. Current in the primary of the selected switch core overcomes a bias current and provides a positive current pulse through the word line. When the primary current is removed, the bias current reswitches the core and provides a negative current pulse through the word line.

During the write mode the bias current is necessarily higher than it is for the read mode. To reduce power dissipation, the bias current is pulsed to the proper amplitude during the write operation.

Because the Permalloy cores are constantly biased, sneak currents through unselected word lines are virtually eliminated. Any such sneak current on the core primary would have to be of sufficient magnitude to overcome the bias before it could couple into the word line.

The low impedance of the word lines necessitates an impedance-matching circuit to the semiconductor drive circuits. The match is accomplished by the Permalloy core, which has a ten-turn primary and a one-turn secondary.

The dual-polarity word pulse is used in conjunction with an overlapping positive or negative digit pulse to accomplish writing a "one" or a "zero." The amplitude of positive or negative current for read or write mode is determined by the regulator and the bias current through the switch cores. The selection system functions in a similar manner for both read and write modes.

C. Digit Control

The digit control circuits for the two memory sections are shown in Figure 6. Two types of sense amplifiers are used, one for NDRO operation only and the other for NDRO-DRO operation. The output signal from the DRO film element is smaller than that of the NDRO. A larger step-up ratio on the DRO input transformer compensates for this difference.

The main amplifier consists of three stages of direct-coupled Class A amplification. A d-c feedback loop from the output collector to the input maintains the bias stability. Direct coupling reduces the number of large coupling capacitors. The voltage gain, excluding input transformers, is about 1000. The pass band is from 1 Mc to 3.5 Mc. Typical outputs are shown in Figure 7. Three different strobe pulses are used with an AND-OR network to gate the output of the sense amplifiers into the blocking oscillator which feeds the memory data register.

Also shown in Figure 6 are the DRO and NDRO bit drivers. The DRO memory utilizes a separate bit line. The driver has a transformer-coupled output and utilizes a transistor switch and a current source in the primary of the output transformer. The NDRO sense amplifier and bit driver share the same line in the memory stack. A diode-resistor network on the sense amplifier input prevents bit current from being short circuited by the winding impedance of the sense amplifier input transformer. The NDRO bit driver circuit utilizes two silicon-controlled rectifiers to discharge an energy storage capacitor through a center-tapped transformer to generate the positive or negative current for the write operation. The sense amplifier experiences a large transient during writing and requires a recovery time of several milliseconds before a read operation can be initiated.

V. Packaging

A. Memory Stack

The plane assembly for the DRO section as shown in Figure 8 is made up of four substrates and a three-layer wiring array containing the drive and sense conductors. The conductors, approximately the same width as the film element, are etched from copper laminated on Mylar film. The substrates are arranged to form sixty-four 24-bit words. The "fold-over" wiring array provides the end connections to the word line. Also a part of the plane assembly is a component board containing the diode-switch core word gates, discussed earlier.

The NDRO plane assembly contains four substrates arranged to form sixty-four 48-bit words. Again the "fold-over" is used for this plane; only a two-layer wiring array is required. The word lines and digit lines are parallel when passing a BICORE element, whereas the DRO word lines cross at right angles with the digit and sense lines. The DRO planes are identical in size with those of NDRO, so that both plane types can be placed together within one stack. For rigidity, all wiring arrays are cemented to epoxy boards. Each plane assembly measures 6.8 by 5.7 by 0.11 inches and weighs about 3 ounces.

Eight planes are assembled into a 1.5-pound "eight-pack" unit measuring 6.8 by 5.7 by 1.0 inches.

The next level of assembly is the memory stack (Figure 9). Seven "eight-packs" make up the stack, one of which is divided into DRO (4 planes) and NDRO (4 planes). Enclosed in a magnetic shield, the memory stack measures about 7 by 7 by 7 inches and weighs 16 pounds.

B. Circuits

The package for the circuitry is shown in Figure 10. This building block uses a "cordwood" structure and is encapsulated with a light-weight epoxy. All heat sources (i.e., resistors, transistors, diodes, etc.) are in close proximity to an aluminum heat sink. The rugged, sealed, enclosures for the components increase the reliability substantially. A typical block has outside dimensions of 0.5 by 0.4 by 1.6 inches and contains approximately 40 components. Nickel tabs are provided for circuit connections. Resistance welding is utilized for circuit interconnection. Two hundred fifty building blocks are used for the memory electronics.

Building blocks are assembled in modules as shown in Figure 11. These are thin-walled, finned aluminum castings which may contain as many as 68 building blocks, together with power filters and input-output connectors. The building blocks, cemented to the module frame, are conduction cooled. The module itself is cooled by air passing between the fins.

Interblock connections are made by resistance-welding conductors to the nickel output tabs, with the conductors being routed through the wiring channels. Wires are secured by coating with an epoxy "freeze coat."

A typical assembled module weighs about 2.3 pounds and measures 6.5 by 6.6 by 1.1 inches. Four modules are used to contain the memory building blocks. Figure 12 illustrates how the modules are physically arranged with respect to the memory stack.

VI. Test Results

A complete computer employing this memory is now in operation and is undergoing functional testing. Diagnostic routines have been run satisfactorily and a complete real-time simulation has been performed. This simulation was performed over a continuous 48-hour period with no malfunction. Additional tests have shown that the power supply voltages can be varied ±10 per cent with no errors.

Upon the completion of functional testing this computer will undergo extensive environmental testing. These tests will include shock, vibration, temperature, altitude, radio interference, and humidity.

During the design phase the following environmental tests were made on memory subsystems.

TEMPERATURE - Hysteresis loop data of single film elements from -100°C to +150°C indicated no change in their parameters. A 5000-bit NDRO stack and a 960-bit DRO stack operated satisfactorily over a temperature range of -40°C to +71°C.

SHOCK - shocks of 120-g's for 0.0065 second in three axes resulted in no damage to an NDRO memory plane and did not change the stored contents.

VIBRATION - A 5000-bit NDRO stack was operated under vibration at 6.0 g's RMS from 5-2000 CPS in three axes with no malfunction.

HUMIDITY - Film core arrays of 768 bits have passed a ten-day 95 per cent humidity test with no deterioration.

LIFE - Film core arrays subjected to 2000 hours of temperature and magnetic field cycling showed no change in their functional characteristics. NDRO arrays have been interrogated 10^9 times with no change in their stored contents.

VII. Conclusion

The successful operation of this memory has proven the practicality of large-scale film memories. Film memories are particularly well suited for operation under severe environmental conditions because of the flat geometry of the array and because the films are quite insensitive to temperature changes. There is every indication that in the near future even more sophisticated uses will be developed, centering around the use of thin films as memory elements.

VIII. Acknowledgments

The authors appreciate greatly the work of the people who assisted in this effort and the encouragement of those who directed it. Special thanks are due to R. J. Petschauer for his many suggestions throughout the design of the memory system as well as in the preparation of this paper.

[1] Petschauer, R. J. and Turnquist, R.D. "A Nondestructive Readout Film Memory," Proceedings of WJCC, Los Angeles, California (May 1961)

[2] Raffel, J. I., "Operating Characteristics of a Thin Film Memory," J. Appl. Physics 30, 60S-61S (1959).

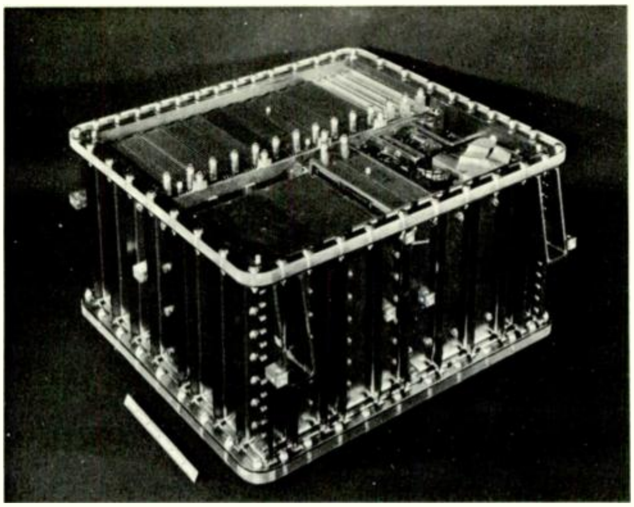


Fig. 1. ADD computer.

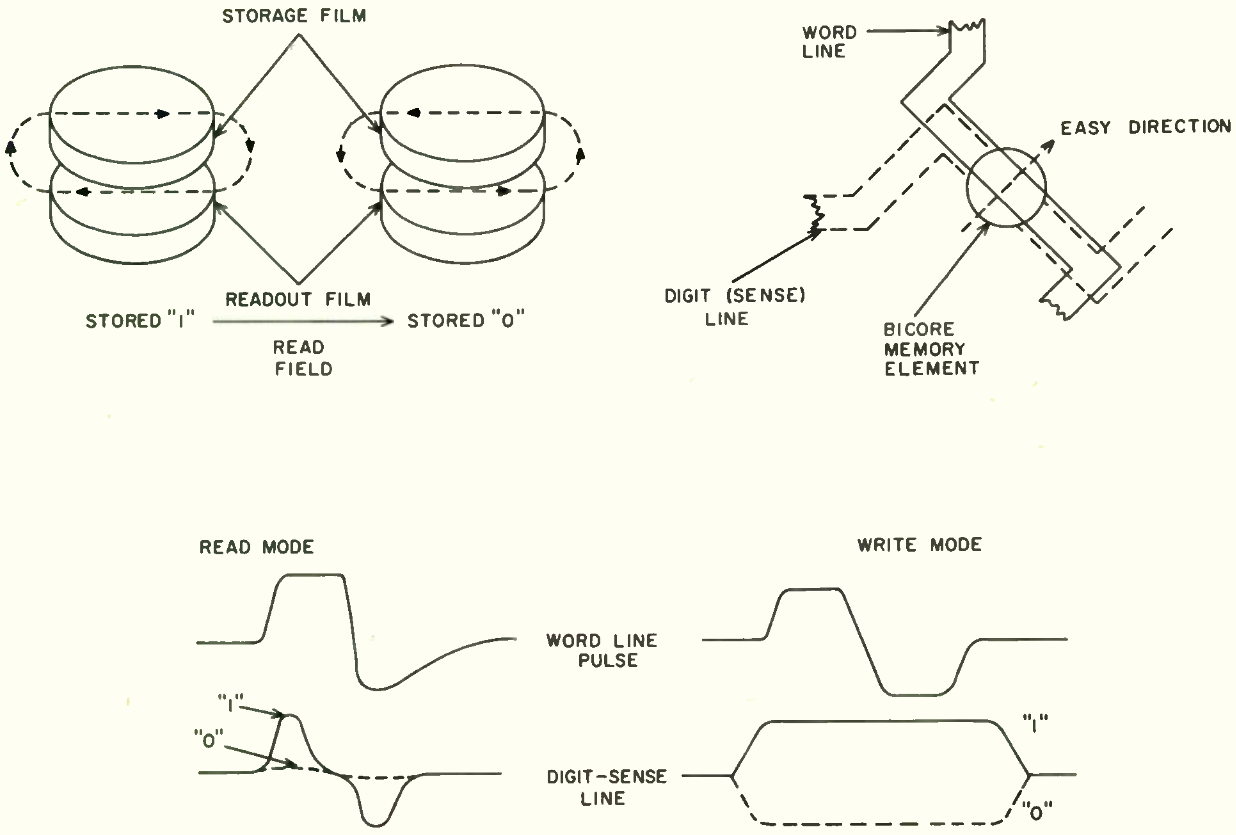


Fig. 2. NDRO BICORE element operation. (a) Coupling of storage and readout film. (b) Geometry of word-digit lines. (c) Pulse schedule.

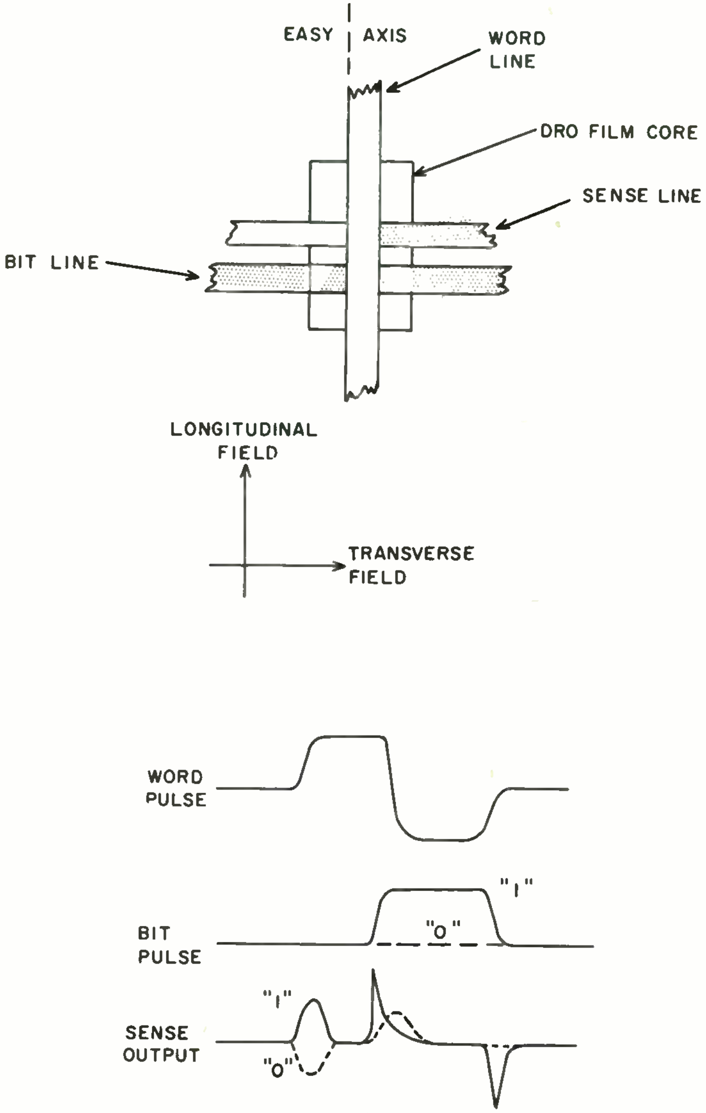


Fig. 3. DRO film element operation. (a) Film element. (b) Read-restore time cycle.

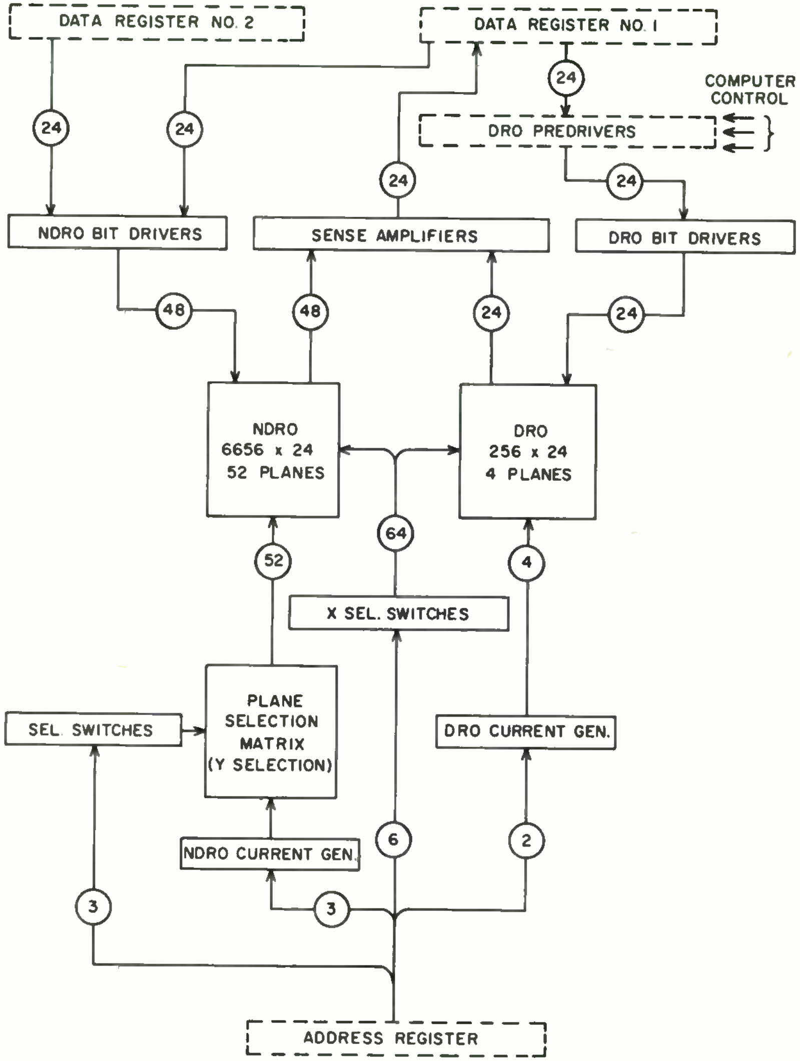


Fig. 4. Memory system logic diagram.

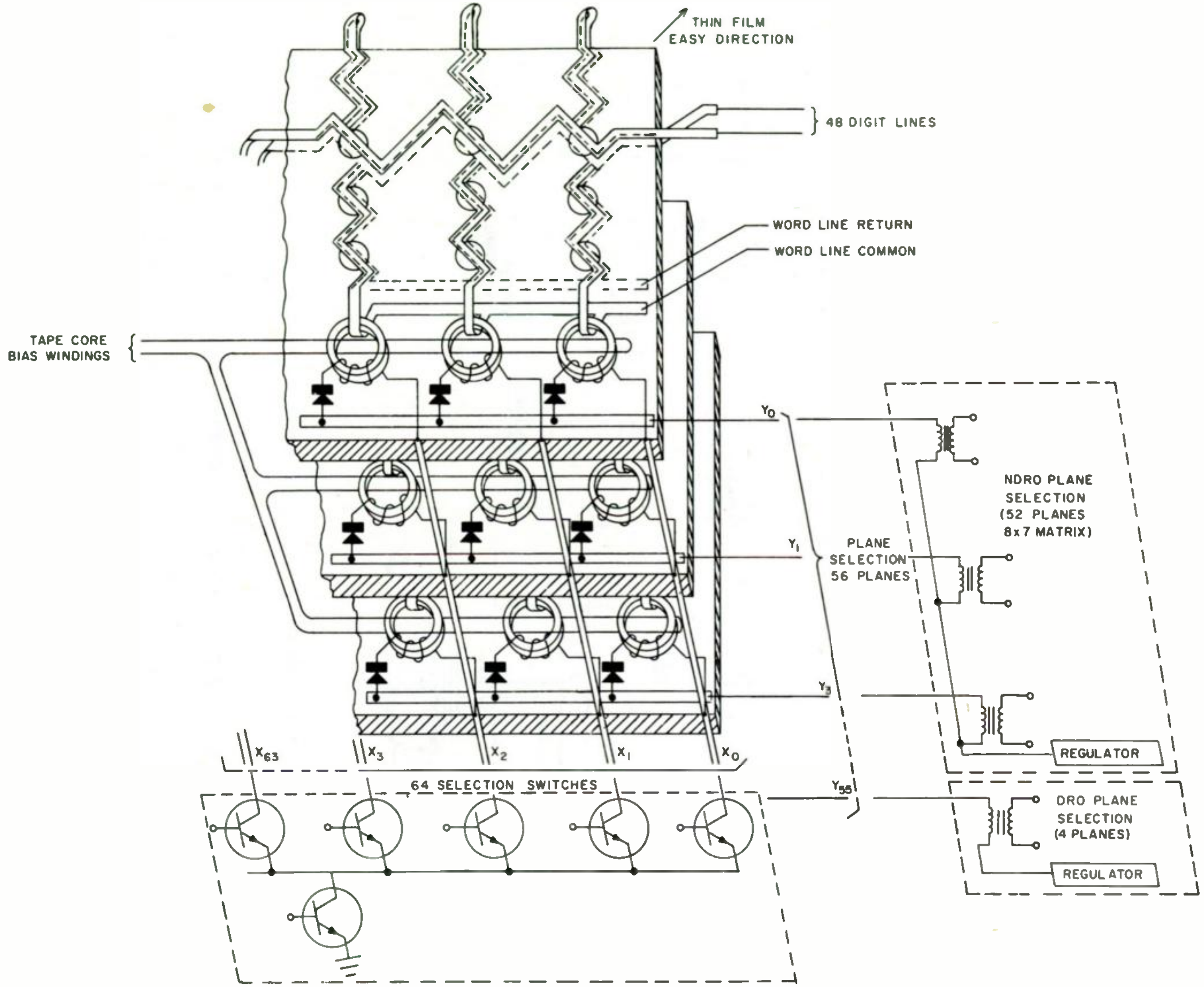


Fig. 5. Address selection scheme.

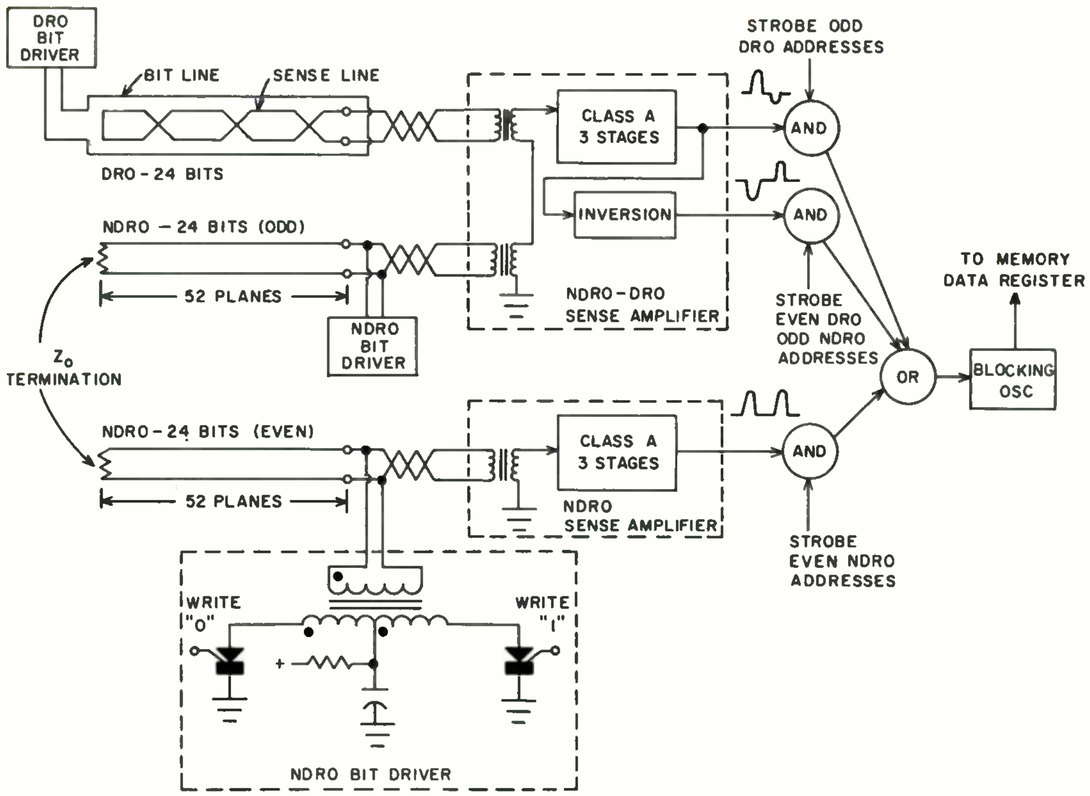
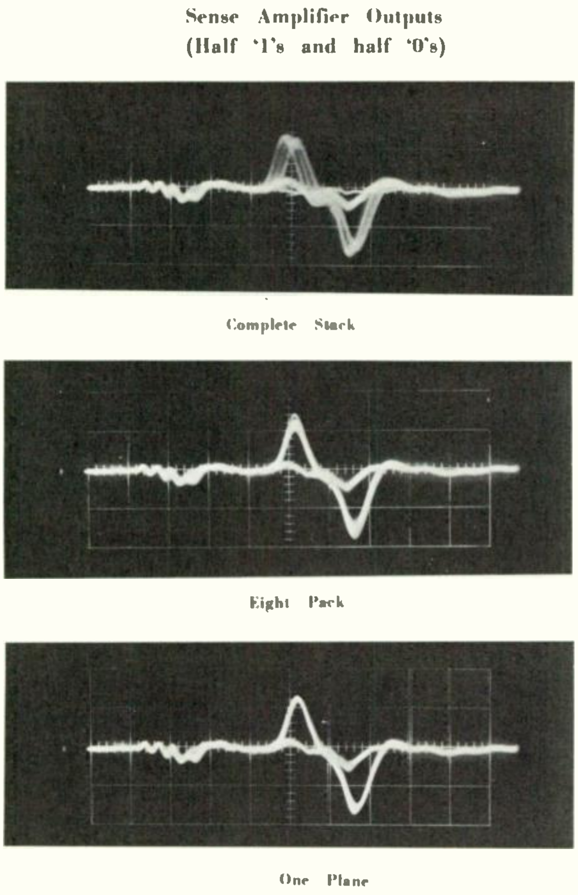


Fig. 6. Digit control circuits.



Complete Stack

Eight Pack

One Plane

Fig. 7. NDRO sense amplifier outputs (half "ones" and half "zeros"). Horizontal calibration: 200 nsec/div. Vertical calibration: 2 volts/div. (a) Complete stack. (b) Eight pack. (c) One plane.

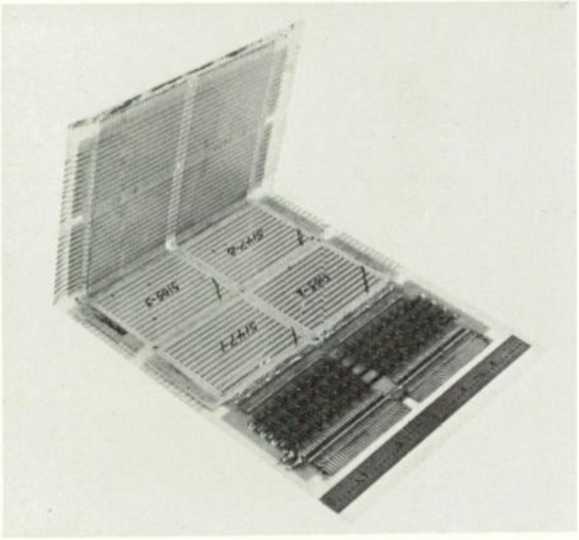


Fig. 8. DRO memory plane.

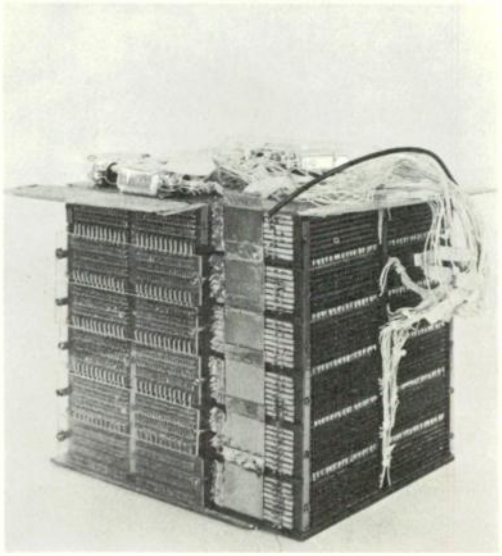


Fig. 9. Stack construction (NDRO section only).

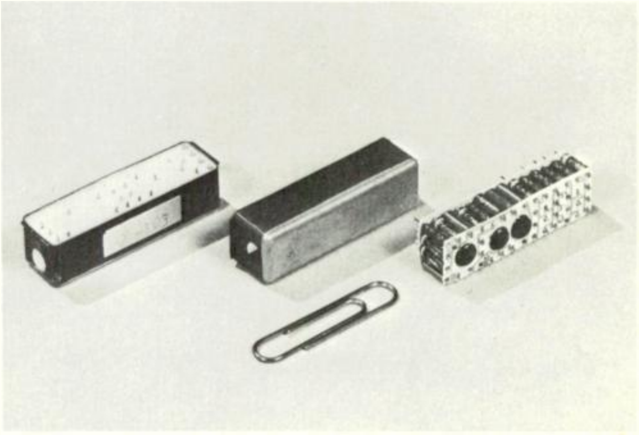


Fig. 10. Building block assembly.

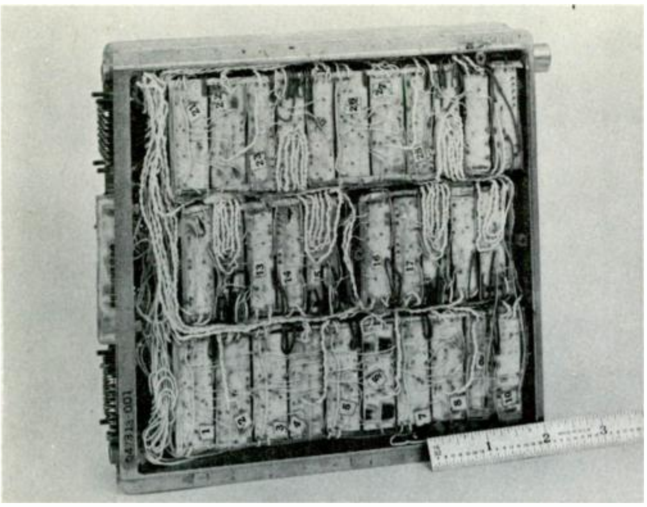


Fig. 11. Memory module.

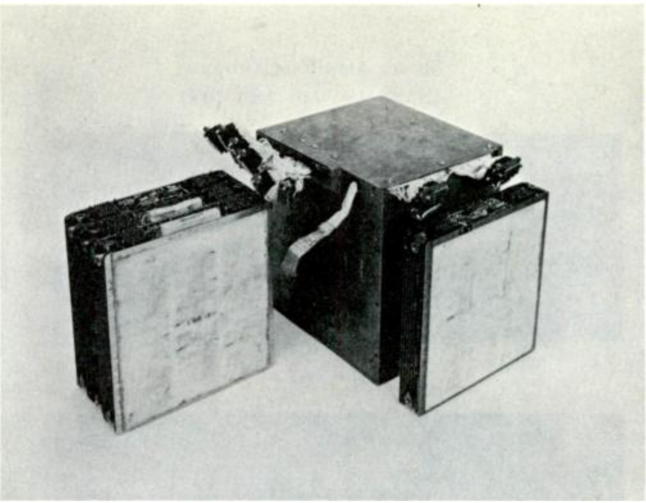


Fig. 12. Memory system assembly.

COMPUTER-CONTROLLED ASW TRAINING FACILITY

Edward B. Boyle, Jr.
Roy L. Edwards, Jr.
Aircraft Armaments, Inc.
Cockeysville, Maryland

Summary

Use of a general-purpose digital computer to control a complex submarine ASW training facility is described. Functioning as an integral part of the system, the computer generates, under the control of program operators, the courses of submarines, ships, aircraft and weapons to create the training environment. Outputs from the computer are used to provide sonar and radar inputs to the submarine crews undergoing training. In addition, the computer activates the program operators' displays and furnishes data for scoring the students' performance. The input/output and programming requirements imposed by using a computer for Real-Time simulation are discussed.

Introduction

The ASW training facility described in this paper is an example of a modern digitally-controlled real-time simulation and training system designed for intensive training in the use of complex equipment. Development of this system is being performed by Aircraft Armaments, Inc., Cockeysville, Maryland, and directed and supervised by the U. S. Naval Training Device Center, Port Washington, N. Y. under contract N61339-949. The facility will provide complete and realistic training in the operation of submarine fire control systems, and in operation of the sensory equipment which provides the input data to the fire control system. Training exercises which would require many ships and extensive periods of time to perform at sea can be performed on land, while preserving much of the realism of actual at-sea conditions and at the same time permitting ease and accuracy of observation and scoring which could not be obtained with the use of actual ships.

Training Requirements

The type of training which is required may be explained by reference to the tasks and the equipment used to perform them. Figure 1 is a simplified block diagram of the organization of the men and equipment used in the attack center in the process of pressing home an attack. The immediate information about the movement of the target must be obtained from the sonar, periscope and radar. Decisions, based on experience, must then be made as to the probable accuracy of data, when to expose the submarine's position by the use of the sensory equipment, and the like. The data thus selected is fed to the fire control system, which is able to compute the probable motion of the target, and the appropriate weapon settings.

At the same time, other command decisions are made, taking into account all the many variable factors of the situation, and orders are given to maneuver the submarine in a manner which will assist the attack. The most advantageous weapons of those available are selected for use in the final stage of the attack, and orders to load these weapons are given to the torpedo room.

While no one of these actions is extraordinarily difficult, the combination of these tasks and the decision-making which accompanies them is sufficiently complex to require the best available training.

The problems of submarine training have long ago been recognized, and many previous simulators have been constructed to provide training to attack center crews. Trainers employing analog computers coupled to simulation equipment have been built and successfully used, but they have severe limitations. Only a small number of targets can be accommodated, the area over which they can operate must be kept small to avoid reducing accuracy, and scoring methods have been unsophisticated. Realism is limited by the small amount of computations which can be performed, and operation and maintenance of this type of equipment is difficult. To compound these problems, newer weapons have complex trajectories, which are extremely difficult to generate by analog means.

The advent of large digital computers, which combine tremendous computing capacity with high reliability at a moderate cost, has made it possible to surmount most of these drawbacks. Realistic simulation, flexible control with a reasonably small number of operators, and accurate scoring of results can all be supplied so that rigorous training can be administered. While some compromises must still be made to keep the total amount of computations within the limits of computer capacity, quite satisfactory results can be obtained. At the same time, the digital computer's capacity for self-diagnosis can be used to simplify maintenance, and by comparatively simple reprogramming, large changes can be made in the trainer with a moderate amount of effort.

The essential items necessary to perform the training task are a mocked-up attack center complete with equipment used during the course of submarine attack, computing and simulation equipment, and control stations for program operators. The simulation equipment is required to activate the sonar, radar and periscope so that the usual sources of input data are available. A computation system is required to generate target motions, score crew actions, and simulate the paths of weapons used by the attack center crew. Human operators are needed to provide the voices and some of the actions of members of the crew who assist in the operation, but who are not present in the attack center. These program operators, as they are called, also act as helmsmen for the submarine, monitor the progress of the training exercise, and control the actions of the target vessels, as they respond to the actions of the submarine.

Facility Arrangement

The building which houses the training complex is designed to afford a logical arrangement of the personnel, and is intended to facilitate the various actions which are required in the course of the training. Figure 2 is a cutaway view of the building which illustrates the location of each area, and the type of equipment provided in those areas.

The attack centers of the two simulated submarines, the sonar rooms, and the program operators' stations are on the first floor of the building. The computer and simulation room is positioned directly above them, so that the periscope simulation can be integrated conveniently when it is added, and so that the leads between the simulators and the submarine equipment can be as short as possible.

Each of the two attack centers has a program operator's station associated with it. A classroom is located adjacent to each program operator's station so that the group being trained can meet after a problem to discuss the results or listen to the instructor's comments. Other rooms located on the first floor are provided to take care of central power supplies, additional laboratory area, housekeeping functions, and the like.

On the second floor, in addition to the computation and simulation equipment, rooms are provided for maintenance, stores, utilities, and the master program operator's station and auditorium.

Program Operator Station

The program operator station is arranged to simplify the process of simulation and control assigned to the program operator. A large console is provided, placed so the program operator can view the interior of the attack center, and facing a projection screen on which is plotted the courses of all the ships and weapons in the problem he is conducting. An additional station is provided for an instructor, who works directly with the crew being trained.

The program operator's console has several work areas, arranged for efficient utilization by the program operator and his assistant. The central area is the control keyboard, which has functional and numerical pushbuttons for reading data into the computer. It is also provided with numerical displays which provide exact information about the speed, range, course and the like of all the vehicles in the problem.

A weapon control panel is placed next to the keyboard so that the program operator can act as the torpedo room crew, loading the selected weapons and participating in the pre-firing and post-firing activities. To the right of the keyboard, a projector control panel is provided so that adjustments in the scale factor and origin of the projected display on the screen can be made. A communications panel is also provided so that the program operator can communicate with the other members of the training staff or with the students, depending on his particular role.

The assistant program operator complements the program operator's functions by observing the operation of the sonar and radar sets through repeaters, and by adjusting the simulators as required during the course of a problem. He is also provided with communications to facilitate his task.

The instructor will observe the actions of the crew in training, and either immediately by use of his communications system, or later in a classroom discussion, will comment and advise the crew on details of their performance. To assist him in his observation, a small readout panel on his console gives range and other data on any target he selects.

Attack Center

The attack center contains all equipment normally used during the course of an attack, including some navigational equipment and all sonar and radar repeaters. Other equipment in the work area is merely mocked up so that the attack center retains the general appearance of the actual submarine. All of the real equipment is activated and performs its functions in the same manner as in the actual ship.

The sonar room, which is mounted in a separate room of the building, also contains real and mocked-up equipment, with the real equipment realistically activated.

Master Program Operator Station

The master program operator station is similar to the program operator station, and uses many of the same type of panels. Controls which parallel those for each of the program operators stations are provided, with override features so that the master program operator can perform operations instead of the program operators.

The projected display for the master program operators operates in the same manner as for the program operators, but is displayed on the screen at the front of the auditorium, with sufficient size to be easily visible to both the master program operator and to the audience. The audience will be composed of other submarine crews and Naval officers who will be able to study the tactics being employed. Tracks of all targets, friendly submarines, and weapons will be displayed, with color coding and symbols to identify each track and important action. Origin and scale controls for this display are located in the MPO console as well as switches which permit the temporary deletion of selected tracks for illustrative purposes. A loudspeaker system is also provided so the master program operator can make comments or explanations.

System Block Diagram

The block diagram of Figure 3 illustrates the interconnection of the major equipment in the ASW trainer. The computer, its accessory equipment, and the input/output equipment are shown in their central position in the system. The program operator's controls and the master program operator's controls are at the left, while the fire control and simulation equipment are at the right.

All computations and most of the system control are performed by the Control Data Corporation 1604 general purpose digital computer. Selection of this computer and details of its operation are included in the discussion of the computer system. The computer communicates with the rest of the trainer complex through its three pairs of input and output buffers. Communication with any of the input/output equipment is asynchronous, but always under the control of the program stored in the computer, and the timing of this communication is one of the major design tasks in writing the program. The computer is designed to share the available communication time between the six input and output buffers, a feature which must be accounted for in system timing. The organization of the hardware which communicates with the computer is influenced by the timing of the data flow, and grouping of similar signals in the different buffers is necessary to reduce the complexities of program organization to a satisfactory level.

Buffers 1 and 2 are assigned to the computer accessory equipment, which includes the magnetic tape units, the typewriter and console. The computer reads data to these units in parallel, and selects one of the units to receive data by providing the correct code on the 12-bit external function lines. All units are designed to ignore the data unless they are selected.

The numerical display, which displays numerical data on the target behavior, is also connected to Buffer 2. The numerical display central equipment has an input buffer designed to accept information at a 200 KC rate from the computer and store it on a magnetic disk. The readout from the computer to the display is performed once per second, and requires from 50 to 170 milliseconds, depending on the activity of the other buffers during the readout period. The display takes the data from its storage, applies it to the character generator and deflection circuits, and distributes it to the three keyboards. Each of the keyboards contains nine 1.5 inch by 3 inch cathode ray tubes which display 90 "words" of numerical data. The generation and cycling rate is 30 times a second, which keeps the eye from sensing any flicker.

Input buffer 3 is assigned to the analog-to-digital converter, which reads in analog voltages from various parts of the system to obtain data for weapon trajectory computations and weapon scoring. In the sensory and fire control equipment, most of the accurate analog information exists as the angular position of synchros in two- and three-speed systems. By measuring the angles of these synchros, input data can be obtained.

The method selected for measuring these angles avoids the modification of standard equipment, as would be required if shaft encoders had been added, and instead samples the synchro stator voltages. Figure 4 shows the two voltages which are measured. When these voltages are digitized and fed into the computer, a program computes the synchro angle directly from these data. Measurement of each synchro of two- and three-speed synchro sets is made at adjacent points in the sequence, so that the first synchro will not turn before the remainder have been sampled. The highest speed synchro is sampled first, and the computer program is designed to take into account slight misalignments, either in the equipment, or because of the sampling delay.

A secondary use is made of the A/D converter, for automatically testing the whole system. In this mode the computer is programmed to send out special test information to all portions of the system, and the automatic test multiplexer sequentially samples test points located in all the external equipment. Signals from the multiplexer are digitized in the A/D converter and read into the computer. Within the computer, these voltages are compared with previously stored values, and if the difference exceeds a predetermined tolerance, which is also stored in the computer, the number of the faulty unit is displayed on the computer console and the test is stopped. The test may be restarted manually, and will run to completion or to the next faulty unit.

The majority of the outputs from the computer to the simulation equipment are provided via buffers 4 and 6. All information required 10 times per second is read out through buffer 4, while the once-per-second and twice-per-second information is distributed through buffer 6.

Data is read out via buffer 4 in blocks of three 48-bit words. A shift register in the buffer 4 interface equipment accepts the three words of each block from the computer, under program control. The three words are then distributed through gating circuitry to the storage elements, and the process is repeated until all words are read out. 490 words are read out each second, in ten intervals, each about 89 milliseconds long. At the end of the complete read out, the computer resets the shift register and gating logic to prepare the interface for the next second's block.

Two varieties of storage elements are used in the buffer 4 system. The first of these flip-flops, which are used to store data on target range and bearing for the radar and sonar simulators. The second elements are latching type reed relays, which are used to provide the data to the synchro voltage synthesizer outputs, which in turn feed the fire control systems.

The fire control systems receive inputs in the form of synchro signals from the sonar and radar sets, and will later also receive periscope inputs. In some modes of operation, these data must be fed directly from the computer, with no intervening manual actions. A computer-controlled servo driving an actual synchro has been used in some previous similar applications, to produce this type of output. Control of digital servos through the computer introduces many design complexities, so an alternate method was chosen.

The output of a synchro can be synthesized with a pair of A-C voltages whose relative amplitude is determined by the rotor position; hence a synchro can be replaced with a tapped transformer and a suitable arrangement for selecting the right taps. This method, termed a digital-to-synchro converter, was selected, and an assembly providing the tapped transformers and relay switching was designed. 120 of these converters are required, and are connected to buffer 4. The computer sets the relays in these converters each 0. 1 second, which is sufficiently often to provide the necessary accuracy on even the highest speed rotation of the synchros.

Data flow through the buffer 6 equipment is gated and controlled in the same general manner as for buffer 4, except that the information is provided only twice each second. The computer words are received by a shift register and distributed to storage elements. Since many of the outputs from buffer 6 are analog voltages, the buffer equipment includes digital-to-analog converters. These converters use reed relays as combined storage and switching devices, to obtain the analog output with the minimum of components. The analog outputs are fed to the sonar simulator, radar simulator and projected display equipment.

Computer System and Programming

The simulator operates in two basic modes, record and playback. The mode of operation is selected by the computer operator and verified by a keyboard command from the Master Program Operator. The above procedure is used to insure that all necessary operations are performed in the computer room prior to the start of a training exercise and to simplify the system executive program. The system executive routine verifies the read in of the overall program into core, initializes certain parameters and flags common to all modes of operation, and then branches on a mode command from the computer typewriter to the record mode executive or playback mode executive. Checks are included in the system executive to insure that paper tape and magnetic tapes required by the mode of operation are loaded prior to the read in of a mode command from the master program operator.

Except for the input data source, the record and real time playback modes of operation are very similar. In the record mode, vehicle control data inputs are provided either from Program Operator Keyboards or preprogrammed paper tape inputs. Data are also provided directly from the fire control system in both digital and analog form. In the playback mode of operation all vehicle and weapon control commands are provided by a previously recorded magnetic tape. During playback, direct keyboard commands are limited primarily to numerical display commands, typewriter print commands or system run-freeze commands.

Three submodes of the playback mode are provided; real time, twice speed, and four times real time. In the real time submode the radar simulator and sonar simulator are activated and the status of the training task may be observed on the radar and sonar consoles. In the second submode all vehicles and weapons are moving at twice normal velocity. This submode will be used primarily for the review of the tactics used in a previously recorded problem, and outputs are provided only to the projected display and numerical display systems. To provide sufficient time to make all control and basic computations twice per second and enter the projected display and numerical display routines once per second, the sonar, radar and other output routines are eliminated during this mode of operation. The elimination of the sonar, radar and other routines does not place a restriction on system operation, since the sonar and radar equipment cannot accept inputs to simulate certain vehicles moving at twice normal velocities. The third playback submode is identical to the second submode except that the vehicles are moving at four times normal velocities.

A change from one playback speed to another may be made in approximately 8 seconds. Although not a feature of the present system, the program is planned to allow one attack trainer to operate in the record mode and the other attack trainer in the playback mode with relatively minor changes in the program.

Computer Characteristics

During the early phases of system design, a Control Data Corporation 1604 Computer was chosen for the training complex. This computer was selected after a thorough analysis of the calculations which would have to be performed each second. The precision of the results, the total number of program steps, and the amount of input and output communication necessary were all considered in determining the size, organization and speed of the computer required For reference, the characteristics of the 1604 are summarized below.

a. Parallel binary machine with 48-bit word length.

b. 16,384 words of magnetic core storage.

c. Three 48-bit asynchronous buffer input channels.

d. Three 48-bit asynchronous buffer output channels.

e. A real time clock.

f. 62 Instructions including floating point, and logical and masking instructions.

g. 4.8 microseconds effective cycle time.

Record Mode of Operation

A simplified flow chart of the record mode of operation is shown in Figure 5. The overall computations cycle is one second, and with the exception of the numerical display routine and small portions of other routines, all computations are made each second. All input/output timing and data transfer are controlled by the executive or monitor routine. The program is planned so that all input data are read in and available prior to the time they are needed by any subroutine.

The simulator program contains three types of routines. The first type includes keyboard control, fire control system decode and other control routines. The second type is the basic routines which include the vehicle courses, the weapon courses and the relative coordinate routines. These routines compute all fundamental data required as a function of vehicle or weapon dynamics and provide all data needed by the output routines. The third type of routine consists of output routines, which are mutually independent and require only data computed by the control or basic routines.

The sequence shown on Figure 5 for output routines is dictated to a large degree by input/output timing requirements. In general, all output data required by a particular type of external equipment such as a radar or sonar simulator are computed by one routine. This concept has the advantage of allowing the program to be easily modified to accommodate new or different types of external equipment. The computer outputs to all external equipment are provided in a format which minimizes the amount and complexity of external equipment. Properly scaled DC and AC analog signals and digital signals are provided to the various output equipments. For example, the design of the sonar simulator was simplified by providing range data as a DC analog voltage and in digital form. All basic computations are performed in units of yards, degrees and seconds. If input data is available in other units the input data is converted before it is utilized by the basic routines. Similarly, the output data is scale factored and optimized for the particular type of output equipment.

In all modes of operation, sufficient time is allowed for the worst case branch for control routines and for the vehicle control and weapon trajectory computations. To keep within the one second time cycle certain output routines are eliminated under worst-case conditions. For example, the vehicle motions section of the numerical display routine is not computed during the second immediately after a weapon firing to allow time for numerical display weapon scoring computations. In addition, certain weapon scoring routines are computed over a period of several seconds. In the high speed playback mode of operation, certain output routines are eliminated to allow adequate time to loop through the control and basic routines four times. The read in and read out of data to external equipment and the characteristics of certain external equipment place additional constraints on the overall one second cycle. For example, a block of data must be read out to the projected display every 500 milliseconds ±10 milliseconds. If the overall one second computation cycle were exceeded, serious errors could occur in the operation of the projected display system. In addition, certain inaccuracies would occur in the vehicle and weapon computations.

In the present application it was necessary to minimize both the computation time and the memory required for each subroutine. During all phases of programming it was necessary to keep a continuous and up-to-date check of time and memory requirements. For some routines, memory was of utmost importance and in other cases time was the important factor. The location of the routine in the overall computation cycle, whether the routine was used in the high speed playback mode where computation time was of first importance, and other related factors determined the relative importance of time and memory. In many cases, a trade of memory for computing time had to be made. An example of this was the sine-cosine routine, which is used many times in the course of each second. By using 1440 words of memory to store a sine-cosine table, the time required for computations utilizing sine or cosine functions was reduced by a factor of 7, from 168 milliseconds to 24 milliseconds. To save both time and memory, a combination interpolation and table look-up method was used to compute e^{-x}.[1]

Careful attention was given to the assignment of memory. All input/output data blocks were grouped together in lower core to facilitate output memory dumps both during system debugging and later for maintenance purposes. Dump, trace and other service routines were grouped together in upper core. Most of this memory block can be considered spare, since the service routines could be called in individually from magnetic tape when needed. The memory block immediately below the service routines is reserved for future additions to the system. Spare blocks of memory were reserved between each subroutine to allow for minor changes or additions. All routines utilized in only one mode of operation were grouped together so that they could be called in individually from magnetic tape as a function of mode of operation. At the present time, sufficient memory capacity is available to store all routines in core. If the memory capacity or computation time is exceeded when future additions are made to the system, most of the output routines can be moved to a satellite computer, using the high speed input and high speed output channel, designed for communication with a satellite computer.

Record Mode Timing Cycle

Referring to the record mode flow chart, Figure 5, and the computer input/output timing sequence, Figure 6, at the start of each second the real time clock is tested. If the system is not frozen, keyboard commands, paper tape input commands or input commands from the fire control system are decoded and all control signals required by other routines are set. The fire-control input synchro data are then decoded and placed in floating point form for use by the weapons and the weapons scoring routines. At this point in the computation cycle all control information needed by other routines is available. All basic own ship and target parameters such as climb/dive, turn, ordered course and similar data are computed and an integration is performed to provide X, Y, Z coordinate data for the vehicles for the current second. The weapon routine then computes the trajectories of torpedoes and other submarine attack weapons and performs scoring computations to determine whether a hit or miss has occurred. During the second immediately after a weapon has been fired, gyro angle, course, speed, and other data for the true fire control solution are computed by the weapon routine. The true fire control solution data are displayed on the numerical display for comparison with the actual inputs from the fire control system, to aid in the evaluation of crew performance.

All relative range and bearing computations both present and apparent between the two own ships and all other vehicles and weapons are computed by the relative routine. At this point in the computation cycle all basic parameters required by the output routines are available. Range and bearing data properly scale-factored and other data required by the periscope simulator are then computed and read out to the periscope simulator via buffer 2. To meet the dynamic requirements of the radar and sonar equipment and certain portions of the fire control equipment, range and bearing data in synchro form is read out every 100 milliseconds. Synchro data for the next second is computed at this time by the Synchten routine and the first 100 millisecond block of data is stored in output memory for readout during the first 100 milliseconds of the next second. Since data are read out continuously via buffer 4 during 89 milliseconds of each 100 milliseconds, data for the other 9/10 of a second are stored in the buffer 4 output memory block during the first 100 milliseconds of the next second while the first 1/10 second block of data is being read out.

During certain phases of training, operators are not provided at the radar and sonar equipment, and range and bearing outputs from the computer are provided directly to the fire control system. These direct outputs are more accurate than would normally be obtained from the radar or sonar systems. To provide realistic simulation under these conditions, pseudo random errors in both range and bearing are added to the radar and sonar output to realistically simulate the action of a human operator at the radar and sonar consoles.

The next routine, the projected display routine, computes properly scale factored and offset X and Y coordinate data for each operator's projected display system. The data includes the X and Y coordinates of all vehicles and weapons and certain alpha-numeric data which is displayed adjacent to the projected display tracks when a weapon hit, a sonar contact, a radar contact or any other noteworthy event occurs. When the projected display routine is completed, buffer 6 is activated and the projected display data is read out via buffer 6 at the rate of approximately one word per millisecond.

The magnetic tape recording routine is then entered and all data needed to later recreate a problem in the playback mode of operation are recorded. To minimize errors in recording and playback of magnetic tape data the following method is utilized. A block of data is recorded on magnetic tape each second. If parity or word length errors occur, the tape unit is backspaced and the block rewritten with a "flag" in the block to indicate that the block should not be utilized during playback. The block is written again on the next section of magnetic tape. If parity or length error occurs during the second recording of the blocks, the tape unit is backspaced and the block recorded a second time in the same position on the magnetic tape. A third error operates an indicator, but is otherwise ignored. Vehicle control commands, weapon firing commands and other critical data are recorded three times in three separate locations in each data block. During the playback mode of operation the three blocks of data are compared and if the three words are not identical the two out of three which agree are utilized.

The methods outlined above are necessary to insure that the playback of a training exercise is not invalidated by an error in the data read back from magnetic tape. For example, if early in a training exercise a command was recorded to change the course of a vehicle from zero to 180 degrees the remainder of the problem could be meaningless if this command was not read in correctly from magnetic tape and the vehicle moved off in the wrong direction. In the rare case when all three input words disagree during playback, means are provided for the training director to freeze the system and manually reinsert an appropriate command.

Radar attenuation functions required by the radar simulator are then computed and the data stored for later readout. Data unique to the sonar simulator are then computed and stored for readout. All the above data are read out via buffer 6. Close timing coordination is required to insure that all data are available prior to the time the particular word is read out. A computer test routine is then entered, which over a period of several seconds checks all computer instructions, index registers and other features of the CDC 1604 computer.

The numerical display routine is then entered. This routine provides data in BCD format via buffer 2 to the numerical display system. The data computed includes the range and bearing between each own ship and all other vehicles. Course, speed and X, Y and Z coordinates are presented for each own ship and two selected targets, and certain scoring information required for fire control crew performance evaluation is also displayed.

The typewriter print routine is then entered. Once each minute, this routine prints out task time for each of the two tasks, the range and bearing between each own ship and one selected target, and any computer system errors which may have occurred during the previous minute. Since the typewriter time shares buffer 2 with the magnetic tape system, the numerical display system and the periscope system, only 500 milliseconds are available each second for typewriter output. To allow an adequate safety margin for variation in typewriter characteristics, only three characters or a carriage return, or an upper/lower case shift and one character are printed out each second. To provide a continuous monitor of system operation a number of tests are made to check the validity of input data, the action of external logic attached to the buffers, the real time clock and other portions of the computer system. If any fault has occurred during the previous minute, the type of fault is printed out on the typewriter. In addition, "trouble counters" are set to give a continuous tabulation of the numbers of times that each type of malfunction has occurred, for use in maintenance.

Conclusion

The foregoing discussion outlines the requirements for a specific real time computer and simulation system, and describes the techniques applied to the design of this system. Many details such as descriptions of the simulation equipment which could have been included were omitted for the sake of brevity, but the essential interrelation of the computer and the external equipment has been given the emphasis it requires. Other approaches to the task could have been employed, especially in the organization of the computer program, but it is believed that the method described approaches the optimum with the particular constraints which were placed on the design. It is hoped that some of the approaches described may be of use in the design of similar systems.

The authors wish to express thanks to their associates at Aircraft Armaments who have provided many of the design concepts and performed much of the work on the system. Without their efforts, this paper would not have been possible.

References

1. Putting a Hex on e^x. W. Fenozeig. Communications of the Association for Computing Machinery. Vol. 4 No. 9, September 1961.

2. Handbook of Automatic Computation and Control. Volume 2. E. M. Grabbe, S. Ramo, D. E. Wooldridge, John Wiley and Sons.

3. Control Programming - Key to Synthesis of Efficient Digital Computer Control Systems, by A. S. Robinson. AIEE Transactions, Part 2 (Applications and Industry). March 1960, pages 475-502.

4. UDOFT Simulation Program. Final Report, by John Prutsalis, Sylvania Electronic Systems, Needham, Massachusetts.

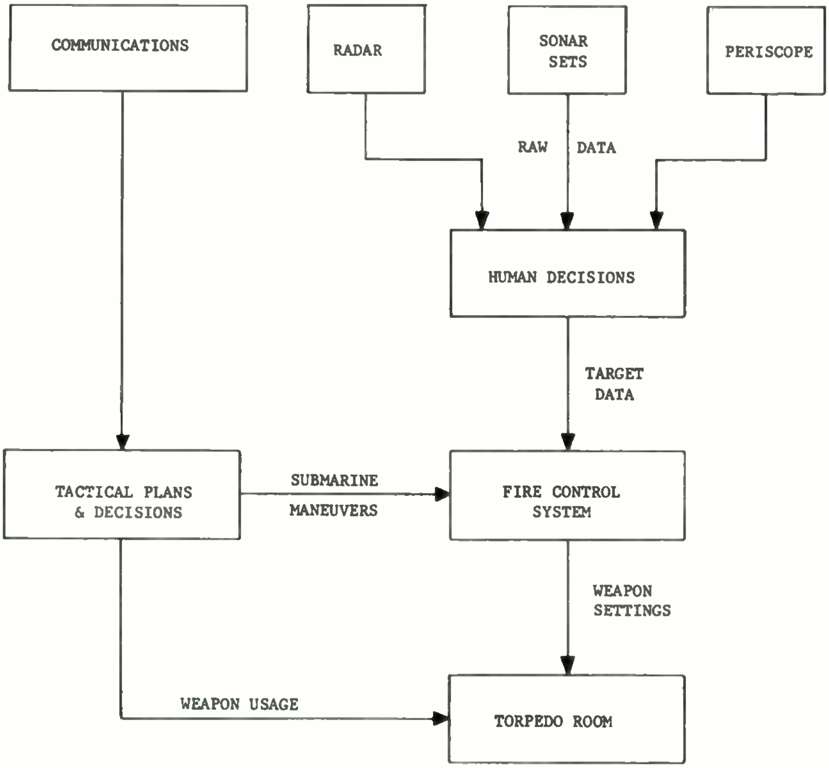


Figure 1

BLOCK DIAGRAM ATTACK CENTER ORGANIZATION

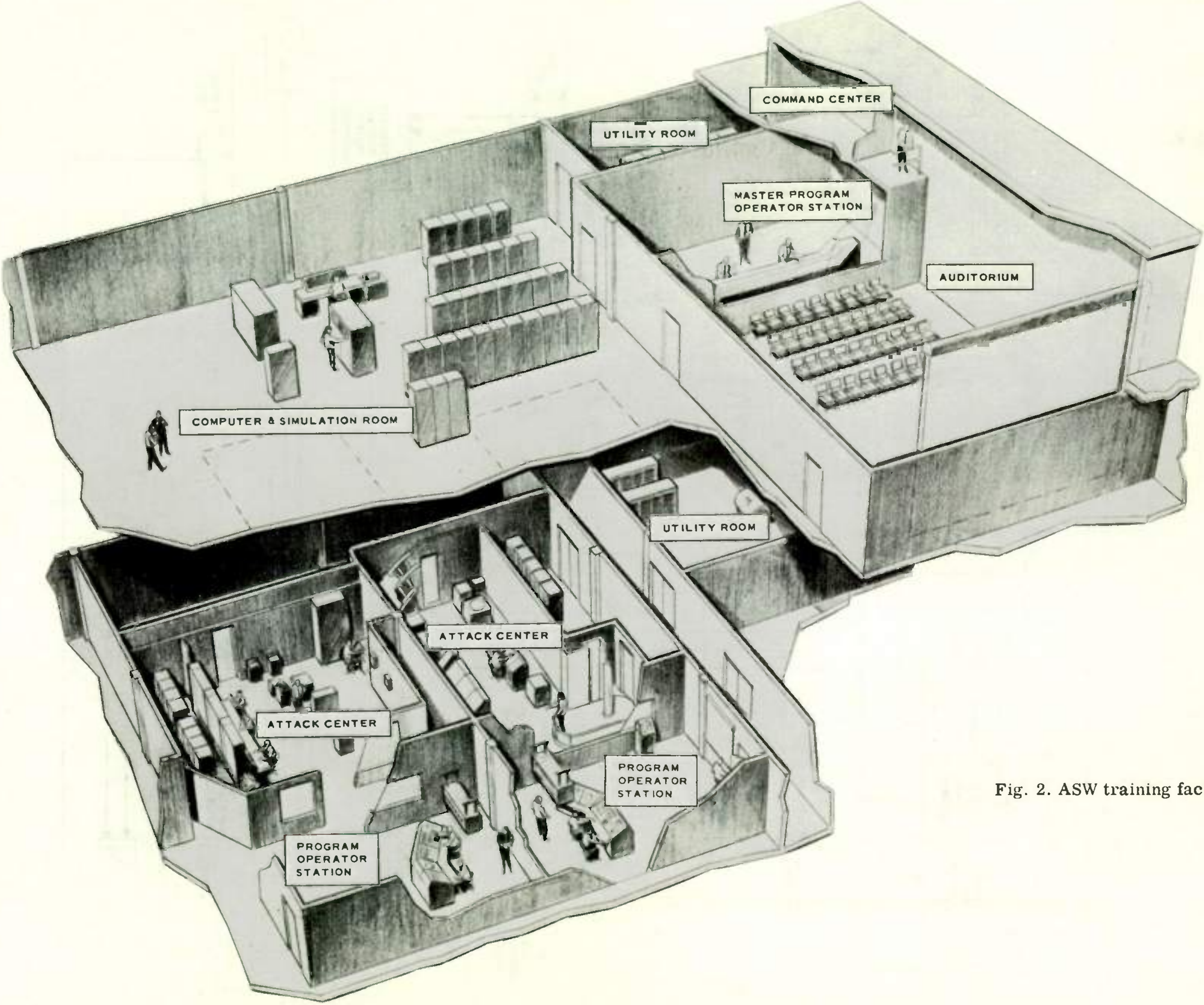


Fig. 2. ASW training facility.

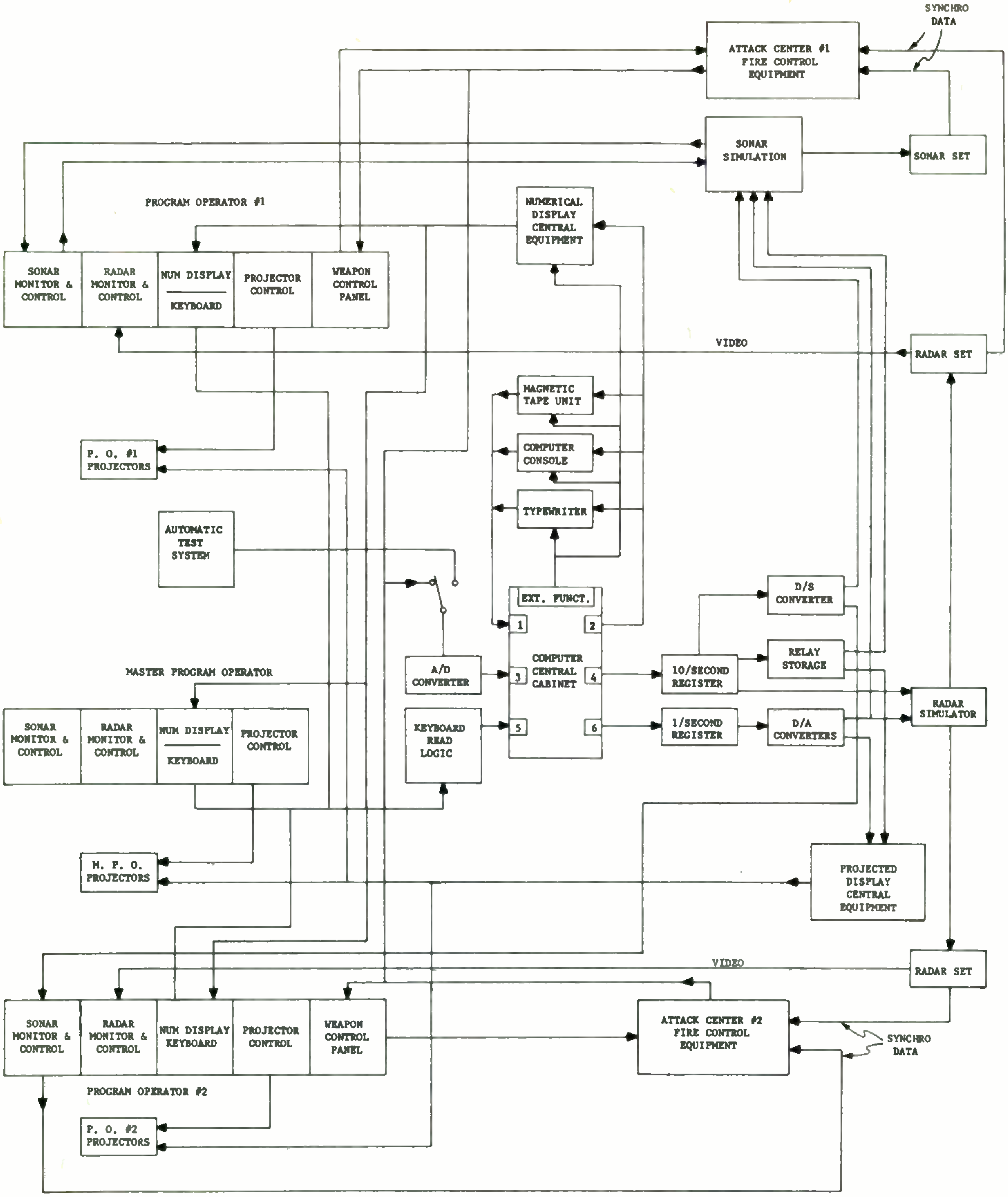


Figure 3

ASW TRAINER BLOCK DIAGRAM

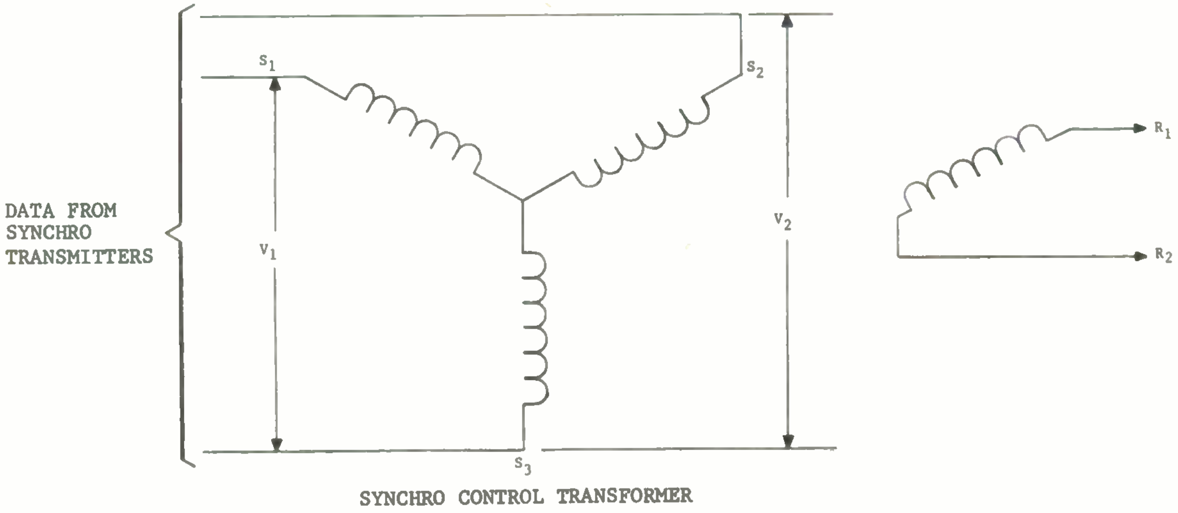


Figure 4

SYNCHRO VOLTAGE MEASUREMENT

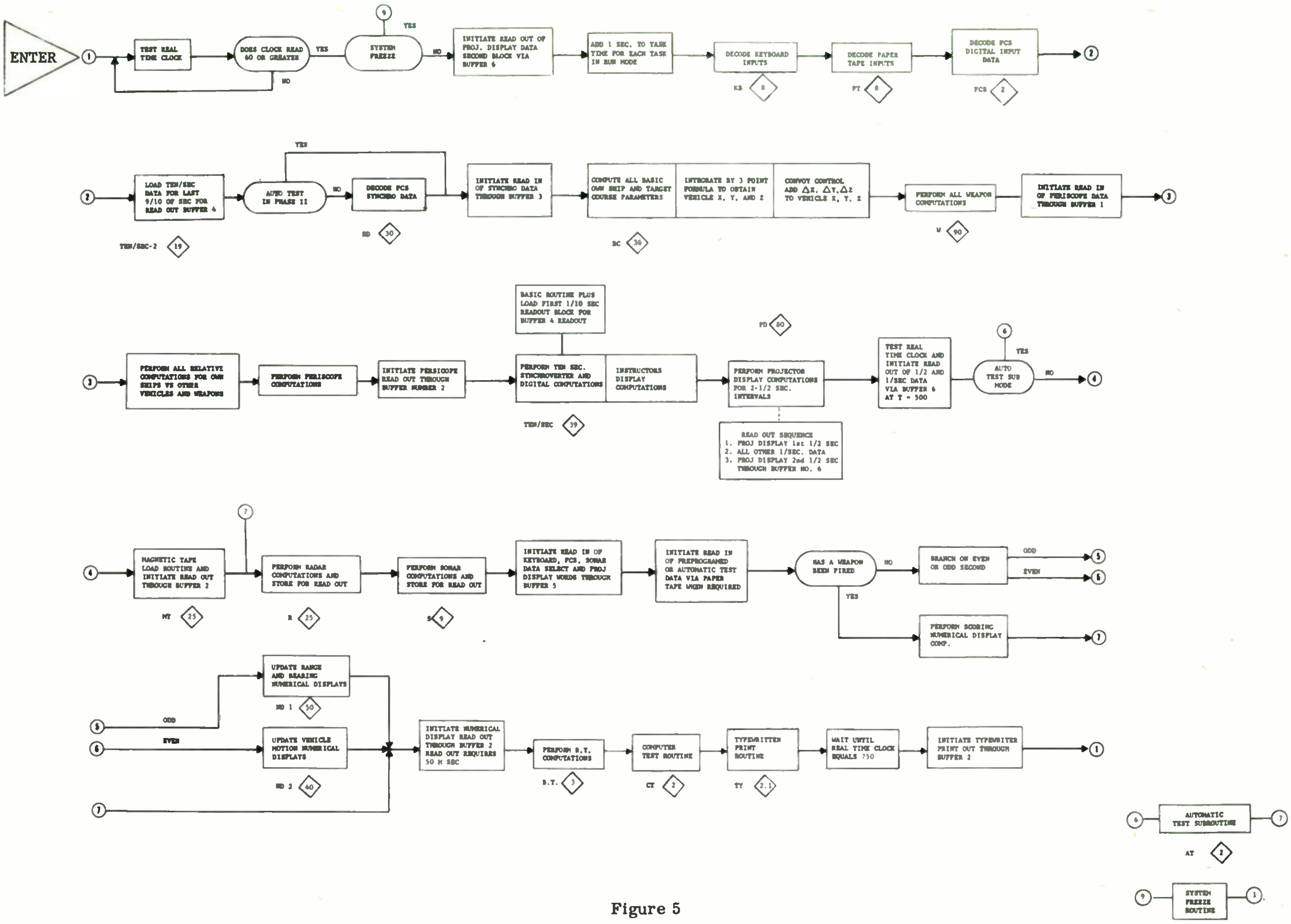

ENTER
1
TEST REAL TIME CLOCK
DOES CLOCK READ 60 OR GREATER
YES
NO
9
SYSTEM FREEZE
YES
NO
INITIATE READ OUT OF PROJ. DISPLAY DATA SECOND BLOCK VIA BUFFER 6
ADD 1 SEC. TO TASK TIME FOR EACH TASK IN RUN MODE
DECODE KEYBOARD INPUTS
KB 8
DECODE PAPER TAPE INPUTS
PT 8
DECODE FCS DIGITAL INPUT DATA
FCS 2
2
2
LOAD TEN/SEC DATA FOR LAST 9/10 OF SEC FOR READ OUT BUFFER 4
TEN/SEC-2 19
AUTO TEST IN PHASE II
YES
NO
DECODE FCS SYNCHRO DATA
SD 30
INITIATE READ IN OF SYNCHRO DATA THROUGH BUFFER 3
COMPUTE ALL BASIC OWN SHIP AND TARGET COURSE PARAMETERS
INTEGRATE BY 3 POINT FORMULA TO OBTAIN VEHICLE X, Y, AND Z
CONVOY CONTROL ADD ΔX, ΔY, ΔZ TO VEHICLE X, Y, Z
BC 36
PERFORM ALL WEAPON COMPUTATIONS
W 90
INITIATE READ IN OF PERISCOPE DATA THROUGH BUFFER 1
3
3
PERFORM ALL RELATIVE COMPUTATIONS FOR OWN SHIPS VS OTHER VEHICLES AND WEAPONS
PERFORM PERISCOPE COMPUTATIONS
INITIATE PERSICOPE READ OUT THROUGH BUFFER NUMBER 2
BASIC ROUTINE PLUS LOAD FIRST 1/10 SEC READOUT BLOCK FOR BUFFER 4 READOUT
PERFORM TEN SEC. SYNCHROVERTER AND DIGITAL COMPUTATIONS
INSTRUCTORS DISPLAY COMPUTATIONS
TEN/SEC 39
PD 80
PERFORM PROJECTOR DISPLAY COMPUTATIONS FOR 2-1/2 SEC. INTERVALS
READ OUT SEQUENCE
1. PROJ DISPLAY 1st 1/2 SEC
2. ALL OTHER 1/SEC. DATA
3. PROJ DISPLAY 2nd 1/2 SEC THROUGH BUFFER NO. 6
TEST REAL TIME CLOCK AND INITIATE READ OUT OF 1/2 AND 1/SEC DATA VIA BUFFER 6 AT T = 500
6
YES
AUTO TEST SUB MODE
NO
4
4
MAGNETIC TAPE LOAD ROUTINE AND INITIATE READ OUT THROUGH BUFFER 2
MT 25
7
PERFORM RADAR COMPUTATIONS AND STORE FOR READ OUT
R 25
PERFORM SONAR COMPUTATIONS AND STORE FOR READ OUT
S 9
INITIATE READ IN OF KEYBOARD, FCS, SONAR DATA SELECT AND PROJ DISPLAY WORDS THROUGH BUFFER 5
INITIATE READ IN OF PREPROGRAMED OR AUTOMATIC TEST DATA VIA PAPER TAPE WHEN REQUIRED
HAS A WEAPON BEEN FIRED
NO
YES
BRANCH ON EVEN OR ODD SECOND
ODD
5
EVEN
6
PERFORM SCORING NUMERICAL DISPLAY COMP.
7
5
ODD
UPDATE RANGE AND BEARING NUMERICAL DISPLAYS
ND 1 50
6
EVEN
UPDATE VEHICLE MOTION NUMERICAL DISPLAYS
ND 2 40
7
INITIATE NUMERICAL DISPLAY READ OUT THROUGH BUFFER 2 READ OUT REQUIRES 50 M SEC
PERFORM B.T. COMPUTATIONS
B.T. 3
COMPUTER TEST ROUTINE
CT 2
TYPEWRITTER PRINT ROUTINE
TY 2.1
WAIT UNTIL REAL TIME CLOCK EQUALS 750
INITIATE TYPEWRITER PRINT OUT THROUGH BUFFER 2
1
6
AUTOMATIC TEST SUBROUTINE
7
AT 2
9
SYSTEM FREEZE ROUTINE
1

Figure 5

MASTER EXECUTIVE FLOW CHART RECORD MODE

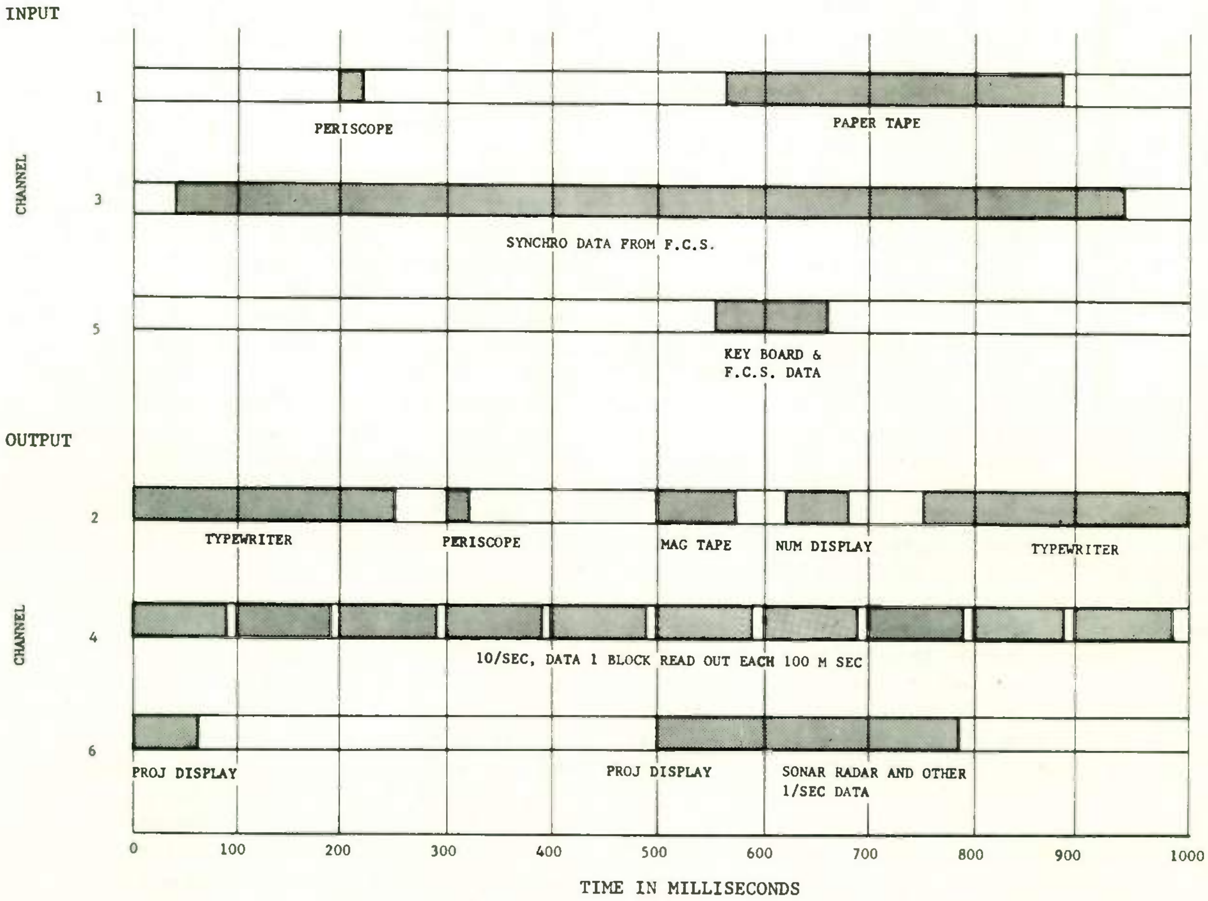


Figure 6
COMPUTE INPUT-OUTPUT TIMING RECORD MODE

THE ITERATIVE CONTROL SYSTEM FOR THE ELECTRONIC DIFFERENTIAL ANALYZER

M. C. Gilliland
Beckman/ Berkeley Division
Richmond, California

Summary

A general discussion is presented regarding new applications for the iterative differential analyzer. The control system is described in some detail. The past development, advantages and disadvantages of the electronic differential analyzer are reviewed. It is shown how the iterative computer surpasses the older non-iterative machine. Some comment is given regarding the limitations of the new computer and also its relationship to digital computers.

The Electronic Differential Analyzer

The EDA (Electronic Differential Analyzer) has enjoyed extensive usage in the computing community. This type of computer has been particularly suitable for the solution of systems of ordinary differential equations. The EDA has produced answers to complex problems in a relatively economical and convenient manner. Because of its ability to integrate directly and its parallel operation, it has been a faster and less expensive machine for the implementation of simulation or optimization studies. Communication with the computer through a visual or graphical data presentation system frequently allows the user to make an immediate evaluation of results. Consequently, it has been a tool without equal for certain types of mathematical analysis.

Major Advantages

Historically, much of the motivation for the development of analog computers of various types derived from the capability of these machines to integrate directly, thus avoiding the difficulties inherent in numerical integration. These computers also were innately parallel machines. That is, all computation proceeded simultaneously as contrasted to the serial operation of most digital computers. This feature, together with an integration capability, resulted in a considerable reduction of computer time necessary for the solution of many problems. In many instances the machine could be made to operate in "real time", thus permitting its interconnection with physical hardware. One of the prime examples of this type of application is the use of analog computer subsystems in cockpit simulators.

Of these machines, the EDA has undergone the most improvement and has become by far the most widely used type of analog computer. Two types of EDA's have been available. The more common of these is a machine with relatively narrow bandwidth (about 100 cps) but with precision computing elements (.01 to .02 percent for linear equipment). The other is known as a high speed repetitive operation analog computer. It has much greater bandwidth (100 kc to 1 mc) but lacks the d-c accuracy required for many types of applications. The latter computer generally solves the same problem repetitively requiring only a very short time for each solution. For many parametric studies it is an extremely powerful tool.

Some Disadvantages

The range of application of the EDA has been restricted by the lack of a flexible control system. Due to the strong influence exerted on computer design by certain areas of the aircraft and missiles industry, some natural applications for this type of machine have been either overlooked or not sufficiently emphasized. During past years, emphasis has been placed on component accuracy rather than flexibility of control. Although improvement of the d-c accuracy and reliability of the EDA is desirable, the lack of flexible control has prevented the broadest use of hardware which has been developed. Some of the more pertinent of these limitations are noted here.

The EDA has had no convenient memory. Thus, computational results could not be stored in an automatic way for future use in decision making or as input data for subsequent calculation. In addition, the lack of memory preempted the use of numerical techniques to augment the already existing repertoire of mathematical functions.

The control system of the EDA is of the synchronous type; that is, all computation is performed simultaneously at the same rate. Thus, the machine cannot simulate sampled-data systems or hybrid sampled and continuous data systems. More generally, any sub-computation which is of closed form cannot be introduced into the main computation.

The EDA control system, because of its synchronous nature, creates unnecessary problems when it is integrated with hybrid (analog-digital) interface equipment. Multiplexing rates must be high or computation must be interrupted if sample simultaneity is important. As will be discussed later, the number of channels required for interface equipment is usually higher than should be necessary.

The synchronous control system largely obviates a requirement for both precise, limited bandwidth and less precise broad bandwidth computation equipment in the same machine. In a synchronous machine an attempt is made to balance the computational bandwidth of the various computing modules. Thus, as noted above, a high-speed repetitive operation computer was developed to solve problems where speed rather than accuracy was desired. It is apparent that if the restriction of control synchronism is removed, then it is advantageous to have both types of equipment available in the same machine.

An Asynchronous Fine Control Structure

Fuller realization of the innate capabilities of the EDA obtains, if an asynchronous control structure is used. A control system of this type was first introduced by one manufacturer[1] about 1958. Since that time, the concept has been developed further, resulting in what is called an iterative control system. The terminology derives from the capability of the machine to perform calculations which are iterative in a mathematical sense. The control structure is better described as asynchronous and "fine". The result, however, is that the machine is called either an IDA (iterative differential analyzer) or an iterative analog computer.

In an IDA, as in an EDA, there are various arithmetic computation modules as well as several basic normal machine modes. The operating modes are Standby, IC, Hold and Compute. A little reflection will show that all arithmetic modules are mode independent except for integrators. Thus, asynchronous operation of a machine largely centers around integrator mode control. In an IDA the integrators are controlled independently in separate groups. Because each of these groups contains only a small number of integrators, the control system is said to have fine structure.

Memory can be obtained by means of integrators. An integrator can "learn" or store information in the IC mode and "remember" or hold the information during the hold mode and even the compute mode if it has no normal computation input. It can be seen that asynchronous control alone will not provide memory. If all integrators are reset simultaneously, no information can be stored for subsequent use. Consequently, the control system must contain complementary logic in addition to asynchronous structure in order for integrators to be used as an analog memory. Integrators under the control of complementary logic can store information while those under normal logic are holding information and conversely. This type of logic and control will be described later in more detail.

Applications of the IDA

The IDA control system is an extension of that for the EDA. Consequently, the IDA can be used as an EDA. Any problem which can be solved with the EDA can also be solved with the IDA. The converse is not true. EDA applications are relatively well-known and the discussion herein will be limited to those which are added as a result of the iterative control system.

Since the IDA has been in use only a short time, it is premature to compile a lengthy list of specific applications. The capabilities of the machine are best illustrated by discussion of its general types of usage. A problem of particular interest is the simulation of combined sampled and continuous data systems. Such systems frequently will contain several sampled-data subsystems operating with different sample rates or on command from some other subsystem. Simulation of the overall system can be implemented only if the computer has an asynchronous control system with sufficiently fine structure. Some guidance systems contain a sampled-data subsystem which services several continuous data systems. This process can also be represented with an IDA since the computer has the necessary command-controlled switching in conjunction with its asynchronous structure.

Perhaps one of the most important features of the IDA control system is its ability to be integrated in a hybrid system with less difficulty than the EDA. The use of memory as a buffer simplifies the interface system. The integrators in the IDA can be used as sample and hold devices for storing information to be multiplexed into the digital computer. Analog memory can provide a simultaneous set of sampled values, thereby reducing the multiplexor rate in some instances. Only one D/ A converter is required whose output can be multiplexed into the analog memory used as a buffer.

An even more important aspect of the IDA is that all of the control structure can be operated in a simple manner from a switching matrix which is readily adaptable to commands from a digital computer. Thus, the digital computer can completely control the analog computer, including the selection of the fine asynchronous structure of the machine.

Independent domain control and specification, iteration rates, sector integration rate specification, computing module transference selection, internal analog logic path selection, main machine mode control, and many other commands can be given to the IDA through a switching matrix. This switching matrix does not accompany the standard IDA, since a pinboard is generally used to perform this function.

An additional side feature of this type of control is that a complex problem can be checked in sections because the integrators can be disabled by pairs from a pinboard. This and many other checkout conveniences are a natural result of the type of control used in the IDA.

The Iterative Control System

In order to explain the foregoing in more detail, a capsule description is given of the control system. As noted above the IDA differs from the EDA only with respect to control capability. Both the EDA and IDA contain integrators, summers, multipliers, dividers, function generators and decision making modules as arithmetic elements. An integrator under complementary logic or C-integrator differs from the normal or N-integrator only with respect to the mode condition. The C-integrator generally finds application during iterative operation of the computer. Figure 1 shows the iterative operation mode sequence of the IDA for both the N- and C-integrators. Note that whenever an N-integrator is in the compute mode a C-integrator is in the initial condition mode and conversely, with the exception of the first cycle. Obviously the C-integrator cannot remain in the compute mode when the IDA is statically reset, since initial conditions which are established for the N-integrators may cause the C-integrators to overload. Thus, initially when the N-integrators are in the initial condition mode, the C-integrators are in the hold mode. When manual mode selection is used the C-integrators may be reset in the pot set mode. A standby tablet switch is provided so that the N- and C-integrators may be put in the initial condition and hold modes respectively when the computer is operated manually. For both iterative and non-iterative operation, machine control nomenclature is based on the N-integrator mode condition with the exception of the special standby mode. From the mode logic it is seen that during iterative operation, data computed by the N-integrators may be used as initial conditions or inputs for the C-integrators, and conversely.

If the only input to an integrator is a time varying voltage applied to the IC terminal, then the integrator will learn or track this voltage during the initial condition mode and hold or retain its value during the compute or hold modes. Thus, an integrator may be used as a memory. The mode duty cycles for the N- and C-integrators are shown again in figure 2. It is seen that these integrators may be used as normal and complementary memories with the proviso that no input be applied to the resistors connected to the summing junction. Thus, as shown in figure 2, the N- and C-memories are essentially either in the initial condition or the hold modes. Note that as stated previously, all other arithmetic modules contained in the IDA are mode control independent. For example, no distinction can be drawn between a summer and a complementary summer.

Figure 3 shows a simplified schematic diagram of an operational amplifier under integrator logic together with its associated networks and control relays. The required relay duty cycles are selected automatically to provide normal or complementary logic for the integrators. The IJ terminal (which appears on the patchboard) may be connected to a spare resistor network to provide a capability for summing inputs to an integrator used as a memory. A solid state module can be used to provide high-speed IC capability.

It is frequently desirable to replace the I, C, $\overline{C}$ relays with solid-state switches to implement calculations with a high iteration rate.

A number of physical processes are characterized by a mathematical model which requires the cyclic solution of two or more interdependent sets of differential equations. A simple example is a model of the heart. Simulation of the diastole depends on input data which were computed during the previous systole. Each systolic response in turn uses information from the previous diastolic response. Thus, the computer is required to operate in two separate sections, each providing input data for the other on an alternating basis. The IDA control system together with its capacity to provide memory is conveniently applied to this type of problem.

Other mathematical models result in the redundant use of computing elements if the calculation is performed on a machine with a synchronous control structure. In particular a number of unilateral sub-computations are required which have identical form. The IDA can use a group of computing modules to perform a single calculation, and store the results to be used as input data for the next. Distillation column models which depend on tray-by-tray computation will exhibit this behavior. An asynchronous machine with memory can be used to perform the required calculations successively down the column, using an iterative procedure to match boundary conditions.

Frequently the mathematical statement of a problem to be solved on a computer will contain some dependent variable defined in terms of another in closed integral form. The EDA has difficulty integrating with respect to a dependent variable. However, with the IDA the integral can be calculated using time as a dummy variable where the limits of integration are determined from the desired value of the input dependent variable. The extension of this technique is a valuable computation aid, but generally depends on the accuracy required which in turn is a function of the computational bandwidth of the equipment.

Some mathematical problems are best solved using iterative techniques. Iteration involves the determination of a set of values which in turn are used as input data for the next calculation. The procedure is continued until the process converges or satisfies some criterion. The use of an analog computer for this type of application depends on the availability of memory as well as asynchronous structure to some extent. Automatic optimization and parameter determination frequently make use of this technique.

A closely related area of interest is adaptive control system analysis. Here, the IDA can be an invaluable simulation tool. Adaptive system studies have been performed with the EDA but only in an uneconomical and complicated manner.

The IDA can perform numerical calculations because it has algebraic computation modules, as well as memory. Thus, a mathematical model of a hybrid nature having both numerical and differential form can be solved with this machine. It should be noted however, that the scope of numerical computation which can be encompassed is rather limited. Extensive numerical calculation should be carried out with a digital computer, perhaps used in conjunction with the IDA. The accuracy of linear analog computing modules is at best only .01 percent of full scale and the non-linear modules are generally worse. Difficulties which result from roundoff and truncation error will be many times magnified for the analog computer as compared with the digital computer. However, an interesting and novel application of the IDA is the study of the stability of numerical processes. Due to the fact that the error is worse step-wise and machine time is relatively inexpensive, much can be determined about a digital process such as numerical integration in a relatively economical and convenient manner.

IDA vs. The Digital Computer

The IDA is not a panacea for the analog computer user who requires the use of memory or hybrid techniques. There is a break-even point beyond which the digital computer is more suitable from either an economic or practical standpoint or both. For example, although the IDA has memory, the "cost per storage location" is relatively high. If a few hundred values must be stored simultaneously, it is less expensive to use a digital computer with a suitable interface system. Of course, the development of less expensive analog memory would shift the break-even point somewhat.

When an extensive number of calculations must be performed which are best implemented with numerical techniques, again, a hybrid system can be justified on an economic basis.

Solid-state, two-terminal switches are available with the IDA system for this purpose. If it is desirable to increase the bandwidth of some of the integrators, special equipment can be obtained.

For the iterative operation system a hold interval is provided between each pair of initial condition and compute intervals. This is done to allow sufficient time for information transfer between N- and C-integrators and to avoid relay tracking difficulties. The holdinterval is chosen to be ten times the initial condition mode time constant of the integrator. If an integrator is used as a memory and is required to retain information throughout a considerable number of iterative operation cycles, then integrator drift can become a problem. Sometimes drift leads to seriously inaccurate computation due to the numerical character of the mathematical model being implemented. Consequently, it may be desirable to use a larger than normal feedback capacitor for the memory integrator with the same IC time required. For the accurate transfer of information under these conditions, a different type of initial condition circuit must be used.

Control of the IDA is implemented with four units which are the CEDAC (Central Differential Analyzer Control, figure 4), the IDAP (Iterative Differential Analyzer Pinboard, figure 5), the IDACON (Iterative Differential Analyzer Control, figure 6), and the IDAS (Iterative Differential Analyzer Slave, figure 7).

Manual machine operation is implemented with the CEDAC. This unit is an expanded version of the familiar EDA control unit. The Standby switch places N- and C-integrators in IC and Hold modes respectively. The Initial Condition, Hold and Compute switches place the N-integrators in the mode indicated and the C-integrators in the complementary mode. These switches allow the machine to be used as an EDA. They can also be used to "single step" the computer through an iterative program.

The choice of control logic for various computing elements is made with the IDAP. A layout of the pinboard is shown in figure 5. Multipliers are used as multipliers, dividers, square root devices or velocity sine/ cosine generators depending on whether or not a pin is inserted in the board at the appropriate location. The function generators can be converted to 20 segment units by inserting pins in the function generator area, thus freeing two amplifiers for other computation use. Resolver mode is selected by appropriate pin location. When no pin is used for a particular resolver, the multipliers and sine/ cosine generators are available independently.

The standard 2133 IDA contains 72 integrators which are grouped in 6 domains on the patchboard. The integrators in each domain are placed under the control of either the IDACON unit or one of five IDAS units (to be described later) by the appropriate location of pins on the pinboard. In order for a particular integrator to be under the control of either the IDACON or IDAS specified for its domain, a pin must be inserted in the IO location for that integrator in the mode control area of the pinboard. If the IO location is not pinned, then the corresponding integrator is subject only to manual mode command from the CEDAC. The function of integrator pairs is determined by pin location in the function control region of the board. An integrator pair can be used as N-integrators, C-integrators or high-gain amplifiers by inserting a pin in the $\int$, $\overline{\int}$, or HG locations respectively. If no pin is used, then the pair acts as summers.

The summer-high gain amplifier pairs are converted from the summer to the high-gain function by inserting pins in the function control area of the board for these amplifiers.

Access to the coils of the C, $\overline{C}$, I relays for the summer-integrator pairs is obtained with pincords inserted in appropriate locations in the coil control area. Coil terminals can be brought out to the patchboard through P-trunks from the pinboard. If no pin is used the relay coils are connected to the control busses and undergo normal operation. Function switch and comparator relay contacts on the patchboard are duplicated on the pinboard.

Iterative operation is implemented with the IDACON and IDAS units. The IDACON provides the basic timing for the IDA. This unit (figure 6) sets the hold interval equal to the time base which may be 0.001, 0.01, 0.1 or 1.0 seconds for the integrators under its control. The initial condition and compute intervals may be set independently by means of thumb switches to an integral multiple of the time base ranging from 1 to 99. The iterative operation switch initiates the iterative operation cycle through a master clock. The compute stop and IC stop switches command the computer to remain in the hold mode at the end of the compute and initial condition modes respectively. This provides a convenient method of temporarily interrupting an iterative computation to observe or record voltages in the computer. The compute coincidence switch accepts a hold command from an external source and bypasses the internal hold-after-compute command. The IC coincidence switch performs the same function with respect to the hold-after-IC command. The I, H, C lights provide a continuous indication of the mode condition of the IDACON. The external position of the time base switch allows the IDACON to be driven from an external oscillator, voltage to frequency converter or some arbitrary series of pulses, instead of its own internal clock. The external drive frequency cannot exceed 100kc. Broad bandwidth equipment can be used with the IDA under the control of the IDACON or one of the IDAS units. Delayed compute and initial condition command busses are available on the pinboard.

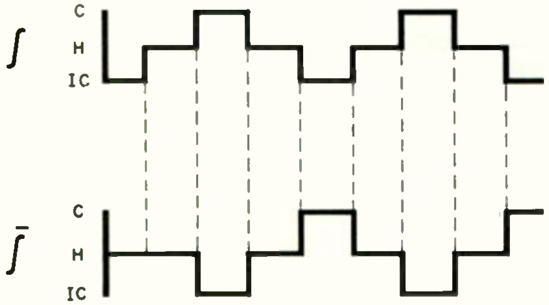
Generally, it is desirable to have available various iteration rates within the same computing system. In order to provide other iteration rates, use is made of IDAS units (figure 7). These slaves are similar in appearance to the IDACON. Each slave can be driven by the IDACON. The IDACON can drive any number of slaves. The slave has a time base selector switch which supplies the IDAS unit with appropriate timing pulses from the IDACON.

The IDAS time base multiplier thumb switches set the slave time base equal to an integral multiple, ranging from 1 to 99, of the time base switch setting. The slave IC and compute interval thumb switches independently set these intervals to an integral multiple of the composite slave time base ranging from 1 to 99. The IDAS coincidence switches and the external position of the time base switch operate in the same manner as those of the IDACON.

An automatic time scale option is provided which speeds up the computation by a factor of ten. The IDA has a far more flexible integrator capacitor selection system under time scale control then did the EDA. For the IDA, integrator capacitors may be switched for each amplifier separately, if required. Additional special capacitors may be added as desired. Integrator capacitor selection is accomplished either automatically or with the IDAP. The standard pinboard arrangement for the manual capacitor selection is shown in figure 5. As can be seen, any of four different capacitors can be selected arbitrarily for certain amplifier groups depending on the time scale condition. Without pinning, the integrator capacitor is automatically selected by the IDACON or CEDAC units in conjunction with the time scale switch.

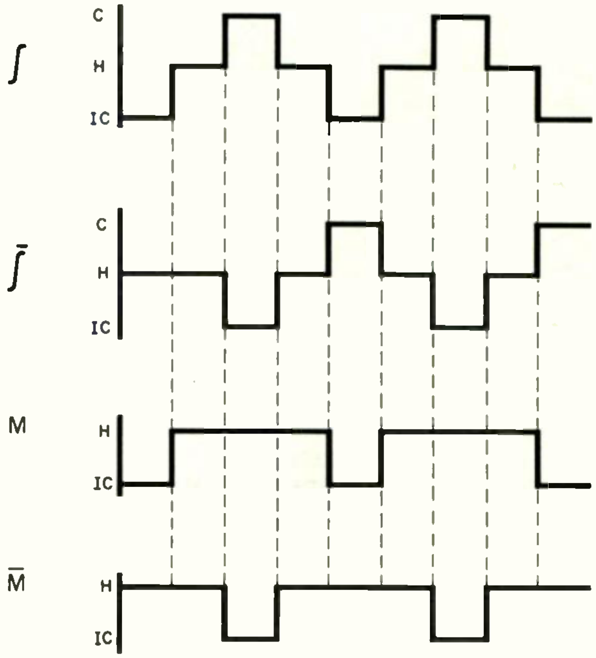
References

1. Andrews, J., Mathematical Applications of the Dynamic Storage Analog Computer, Proceedings of the Western Joint Computer Conference, 1960.



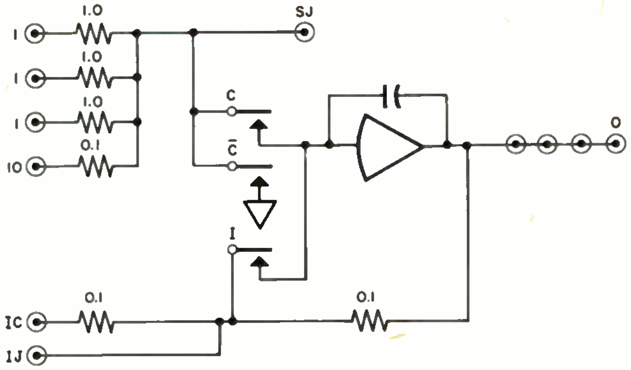
ITERATIVE OPERATION MODE CONDITION
FOR NORMAL AND COMPLEMENTARY INTEGRATORS

FIGURE 1.



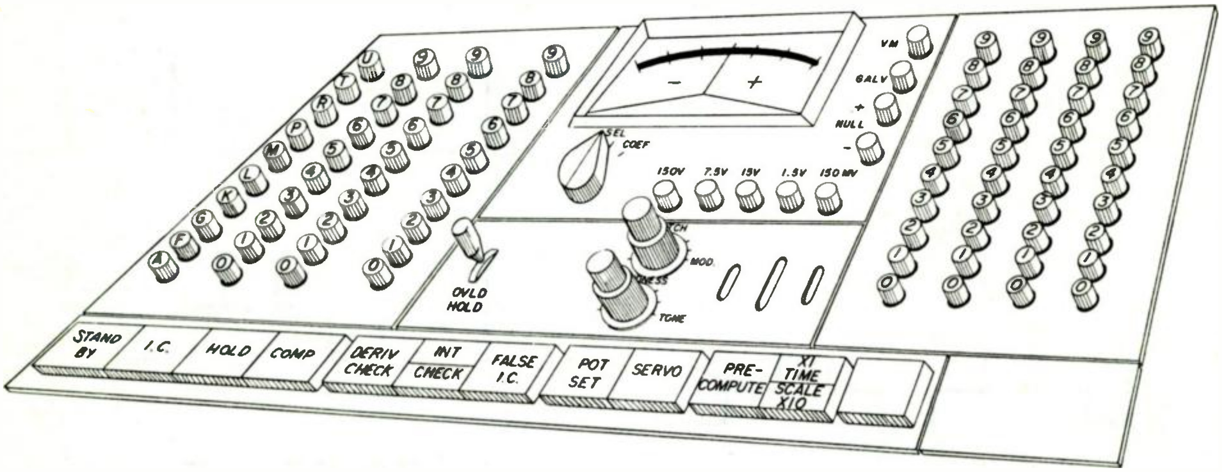
NORMAL AND COMPLEMENTARY
MEMORY-MODE DUTY CYCLES

FIGURE 2.

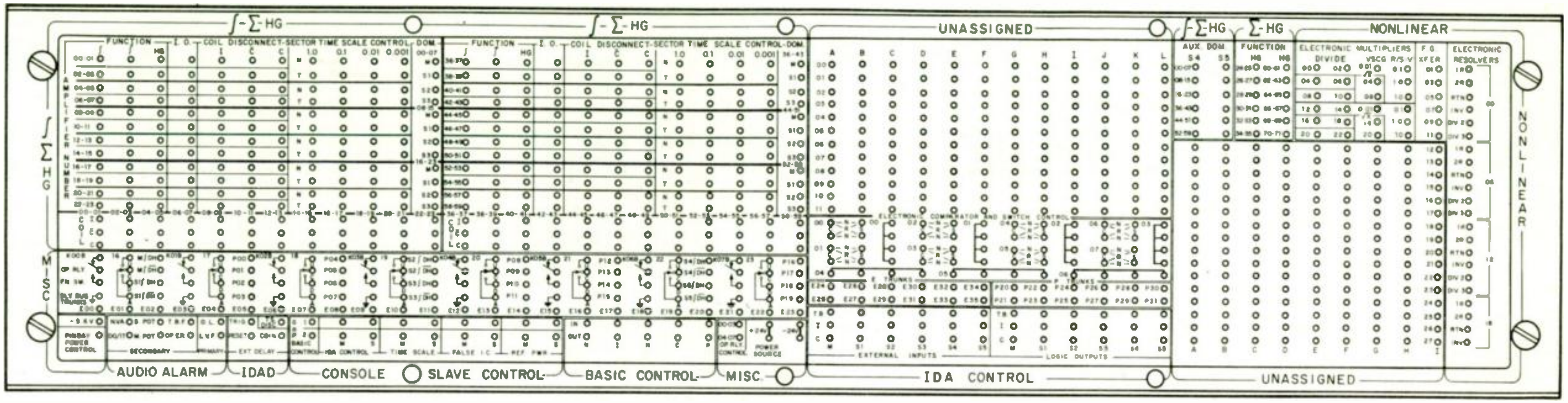


SIMPLIFIED SCHEMATIC
FOR NORMAL AND COMPLEMENTARY INTEGRATORS

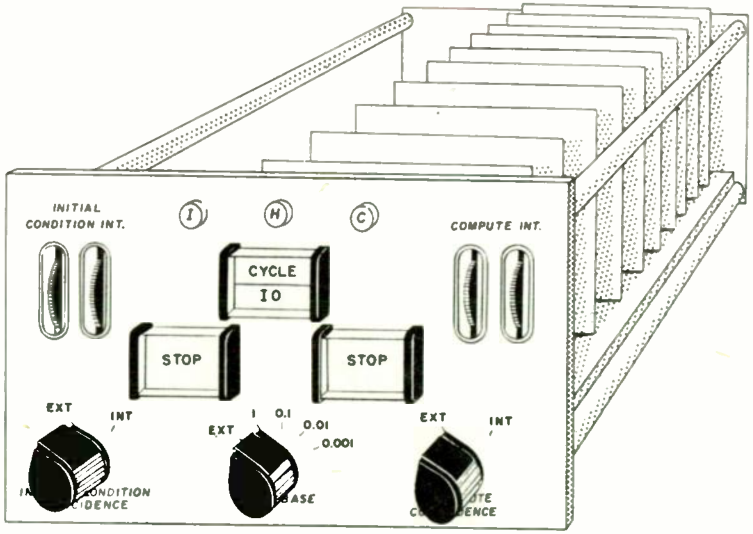
FIGURE 3.



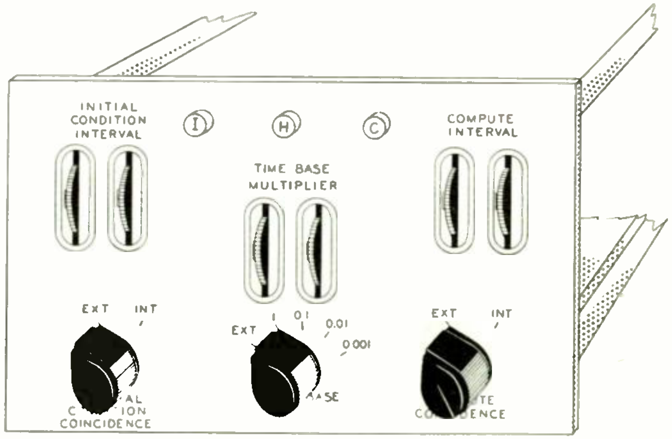
CEDAC UNIT
(CENTRAL DIFFERENTIAL ANALYZER CONTROL)
FIGURE 4.



IDAP UNIT
(ITERATIVE DIFFERENTIAL ANALYZER PINBOARD)
FIGURE 5.



IDACON UNIT
(ITERATIVE DIFFERENTIAL ANALYZER CONTROL)
FIGURE 6.



IDAS UNIT
(ITERATIVE DIFFERENTIAL ANALYZER SLAVE)
FIGURE 7.

DIGITAL SIMULATION OF PULSE DOPPLER TRACK-WHILE-SCAN RADAR

W. A. Bishop and W. A. Skillman
Westinghouse Electric Corporation
Baltimore, Maryland

Summary

The expense involved prohibits the complete evaluation of a complex airborne pulse doppler track-while-scan radar system in the tactical environment for which it is designed by actual flight testing. By employing a high-speed digital computer the radar system and its tactical environment can be simulated resulting in a reliable prediction of the system capabilities at a fraction of the cost of flight testing. Also, a simulation program allows greater flexibility and more accurate control of the test parameters than an actual flight test and is not susceptible to the vagaries of weather and equipment.

The digital computer simulation of such a radar system is described in this paper. A realistic signal output from the filter bank is obtained by simulating one or more dynamic targets with independent scintillation and then adding simulated receiver noise. This output data is then translated into target tracks by the radar data processor.

In addition to predicting the radar system's range performance, angular measurement accuracies, and ability to generate and maintain target tracks, a study of the simulation results led to the optimization and improvement of some system parameters.

Introduction

This paper describes the digital simulation of a pulse doppler track-while-scan radar system. A brief description of the radar system is included. A realistic signal output from the filter bank is obtained by simulating one or more dynamic targets with independent scintillation and then adding simulated receiver noise. This output is then used as an input to the radar data processing unit which separates the data into groups.

Each of these groups represents the returns from a single target. This single target information is used to update existing target tracks or establish new target tracks. Information on target tracks and the radar output are recorded on tape for later printout and analysis.

Radar Description

This pulse doppler track-while-scan radar is used to track several targets within a specified volume of space which is being systematically scanned by radio frequency energy concentrated in a pencil beam by the radar antenna. A target is said to be tracked if its coordinates are measured periodically by the radar so that the future position of the target may be predicted.

The doppler technique makes use of the doppler shift in frequency of radiation reflected from moving targets. This allows low-flying targets with sufficient range rate to be distinguished from the ground clutter, that is, the radar return from the ground. Time sharing of one antenna by use of pulse transmission avoids the leakthrough problems associated with a continuous wave radar. However, a high pulse repetition frequency (PRF) must be employed to avoid doppler ambiguities, resulting in pulses returning from the target after successive pulses have been transmitted. Consequently, range cannot be found directly by measuring the time between the transmission and reception of pulses. In order to determine the true range to the target, it is necessary to put secondary modulation on the transmitted waveform in addition to the basic pulse modulation. The secondary modulation selected is a multiple PRF scheme which requires a range measurement in each PRF before the true range can be determined. This requires that the PRF be switched through each of its multiple values during the time the pencil beam passes over the target and that the radar output be sampled at least once during each PRF.

To measure the range in each PRF, the radio frequency portion of the radar is terminated in a set of contiguous, non-overlapping range-gated channels, see Figure 1. All of the range-gated channels are switched off while a pulse is being transmitted. However, between the transmission of pulses the range-gated channels are switched on in sequence for a period of time equal to pulse transmission time. At a given time only one range-gated channel is in operation. Range in each PRF is then determined by noting in which channel the signal appears. Each of the range-gated channels is in turn terminated in a set of contiguous velocity filters which are used to obtain the velocity of targets by measurement of the doppler shift of the returned signal.

During passage of the antenna beam over the target, many pulses are transmitted and received and the target energy in the velocity filters is integrated for a "look time". At the end of the look time, the filter bank integrators of all channels are sampled and a detection threshold applied to each filter in turn. If the threshold is exceeded, a detection is scored for that filter. These detections and other relevant radar data are sent to the radar data processing unit which sorts the mass of data into groups of single target data. Each group of data is further processed to decide whether it correlates with existing tracks, in which case it can be used to "update" the track; whether it represents a new target, in which case it is used to start another track (acquisition); or whether the data is insufficient for either purpose, in which case it is discarded.

The target coordinates are calculated from the raw data by the data processing unit. True range is a function of all the PRF ranges. Azimuth and elevation are calculated by a centroiding technique which is a function of the antenna scan pattern, the antenna scan rate, the look time, and the amount of data.

Simulation Description

Introduction

A simplified block diagram of the pulse doppler track-while-scan radar simulation is shown in Figure 2. A number of target aircraft are simulated, each with independent motion in three dimensions and with independent scintillation. These may be placed in various raid configurations as desired. The total signal power in each filter is obtained by summing the contribution from each aircraft. The signal power is converted into an output current which is modified by the AGC action of the range gated channel and compared to a detection threshold. After this threshold process has been applied to all the filters containing signal power in the range gate - filter bank matrix, the resulting detections and other pertinent radar data may be converted into target tracks by the radar data processor. This routine is, repeated for each look time as the radar antenna scans the given volume of space.

Input Parameters

Initially the radar simulation program and input parameters are stored in the digital computer. The length of the simulation, the number of targets, and certain target and radar parameters may be varied from run to run in order to provide a high degree of flexibility in the simulation.

Target Parameters. One set of the following parameters is necessary for each target that is simulated.

1. Initial position.
2. Initial velocity.
3. Acceleration.
4. Average radar cross-sectional area.
5. Type of scintillation: scan-to-scan or look-to-look.

Radar Parameters. The following radar platform, antenna, and receiver parameters can be varied from run to run.

1. Radar platform parameters.
   a. Initial position.
   b. Initial velocity.
2. Radar antenna parameters.
   a. Scan pattern.
   b. Scan rate.
   c. Antenna beamwidth.
3. Radar receiver parameters.
   a. Look time.
   b. Transmitted pulse width.
   c. Range gate width.
   d. Number of PRF's.
   e. Number of range gates for each PRF.
   f. Looks per PRF.
   g. Idealized range, R_o, (unity target cross-section).
   h. Frequency of first filter in filter bank.
   i. Filter width (frequency spacing of filters).
   j. Number of filters in filter bank.
   k. Threshold level.

Target Signal Strength

During a look time the total energy received from all targets is integrated in the range gate - filter bank matrix of the radar. Since the range gate width is equal to the transmitted pulse width, energy from a single target will, in general, appear in two adjacent range gates. Some energy will be present in all the filters in each of these range gates due to the filter skirts. However, the independent automatic gain control of each range-gated channel and the steep skirts of the filters result in significant energy being present in only the five contiguous filters nearest the frequency of the signal. Therefore, in the simulation the signal strength in five contiguous filters in two adjacent range gates is calculated for each target. Following is the step by step procedure used to arrive at the target signal strength appearing in the filter bank.

Basic Signal Strength. On each look, the antenna is positioned, the first target is selected and its angle with respect to the peak of the antenna beam is calculated. If the target is not illuminated by the main beam of the antenna, the signal from this target is not calculated (to reduce running time) and the next target is selected. Omission of possible side lobe returns causes negligible error if the target remains at relatively long range which is the case of interest here. If the target is within the radar antenna's main beam, the range of the target with respect to the radar platform is calculated and the basic signal strength of the target computed using the following equation:

$$(S/N)_1 = {\frac{R_o}{R}}^{4} \qquad (1)$$

where R is the target range and R_o is the range at which the signal-to-noise ratio is unity for unity target cross section.

R_o is a function of the radar system parameters and it is arrived at under the following conditions:

1. The target cross section does not scintillate and is unity.
2. The target is located at the peak of the radar antenna main beam.
3. The target energy is centered in a range gate.
4. The frequency of the target energy coincides with the peak gain of a doppler filter.

Since these assumptions do not hold in the general case this basic signal strength is modified to remove each of these restrictions.

Target Cross Section and Scintillation. The basic signal strength is first modified to account for the target cross section and scintillation as follows:

$$(S/N)_2 = (S/N)_1 \sigma (SF) \qquad (2)$$

where σ is the mean target cross-section and SF is the target scintillation factor.

Target scintillation is essentially a random variation in target cross-sectional area due to small short term variations in aspect angle. The assumption is made that the cross-section has an exponential distribution. Jet aircraft normally scintillate slowly so that the further assumption is made that the cross-section is constant during the time the antenna main beam is passing over the target. This is called scan-to-scan scintillation. If the cross-section changes very rapidly it is called look-to-look scintillation. Both types of scintillation are accounted for in the simulation by generating random numbers with an exponential distribution with an average value of unity. This number is obtained as follows:

$$SF = \frac{1}{2} (R_1^{\,2} + R_2^{\,2}) \qquad (3)$$

where R_1 and R_2 are independent random numbers having a Gaussian distribution with mean zero and standard deviation of unity. A target scintillation number is computed independently for each target whenever the radar antenna's main beam passes over it, changing either scan-to-scan or look-to-look, depending on the input parameter selection.

Antenna Gain. Next the signal strength is altered by the actual antenna gain in the target's direction by the equation:

$$(S/N)_3 = (S/N)_2 \, G^2 \qquad (4)$$

where G is the one-way power gain of the Gaussian antenna pattern normalized to unity peak gain and is calculated as follows:

$$G = e^{-4 \ln 2 \left(\frac{\theta}{BW}\right)^2} \qquad (5)$$

where θ is the angle between the target and the peak of the radar antenna main beam at the center of the look time and BW is the half-power beamwidth. It is assumed that the antenna beam does not travel through a large angle during the look time, otherwise, the antenna gain would have to be integrated over the look time.

Range Gating. The signal-to-noise ratio at this point assumes the target return is centered in a range gate. If the target return pulse partially overlaps or straddles a range gate, the energy in the velocity filter will be reduced. To calculate the reduction, the position (in units of range gate width) of the return, AR, in the interpulse period (time between transmitted pulses) is first calculated using the slant range to the target, R. See Figure 3.

$$AR = \frac{2}{c\tau} R \text{ (modulo range between pulses)} \qquad (6)$$

where c is the velocity of light and τ is the range gate width.

The integral part of AR is the number of the range gate in which the leading edge of the return first appears and is designated n. The fraction of the target pulse in range gate n is called the straddling factor for the nth range gate. Mathematically, the straddling factor, x, is:

$$x = n + 1 - AR \qquad (7)$$

The straddling factor is 1 - x for range gate (n + 1).

If n is zero, meaning that this portion of the return was received during the transmission of a pulse (eclipsing), x is set equal to zero. Similarly, if n is equal to the highest numbered range gate in this PRF, 1 - x is set equal to zero since that portion of the return was eclipsed.

The reduction in power in the doppler filter caused by the straddling is proportional to the square of the straddling factor since only the power in the central spectral line is used in a pulse doppler radar. The signal-to-noise after range gating for range gates n and n + 1 is then:

$$(S/N)_n = (S/N)_3 \, x^2 \qquad (8)$$

$$(S/N)_{n+1} = (S/N)_3 \, (1 - x)^2 \qquad (9)$$

It is further assumed that:

1. The target is a point so that range and angle noise can be neglected.
2. The target does not move during the look time.

Filter Attenuation. Since, in general, the return signal will not be centered in a velocity filter, an attenuation factor must be calculated. Due to the skirts of the filters a return has some power in all filters. However, due to the AGC action of the range-gated channel and the steepness of the skirts of the filters used, significant power is present in only the five contiguous filters nearest the signal.

The frequency of the return is found from the doppler equation:

$$f = \frac{2\, f_t\, \dot{R}}{c} \qquad (10)$$

where $\dot{R}$ is the range rate of the target with respect to the radar platform, c is the velocity of light, and f_t is the transmitted carrier frequency.

The normalized filter gain function A $(f-f_i)$ (where f_i is the center frequency of the doppler filter), is approximated in the simulation by a polynomial fitted to the actual shape of the filters used in the radar. The signal-to-noise ratio, $(S/N)_{n,i}$, in each of the five filters nearest the target frequency is obtained by multiplying the previously calculated signal-to-noise ratio, $(S/N)_n$, by the appropriate filter attenuation factor:

$$(S/N)_{n,i} = (S/N)_n \left[A\ (f - f_i)\right] \qquad (11)$$

$$(S/N)_{n+1,i} = (S/N)_{n+1} \left[A\ (f - f_i)\right] \qquad (12)$$

Summation of Filter Signal Strength

The value of $(S/N)_{n,i}$ and $(S/N)_{n+1,i}$ is calculated for all targets in the main beam. Targets flying in formation may appear in the same range gate and in the same filters. Assuming the signals are independent, they will add powerwise, hence $(S/N)_{n,i}$ is summed over all targets for each filter. This summing process is performed in steps. That is, as soon as $(S/N)_{n,i}$ and $(S/N)_{n+1,i}$ are calculated for a target, they are added to the $\sum(S/N)_{n,i}$ from previous targets on this look before $(S/N)_{n,i}$ and $(S/N)_{n+1,i}$ are calculated for the next target.

Detection Process

After the total signal in each filter is calculated, only filters which contain signals are tested to see if the detection threshold is exceeded. This thresholding process is applied to the filters in one range gate at a time.

To prevent saturation, a fast automatic gain control (AGC) is used independently in each range gate to maintain a constant amplitude signal into the filter bank. The effect of this AGC action is accounted for in the detection thresholding process as follows:

$$\text{If } I_{n,i} > B \cdot I_n, \text{ a detection is scored} \qquad (13)$$

where $I_{n,i}$ is the total output current in the ith filter of the nth range gate, B is the threshold level, and I_n is the total output current in the nth range gate.

If $I_{n,i}$ is greater than $B \cdot I_n$ a detection is scored for the ith filter in the nth range gate. This process is continued for each filter containing a signal in this range gate. When all filters have been tested the next range gate which contains signals is selected and the above process repeated.

Both output currents I_n and $I_{n,i}$ are calculated using the following equations.

Assuming a linear AGC detector, the output current is simulated by first calculating the direct current, I_{dc}, by use of the usual equation:[1]

$$I_{dc} = \left[(1 + F) \cdot I_o\left(\frac{F}{2}\right) + F \cdot I_1\ \left(\frac{F}{2}\right)\right] \cdot e^{-\left(\frac{F}{2}\right)} \qquad (14)$$

where I_o and I_1 are modified Bessel functions and F is the total signal-to-noise ratio.

The mean square alternating current, $\overline{I^2_{ac}}$, out of the AGC detector is found by means of the equation:

$$\overline{I^2_{ac}} = \left[\frac{4}{\pi}\,(1 + F) - I_{dc}{}^2\right] \qquad (15)$$

Now the output current is filtered effectively over the look time so that $\overline{I^2_{ac}}$ is correspondingly reduced. The reduction factor can be shown to be the square root of 1/1.25 times the bandwidth reduction ratio. Due to the integration, the distribution of the alternating current is approximately Gaussian. A Gaussian random number, RN, is therefore used to modify the value of I_{ac} to simulate the effect of noise. With these effects included the instantaneous alternating current output, I_{ac}, is:

$$I_{ac} = RN \left(1.25\ \frac{\Delta f}{\Delta F}\left[\frac{4}{\pi}\ (1 + F) - I_{dc}{}^2\right]\right)^{1/2} \qquad (16)$$

where Δf is the reciprocal of the look time, ΔF is the bandwidth, and F is again the total signal-to-noise ratio.

Finally, the total output current is arrived at by adding the ac and dc components.

$$I = I_{dc} + I_{ac} \qquad (17)$$

The I in equation 17 is I_n if ΔF is the bandwidth of the range gated channel prior to the filter bank and F is F_n where:

$$F_n = \sum_{j=1}^{m} (S_j/N)_n \qquad (18)$$

where m is the total number of targets.

The I in equation 17 is $I_{n,i}$ if ΔF is the bandwidth of the doppler filter and F is $F_{n,i}$ where:

$$F_{n,i} = \sum_{j=1}^{m} (S_j/N)_{n,i} \qquad (19)$$

Data Processing

After the whole filter bank has been tested, the radar data is operated on by the tracking rules which constitute the logic contained in the data processor for generating and maintaining target tracks. The radar data is separated into groups of data. Each of these groups represents the returns from a single target. These single target data are used to update existing target tracks or establish new target tracks.

When the processing of the radar data is finished, information on target tracks and the filter bank output are recorded on tape for later printout and analysis.

If the run has not been completed, time is incremented by one look time and the calculations for the next look time are performed.

Conclusions

The simulation program described has been used extensively over the past several years to evaluate performance and optimize parameters and tracking rules for a pulse doppler track-while-scan radar. Flight test data has given good agreement with predicted performance over a limited flight regime.

Such a simulation program allows greater flexibility and more accurate control of the test parameters than an actual flight test and is not susceptible to the vagaries of weather and equipment. The ability to generate and maintain target tracks, the angular measurement accuracies, and the range performance of the radar can be predicted for any set of radar and tactical parameters. The effect of present and future tactical raid models on the system can be studied. Optimization of the logic used to set up and maintain target tracks can be accomplished by varying the "tracking rules" while keeping the radar and raid model constant. Finally, various parameters of the radio frequency portion of the radar may be varied in order to optimize them, thereby obtaining best system performance at minimum cost.

Modifications have since been made to the basic program to incorporate various types of electronic countermeasures to evaluate performance in this environment. Further modifications are in progress to evaluate performance enhancement due to various types of pause modes. In these modes the antenna is caused to dwell for a longer time on the target to obtain a better signal-to-noise ratio. The logic involved in deciding when and where to pause can be investigated easily with the simulation. Thus, the basic radar simulation program provides a useful tool both to evaluate performance of present day radars and to investigate more advanced radars in different environments.

References

1. Goldman, Stanford, Frequency Analysis, Modulation and Noise. New York: McGraw-Hill Book Company, Inc. 1948, Chap. 6, p. 246.

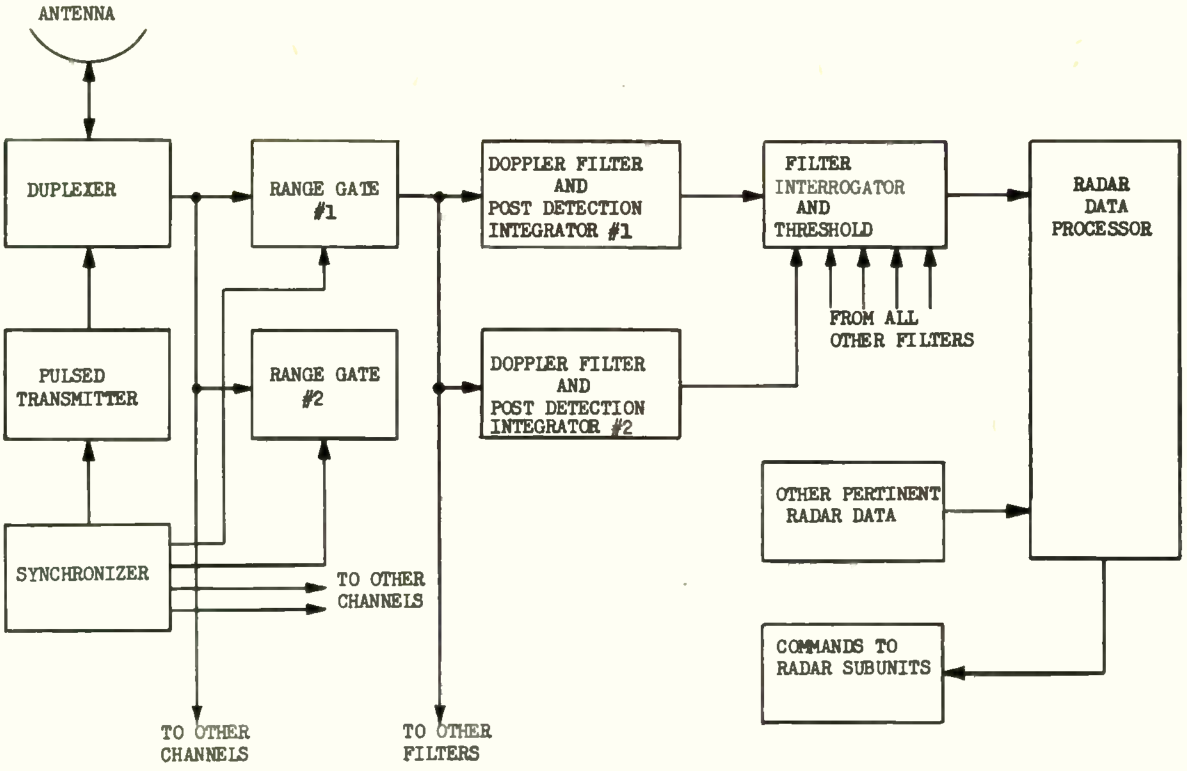

ANTENNA
DUPLEXER
RANGE GATE #1
DOPPLER FILTER AND POST DETECTION INTEGRATOR #1
FILTER INTERROGATOR AND THRESHOLD
RADAR DATA PROCESSOR
FROM ALL OTHER FILTERS
PULSED TRANSMITTER
RANGE GATE #2
DOPPLER FILTER AND POST DETECTION INTEGRATOR #2
OTHER PERTINENT RADAR DATA
SYNCHRONIZER
TO OTHER CHANNELS
COMMANDS TO RADAR SUBUNITS
TO OTHER CHANNELS
TO OTHER FILTERS

FIGURE 1. PULSE DOPPLER RADAR SYSTEM

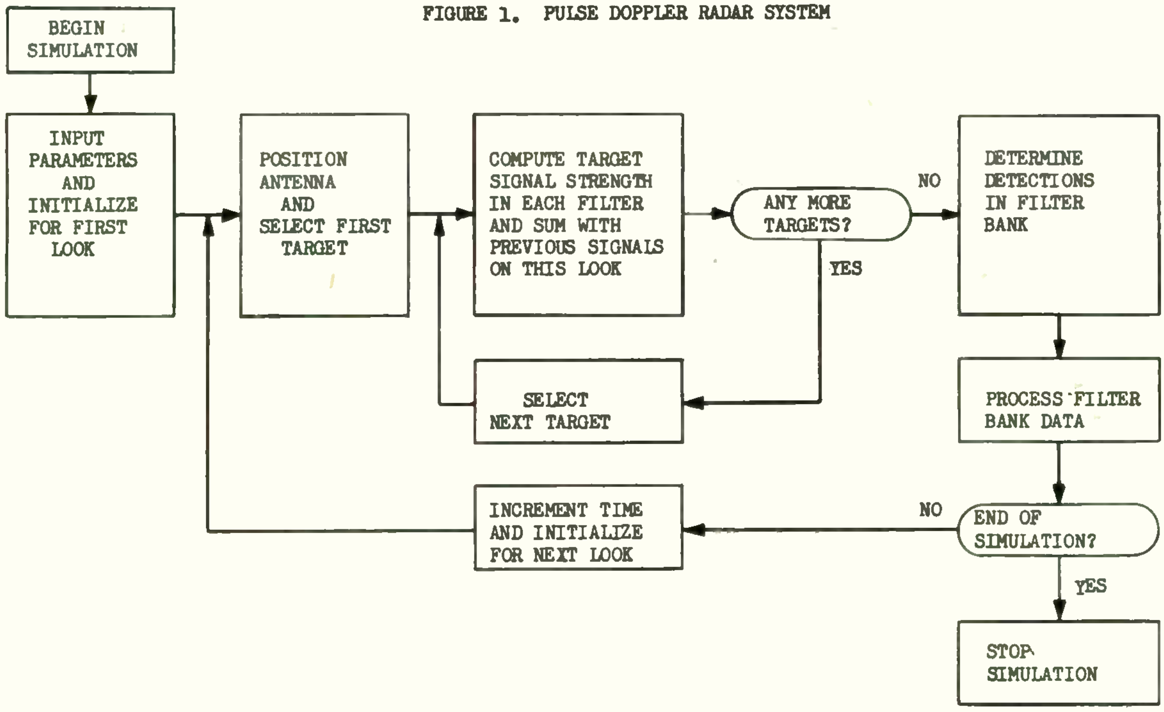

BEGIN SIMULATION
INPUT PARAMETERS AND INITIALIZE FOR FIRST LOOK
POSITION ANTENNA AND SELECT FIRST TARGET
COMPUTE TARGET SIGNAL STRENGTH IN EACH FILTER AND SUM WITH PREVIOUS SIGNALS ON THIS LOOK
ANY MORE TARGETS?
NO
YES
DETERMINE DETECTIONS IN FILTER BANK
SELECT NEXT TARGET
PROCESS FILTER BANK DATA
INCREMENT TIME AND INITIALIZE FOR NEXT LOOK
NO
END OF SIMULATION?
YES
STOP SIMULATION

FIGURE 2. PULSE DOPPLER TRACK-WHILE-SCAN RADAR SIMULATION

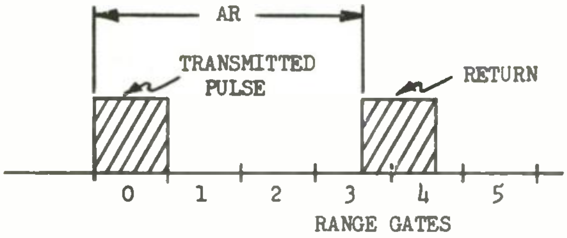


FIGURE 3. INTERPULSE INTERVAL

TELLERTRON

A REAL-TIME UPDATING AND TRANSACTION PROCESSING SYSTEM FOR SAVINGS BANKS

Dan M. Bowers
William T. Lennon, Jr.

William F. Jordan, Jr.
Donald G. Benson

Computer Control Company, Inc.
Framingham, Massachusetts

Summary

This paper describes the overall system design of a large-scale, high-speed information updating and retrieval complex for accelerating teller window transactions and for maintaining up-to-the-minute customer account information.

A central 40-million bit memory stores unposted interest, customer balances, and other information for each savings account. Proper information is visually displayed to the teller and updated by the teller within seconds, thus providing new and improved customer service. Branch office tellers are served from the home office by a dataphone link. Records of all of the days transactions are also made for bookkeeping purposes.

Introduction

Savings bank transactions necessitating teller reference to unposted interest documents and customer balances are generally time consuming. To expedite such transactions, and thus ensure customer satisfaction, more tellers, equipment, and space are constantly required to properly service a growing number of accounts. Further, during periods of peak activity, hasty reference to inconvenient records often results in erroneous entries and monetary mistakes.

TELLERTRON* (figure 1) alleviates many of these problems by automatically providing depositor information in the form of a visual display. In addition, the system automatically updates accounts and provides a complete record of all transactions (as a by-product of its other functions) in a form that is directly admissible to a data processing system without additional intermediate operations.

TELLERTRON accomplishes all of the above functions faster than present methods, and without special effort or training on the part of bank personnel. More important, tellers never need to search files.

When a teller enters a savings account number on his window posting machine (figure 2), an illuminated display automatically provides the following information: old balance; amount of unposted interest; existence of a previously deposited check which may not have cleared; existence of situations requiring special attention; and the number of a teller who has processed a previous transaction on the same account that day.

The information supplied to the teller is current, whether the last transaction was processed years or even seconds ago, since updating is performed simultaneously with the original handling of the transaction (in real-time).

Additional checking and control features are provided for reducing or

*TELLERTRON is a registered trademark of TELLERTRON, INC., a subsidiary of Stone Laboratories, Inc., Boston, Massachusetts.

eliminating such errors as; (1) recording of an incorrect old balance; (2) recording of incorrect unposted interest amount; and (3) posting to wrong accounts. By eliminating these errors, balancing operations are improved and error corrections are unnecessary. Since the teller has acquired information faster and with less effort, more of his attention may be directed to both the transaction and the depositor. This results in superior operation control and customer goodwill.

Modular construction is utilized in TELLERTRON design, so that the number of teller stations, offices, and accounts can be expanded as growth may warrant.

The differences among individual banks in procedures, equipment, and the fundamental parameters of number-of-tellers and number-of-accounts make it necessary to describe the operation of one particular TELLERTRON. TELLERTRON Serial One is being delivered to the Provident Institution for Savings, in Boston, under a contract to Stone Laboratories, Inc., of Boston, and the details following apply specifically to this installation.

Teller Operations

TELLERTRON operations can be more easily understood if a degree of familiarity with teller operations is acquired. The basic sequence of operations, by both teller and system, is basically as follows.

(1) The teller inserts the customer's passbook into a window posting machine, enters the customer's account number into the keyboard of the window posting machine, and depresses a motor bar which causes the window posting machine to print the account number on its journal sheet. Depression of the motor bar places the teller into a "ready" status so far as TELLERTRON is concerned. The teller's posting machine is automatically locked so that he cannot proceed until he has been serviced by TELLERTRON.

(2) TELLERTRON locates the teller who requires service, investigates the condition of his window posting machine, determines that an account number has just been entered, reads the account number from the posting machine, and locates that account number in its mass memory. TELLERTRON then extracts all information pertinent to that account from the mass memory, and sends the following portions of the information (if appropriate) to the teller's visual display unit:

"HOLD" (possible uncleared check)

"STOP" (investigate special instructions)

"INT" (unposted interest exists on this account)

"NB" (unposted "no-book" transaction exists on this account)

"TELLER NUMBER" (the number of a teller who has posted another transaction to this same account that day)

(3) The teller views the displayed information, and if no irregularities exist, reads the customer's balance from the passbook, enters it into the keyboard of his window posting machine, and depresses a motor bar. The posting machine prints the balance on its journal sheet, and the teller is placed into a "ready" status.

(4) TELLERTRON locates the "ready" teller, and determines by the condition of his posting machine that the teller has just entered an "old balance". TELLERTRON reads the "old balance" from the posting machine, and compares it with the "old balance" stored in memory. If the two numbers are identical, the teller is allowed to proceed. If they are not, the posting machine is locked so that no further operations can be performed. The teller is notified, via an "error" light, that he has entered the wrong "old balance", and the correct "old balance" from memory is displayed to the teller.

(5) If an error was committed, the teller must restart. If no error was committed, the teller enters the transaction amount (deposit or withdrawal)

into the keyboard of his window posting machine, and depresses a motor bar. The posting machine prints the transaction amount on its journal sheet, and the teller is placed into a "ready" status.

(6) TELLERTRON locates the "ready" teller, determines that the teller has just entered a transaction amount on his keyboard, reads the amount and type of transaction, (deposit or withdrawal) and stores it for recording on the daily transaction record.

(7) If no errors in the transaction have been detected by TELLERTRON, the window posting machine is caused to print all transaction information onto the customer's passbook, and calculates a new balance from the "old balance" and the "transaction amount". TELLERTRON reads the "new balance" from the posting machine at the proper time, stores it for rewriting of memory, and allows the posting machine to return to its home position.

(8) TELLERTRON determines that the transaction has been successfully completed, writes the new information (e.g., new balance, new flags) concerning this account into the proper place in its mass memory, writes the details of this transaction into the daily transaction file, and allows the teller to proceed with the next account.

(9) If a teller discovers during his processing of a transaction that he has committed an error (e.g., entered an incorrect deposit amount), he will depress the "ER" key, return his window posting machine to the "home" position and restart. TELLERTRON will ignore the erroneous transaction. TELLERTRON's procedure in this case is the same whether the teller discovers the error, or TELLERTRON discovers the error (e.g., incorrect old balance entry).

It is an important function of TELLERTRON to control and verify routine teller operations as described above. However the greatest benefit of TELLERTRON is its handling of special situations, as described below. (Perhaps the single most important benefit of TELLERTRON is its display of unposted interest.)

(1) Interest Posting:

When the "INT" flag (signifying that unposted interest exists on this account) is displayed to a teller following his entry of an account number, the amount of unposted interest will be displayed to the teller immediately following his correct entry of the "old balance". He may then enter the interest amount in place of the normal transaction (i.e., perform an "interest posting transaction"), and TELLERTRON will compare the interest entry with the interest amount stored in its memory, and allow him to proceed only if he has correctly entered the interest amount.

(2) Not-on-Book Outstanding:

When the "NB" flag (signifying that the last transaction conducted on this account was done in the absence of the passbook) is displayed to a teller following his entry of an account number, the "old balance" which the teller enters from the passbook is not up to date. The teller must then bring the passbook up to date by entering the outstanding not-on-book transaction (an optional TELLERTRON feature provides automatic display of the details of the outstanding not-on-book transaction). When the passbook has been updated, TELLERTRON compares the memory "old balance" with the passbook balance. These two amounts must be identical or an error is displayed and the teller's posting machine is locked.

(3) Duplicate Transactions:

The processing of an outstanding not-on-book transaction to update the passbook is a duplication of a previously conducted transaction, and is therefore designated as a "duplicate transaction". TELLERTRON records duplicate transactions in the daily transaction file, but does not update the memory information from them, since the memory already includes this information.

(4) Not-On-Book Transaction:

When a customer appears at the teller window without his passbook, the transaction to be processed will be a "not-on-book", and the teller depresses an "NB" key when he enters the account number. TELLERTRON will display the memory "old balance" immediately, since the teller has no other means of obtaining a balance amount.

(5) Multiple Transactions:

When a customer desires to make several transactions during one visit to the bank (e.g., interest posting and a deposit), the teller needs only to enter the account number at the start of the first transaction, and not for each transaction. TELLERTRON will prepare a daily transaction record for each separate transaction of the multiple group, and update its memory after the last transaction of the group.

(6) Mortgage Transactions:

Mortgage transactions are processed by TELLERTRON Serial One only to the extent of recording them on the daily transaction file. Thus TELLERTRON acts as a teller-to-magnetic tape converter for mortgage transactions.

(7) Special Designations:

Special symbols designating that the transaction is conducted by mail, or by check, may be entered from the teller's window posting machine to the daily transaction record. Other symbols ("hold", "stop") may be entered into the TELLERTRON memory.

(8) Identification Card ("I.D.") Accounts:

Some accounts have no passbook issued, and all transactions are conducted on the strength of an "I.D." card issued to the customer. TELLERTRON will display memory "old balance" to the teller for all I.D. accounts, and then proceed normally.

(9) Error Correction:

Transactions processed by a teller to correct an account are designated "EC" from the teller's window posting machine, and are handled as normal transactions by TELLERTRON except that the "EC" is carried with them.

System Organization

All account information is stored on a large magnetic Disk Storage Unit, (figure 3). Access to, addressing of, reading from, and writing onto the Disk Storage Unit are functions performed by the Main Memory Unit. The address lookup function (given an account number, finds its address on the Disk Storage Unit) is also performed by the Main Memory Unit. Given an account number from the Central Processor, then, the Main Memory Unit locates the stored account information in the Disk Storage Unit.

Transaction information is stored on a portion of the Disk Storage Unit until the end of the business day, at which time the complete daily transaction record is written onto a magnetic tape. These operations are under control of the Main Memory Unit.

The Disk Storage Unit is completely re-written, quarterly, in order to add new interest amounts, delete closed accounts, and add new accounts. This re-writing is done from a magnetic tape prepared by an off-line date processor. This operation is under control of the Main Memory Unit.

The Central Processor is built around a Teller Assembly Register, (or Teller Buffer), and each teller is assigned a portion of this register. When the Main Memory has looked up account information for a teller, all the information is placed into the Teller Buffer. All subsequent communications necessary for processing a transaction are performed purely between the teller and his Teller Buffer, thus obviating repeated time-consuming accesses into the Disk Storage Unit.

All tellers are time-sequenced into the Central Processor, with one teller being given access to his Teller Buffer every 16 milliseconds.

All communications between Central Processor and Main Memory are also accomplished through the Teller Buffer, with the Main Memory Unit time-sequencing the requests for memory lookup or write.

The comparators for assuring that correct amounts are entered by the teller are located in the Central Processor. Comparison, routing of information from a Teller Buffer to the teller's display unit, writing of information from the posting machine into the Teller Buffer, error display, and locking of the posting machine are controlled by a wired-program data processor, which constitutes the bulk of the Central Processor.

Branch bank tellers are accommodated in the Central Processor through the Branch Processor and a Branch Central located at each branch bank. The Branch Processor and Branch Centrals are data collection devices which allow branch tellers to use telephone facilities to communicate with the Central Processor. Each branch teller has a Teller Buffer located in the Central Processor.

Main Memory Unit

The Main Memory Unit (figure 4) stores and looks up all account information, accumulates the daily transaction record, and handles all access to or storage of this information by the teller (via the Central Processor), or the off-line accounting equipment (via magnetic tape).

(1) The Main File, which is located on the Mass Memory device, stores 32 characters of information about each of 225,000 savings accounts. The information sorted for each account includes the account number, old balance, unposted interest amount, date of last transaction and number of the teller who conducted that transaction, and special conditions (flags).

TELLERTRON Serial One employs an IBM 350 Disk Storage Unit for its Mass Memory. Both sides of 50 magnetic disks are used for storing 40 million bits of information, which are sufficient to handle 250,000 savings accounts (25,000 positions are reserved in TELLERTRON Serial One for temporary storage of the daily transaction records). Each disk side contains 100 tracks, each of which is divided into five sectors. Each sector contains the information necessary to describe five savings accounts. Two access arms are operated independently in order to speed up handling of inquiries to the disk file, and provide backup in case of failure of one of the arms. The average access time to an address, using two independent access arms as employed in TELLERTRON, is about 280 milliseconds. An account is located by addressing an access arm to a pair of tracks on opposite sides of a disk, and serially searching until the correct account is located. It is of interest to note that, should all 21 tellers in the main office of the bank served by TELLERTRON Serial One simultaneously request access into the Main File, the last teller to receive his answer will be served within seven seconds.

(2) The Memory Request Scanner sequences the tellers requesting memory lookup, such that one and only one account number is being handled per arm. The Memory Request Scanner scans the Teller Memory Status Register, which is part of the Central Processor Unit.

(3) Two Sequential Core Buffers (DI/AN SA-1A-1NT) act as speed-changing and intermediate storage devices between Main File and Teller Buffer, Main File and Magnetic Tape, Daily Transaction File and Teller Buffer or Magnetic Tape. When operating with the Teller Buffer, for example, on lookup of account information, the desired account information is located in the Main File, and then read through a Sequential Core Buffer into the Teller Buffer of the teller requesting the information. After the account data has been altered by the teller and it is desired to update the file, the file data is read from the file into the Sequential Core Buffer, the new data appended to the buffer contents, and the file rewritten. It is important to note that the Teller Buffer is located on a magnetic drum (Bryant 10005) and its bit rate is 227 kc; the bit rate of information from and to the mass memory is 83.3 kc; the bit rate of information from and to the Magnetic Tape Handler (IBM 729 II) is 41.7 kc.

(4) The address lookup section will quickly produce the address (out of 5000 possible addresses) of the track-pair on the Main File

wherein a given account number and its associated information may be found. The address lookup section consists of an account number register, into which the desired account number is placed (from a Teller Buffer); the file address storage, which is a table lookup device written on the magnetic drum; and the file address register, into which the desired address is placed for addressing of an access arm. In order to speed up handling of requests by the file, the address of the next memory operation is determined while the memory is busy with previous searches.

The address lookup procedure takes advantage of the fact that the account information is stored in the disk file sequentially by account number. It is hampered, however, by the fact that 80 percent of the account numbers are unused (for example, closed accounts), and file capacity limitations prohibit the storage of these unused numbers.

The file address storage is composed of two sections: (a) a primary (coarse) address section, stored on one drum track, which allows the system to determine the correct disk address of any account and directs the lookup system to the correct portion of the secondary address section; and (b) a secondary (fine) address section, stored on 45 drum tracks, which allows the system to determine the correct track-pair address of any account. The total address look-up takes an average of 26 milliseconds. When the complete address is determined, it is presented to the next file access arm that becomes available for use.

(5) The Daily Transaction File, located on five disks of the Mass Memory device, accumulates records of all transactions performed by all tellers during the day. All transaction records are written onto Magnetic Tape at the end of the day for entry into the bank's off-line accounting equipment. This writing of the transaction tape takes a maximum of three minutes if the maximum of 25,000 transactions must be transferred.

(6) A Magnetic Tape Handler and its associated controls, format control, and compatibility circuits control the writing of the Daily Transaction File onto tape, rewriting of the Main File in order to load new interest amounts and new accounts, reading out of the entire Main File for accounting purposes, and selective altering of accounts in the Main File for error corrections, payroll savings, etc. The format is adapted to the requirements of the off-line accounting equipment which the bank chooses; in the case of TELLERTRON Serial One, this is an IBM 1401. These routines involving magnetic tape are off-line operations so far as TELLERTRON is concerned, and may not be done while TELLERTRON is serving tellers in an on-line capacity.

The entire 225,000 accounts in the Main File may be written onto magnetic tape in 24 minutes. The entire Main File may be loaded with 225,000 accounts from magnetic tape in 66 minutes.

Central Processor

The Central Processor (figure 5) is a wired-program data-handling unit which implements the teller program in handling all communications between the tellers and their buffers.

(1) A Teller Assembly Register (Teller Buffer) is provided for each teller. The entire Teller Buffer section is made up of three 2990-bit serial buffer loops on the magnetic drum, and each teller is assigned a 230-bit portion of a loop. A total of 39 tellers is accommodated by the buffer loops in TELLERTRON Serial One. All communications between teller and TELLERTRON, while the teller is processing a transaction, are carried out between the teller and his buffer. Each communication requires 16 milliseconds. Initially, an account number is inserted into the Teller Buffer from the posting machine. This account number is transmitted to the

Main Memory from the Teller Buffer, and the information pertinent to that account is returned from the Main Memory to the Teller Buffer. The information is displayed to the teller and modified by the posting machine, entirely through teller-Teller Buffer communications. Rewriting of the Mass Memory is done from the Teller Buffer. Thus, only two long-time (500 millisecond) accesses to the Mass Memory (lookup and rewrite) are required per transaction, with the many intermediate operations being performed through the short-time (16 millisecond) access to the Teller Buffer.

(2) The Teller Status Register indicates, for each teller, the necessity for the teller's access to his buffer, the nature of operation to be performed, and error conditions.

(3) The Teller Ready Scanner sequences the tellers such that only one teller is operating with his Teller Buffer during a 16-millisecond period.

(4) The Posting Machine Data Selectors "read" the posting machine of the teller who has been given access to his Teller Buffer by the Teller Ready Scanner.

(5) The Posting Machine Column Logic programs the interrogation of various columns of the posting machine.

(6) The Decimal-to-BCD Converter converts the posting machine's decimal outputs to binary-coded-decimal for use by TELLERTRON.

(7) The Write Format Control selects, alters, and arranges posting machine data for writing into the Teller Buffers.

(8) The Teller Buffer In-Out Register allows input to and out-put from the buffer loops. Teller information is interlaced in the Teller Buffers to facilitate real-time communication with the tellers at a slow rate (3 kc) due to the distance and line capacity involved. The Teller Buffer In-Out-Register allows short-term storage of buffer data for display on the teller's Display Unit.

(9) The BCD-to-Decimal converter decodes Teller Buffer information for display on the teller Display Units.

(10) The Display Format Control selects the proper Teller Buffer data for display at the teller Display Unit.

(11) The Display Column Logic is a logical network which generates column distribution signals for the Teller Display Units. These column signals distribute the display data to the various digit positions in the Display Units.

(12) A Teller Memory Status Register informs the Main Memory Unit that a teller requires access to the Mass Memory, and the nature of access (read or write) required.

These wired-program units of the Central Processor control and implement the following processes during the course of a transaction:

(1) Visual display of:

a. Old balance.
b. Unposted interest.
c. Number of a teller who operated on the same account during that day.
d. Information concerning a special nature of the account (flags).
e. Date of last transaction.

(2) Checks on teller operations.

a. The old pass-book balance is the same as that in the file.
b. The interest added by the teller is the same as the unposted interest on the account.
c. The new pass-book balance is the same as the central memory's balance, following a transaction to bring the pass-book up to date.

(3) Updating of the account by writing buffer information, such as:

a. Bank and/or branch at

which the transaction is being performed.
b. Type of transaction.
c. Changes to flag information.
d. Date.
e. Teller identification.
f. Transaction amount.

(4) Locking of the teller's posting machine until the data has been processed correctly.

(5) Signalling of errors to the tellers.

Teller Station

The teller has two devices at his disposal for automatic account processing--a posting machine and a Display Unit (figure 2). In TELLERTRON Serial One, the posting machine is a Burroughs "Sensimatic". This machine allows for entry of data into the buffer, printing of a transaction record, printing on the customer's pass-book, and punching of a paper tape. The Sensimatic data is used by TELLERTRON in the form of 13 decimal columns. The teller Display Unit uses decimal projection indicators, and contains storage and indicator-driving circuit cards.

To interrogate the teller posting machines, pulses are delivered sequentially to the columns of all Sensimatics by the Central Processor, which then selects the decimal data from a particular teller for processing. To display, information is delivered to the Display Units of all tellers, along with a gate to enable the correct digit. A pulse is then delivered only to a particular teller, so that his display unit is the only one which displays the information sent. The Display Units are located at distances of up to 500 feet from the Central Processor.

Branch Operation

Tellers at remote branch offices may communicate with the Central Processor via standard telephone lines and A T & T Dataphone equipment (figure 6). A Branch Central unit (figure 7) is required at each remote office to organize and control the data traffic. This device contains a scanner to accommodate branch teller priorities, a data serializer, and checking facilities. Only one branch teller at a time may communicate with the Central Processor, an operation which requires 140 milliseconds.

A Branch Processor unit is employed at the home office in order to control the communication links with all branches. Modified full-duplex operation (transmitting to one branch while receiving from another) is employed. The Branch Processor also contains 1 mc recirculating magnetostrictive delay lines for temporary storage of data for and from branch tellers, so that Central Processor operations need not be modified due to the slow (1000 bits/second) rate of data transmission over the telephone lines.

Perforated Tape Operation

In case of TELLERTRON failure, perforated tape is created by each teller (via his posting machine) while the TELLERTRON is inoperative, and the Main File is later updated by entering the perforated tape into the system. In addition, paper tape from other banks may also be entered into the TELLERTRON system. This latter facility enables banks to have their accounting performed by off-line equipment of the bank possessing the TELLERTRON system.

Reliability, Redundancy, Checking, and Maintenance Features

Reliability and Checking

The TELLERTRON System is built on the basic philosophy that no single electronic component failure can disable service to the entire bank. Consequently, the majority of the system is built in duplicate. In the normal case, when both of the duplicate systems are operative, one channel is designated the "main" channel and performs all system operations. The second channel operates in parallel with the main channel, and the results of the two channels are compared. Both channels must independently arrive at the same answer or operation cannot proceed. The duplication philosophy provides for two independent access arms on the Disk Storage Unit, two core buffers, and redundant power supplies.

Checking of teller operations has been previously described. Operationally, in addition to the comparison of the two redundant systems described above, the following checks are made on the transfer and storage of information:

(1) Parity bits are carried with all information, and are checked each time information is used or handled.

(2) A check is made throughout the machine for any alteration of information which produces an illegal binary-coded-decimal code configuration.

(3) When information is read from the teller posting machine, it is checked to ensure that one and only one line out of the ten decimal lines is active. This effectively checks each posting machine, which is primarily a mechanical device.

(4) Information is sent to the Teller Display Unit in decimal form so that a one-out-of-ten check can also be made at the teller station.

(5) A read-after-write check is performed each time information is written onto a magnetic surface (drum, disk, tape) in the system. The source of information is not destroyed until after this check has been made, so that information can be rewritten if necessary.

The cardinal rule of TELLERTRON is that no account information in the Main File may be erroneously altered. An account in the Main File is not erased and rewritten until it has been determined that a transaction on that account has been completed without error either by the teller or the system. In the event that any errors are detected (as described above), TELLERTRON causes the operation to be repeated. After three unsuccessful attempts, TELLERTRON then takes one of the following courses of action:

(1) Determines that a teller is in error and directs him to repeat the transaction.

(2) Determines that one of its two channels is defective, and so informs the maintenance personnel.

(3) Determines that a defect exists in the system such that it cannot proceed without danger of erroneously altering Main File information, and so informs maintenance personnel.

Panel Design

Four maintenance panels are built into the Central TELLERTRON complex (exclusive of the Master Control Console) and one maintenance panel is built into each Branch Central (figure 7). These panels are so designed that maintenance personnel may operate the TELLERTRON, or any of its sub-units, in either of two test modes, the "cycle" test and the "single step" test.

In the "cycle" test, TELLERTRON is exercised in a repetitive or cyclic operation, at speeds convenient for oscilloscope monitoring, and an illuminated display of results is presented. The "cycle" test is used to completely exercise the system at the beginning of each working day in order to establish that the system is functioning properly. The "cycle" test is also used to locate faulty subunits, and to exercise these units in order to locate any faulty component parts.

The "single step" test allows TELLERTRON to be operated one-bit-at-a-time, or one-operation-at-a-time, under manual control. Data is inserted and extracted manually, and "single step" operation is observed on maintenance panel indicators.

The maintenance panels are organized, insofar as is possible, as functional flow or block diagrams of the TELLERTRON units, to aid in ease of maintenance operations. Functional color coding of operational and test switches is observed.

Digital Modules

The digital logic of TELLERTRON was designed and constructed using the standard transistorized digital modules manufactured by Computer Control Company (figure 8). Approximately 1700 plug-in modules of the T-PAC and S-PAC series, representing approximately 15,000 transistors and 75,000 diodes, are used in TELLERTRON Serial One. Some vacuum tubes are also used.

The T-PAC Series uses synchronous (1 mc clock rate) modules, the basic feature of which is a universal logical element (Model LE-10). All functions of three binary variables, and most functions of four or five binary variables may be implemented on the "and-or-not" structure of a single logical element; system design and minimization is thus easily accomplished. Serial magnetostrictive delay line storage is also provided. Over five years of operating experience representing many millions of module-hours, offer proof of T-PAC reliability.

The S-PAC series uses "nand" logic, available in a variety of module configurations. The use of clamped circuits throughout the S-PAC series provides maximum protection against noise. The variety of S-PACs available permits efficient and economical utilization of circuits, and provides the flexibility necessary to accommodate a variety of peripheral problems, such as exist in TELLERTRON. Conservative design has resulted in proven reliability of the S-PAC series.

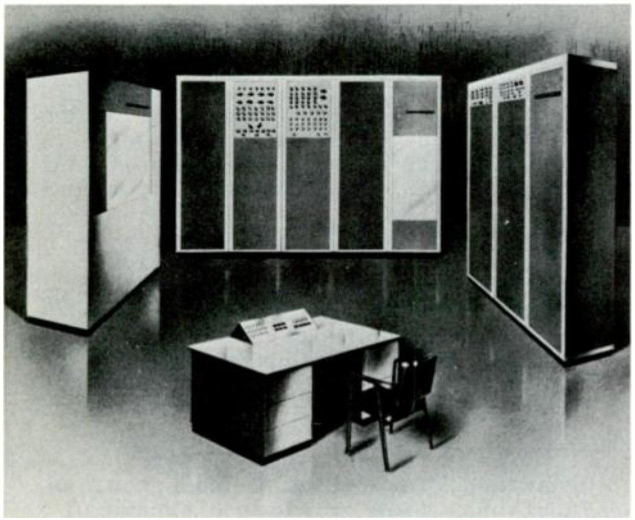


Fig. 1. TELLERTRON serial one.

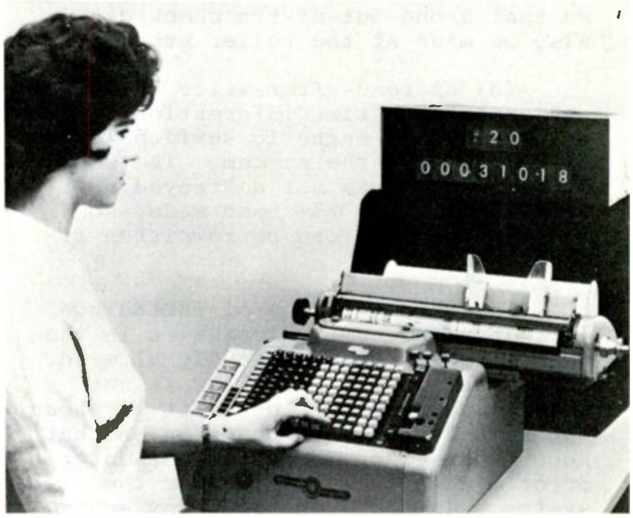


Fig. 2. Teller station equipment.

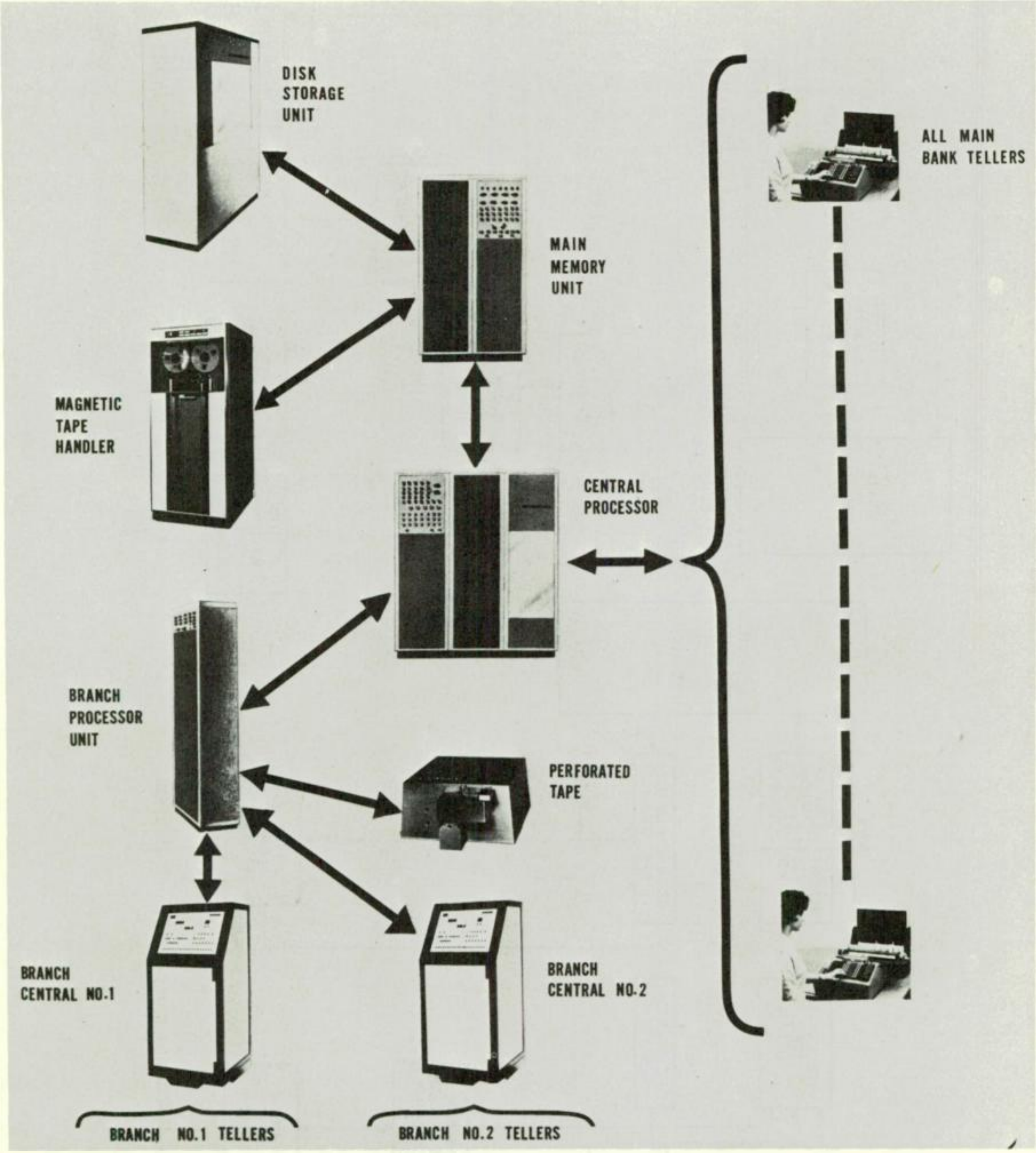


Fig. 3. TELLERTRON functional block diagram.

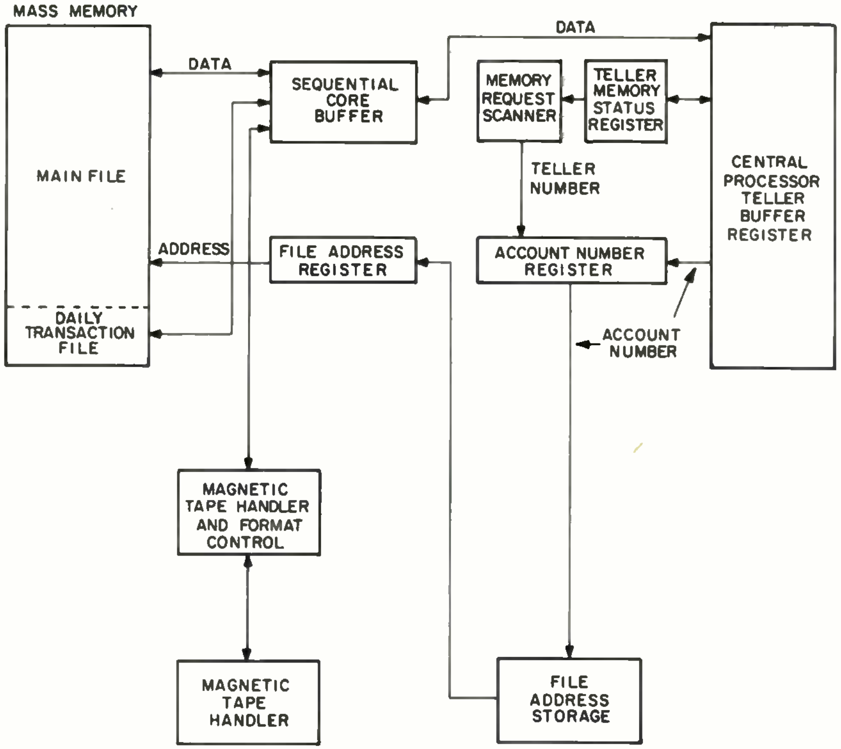


Fig. 4. Main memory unit, block diagram.

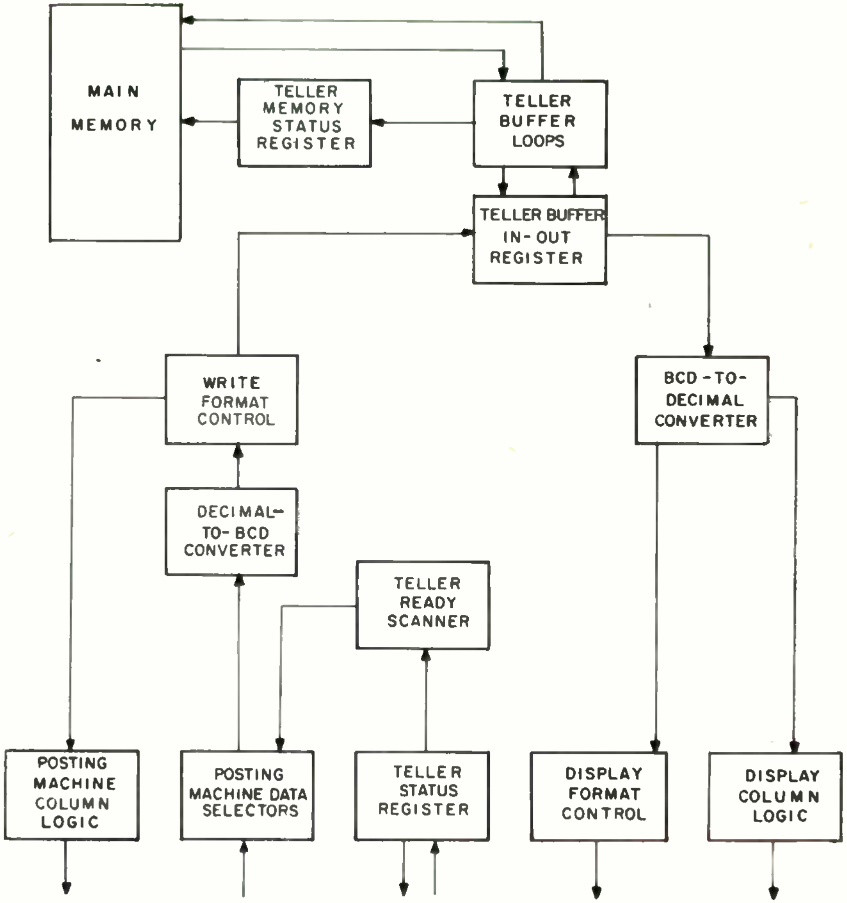


Fig. 5. Central processor unit, block diagram.

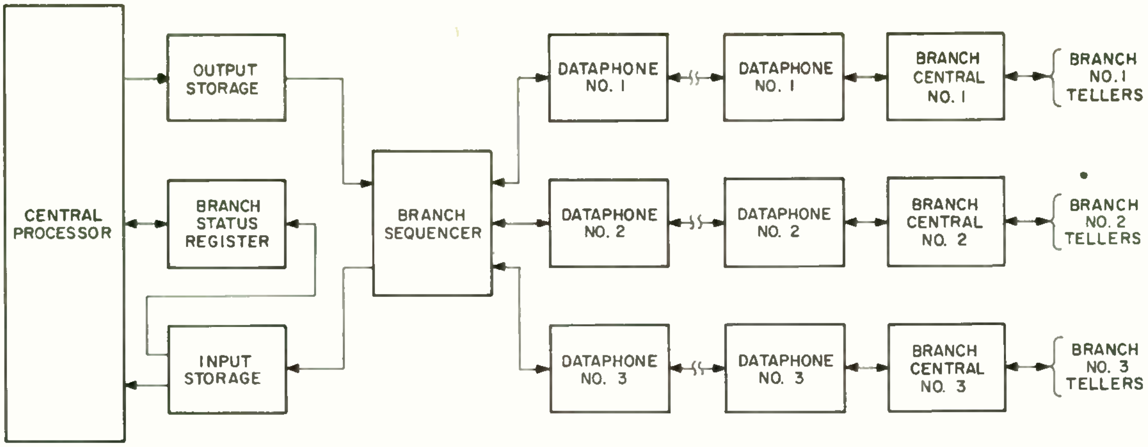


Fig. 6. Branch operations, block diagram.

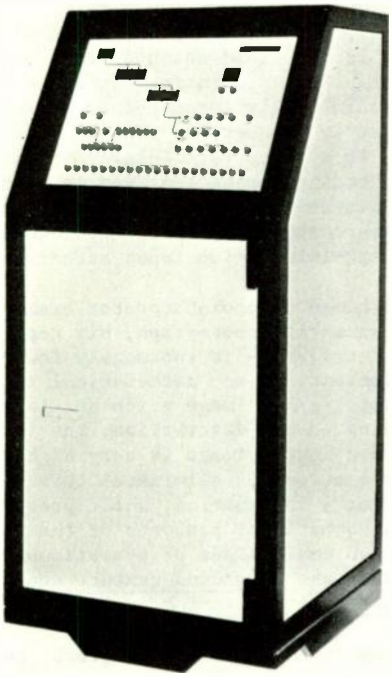


Fig. 7. Branch central unit.

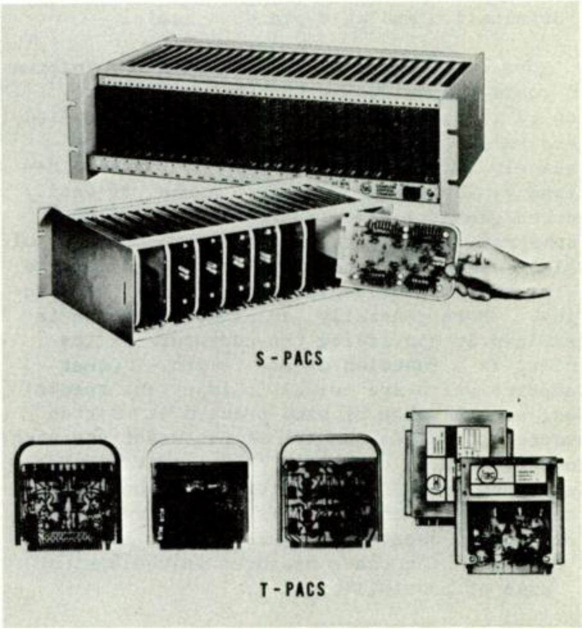


Fig. 8. S-PAC and T-PAC series of digital modules.

AUTOMATIC RECOGNITION TECHNIQUES APPLICABLE TO HIGH-INFORMATION PICTORIAL INPUTS

Azriel Rosenfeld
Budd Electronics, A Division of The Budd Co., Inc.
Long Island City 1, New York

Summary. This paper describes a class of automatic recognition techniques which are applicable to inputs having high information content, such as aerial photographs. These techniques are based on statistical analysis of the input image. This analysis is used to determine the boundaries of the conspicuous figures which the image contains and to generate a simplified description of the "visual texture" of various parts of the image. * * * *

Most of the past work in the field of automatic shape and pattern recognition has involved inputs having a relatively low information content. In those cases which have received the most concrete attention, the information content involved is so low that it becomes practical to use a process of matching or correlation for recognition purposes, in which the actual input is compared with a duplicate of every possible input. Typical of these cases is the problem of recognizing characters coming from a specific set of type fonts and presented in a standard position and orientation and at a standard scale.

More complicated cases, in which the information content involved is too high for simple matching to be practical, have also been studied extensively. Even these cases, however, are relatively simple; in them, the input can be described by specifying a limited number of easily measured geometrical parameters. For example, an arbitrary point pattern (or figure made up of straight lines) is determined if certain lengths of line segments or angles between them are speccified. More generally, an arbitrary shape is determined by specifying the curvature of its boundary as a function of arc length. Other parameters which are useful in important special cases, such as that of hand printed or written characters, include numbers of intersections with a raster, numbers of branch points of various orders, and so on. Recognition can be effected in cases such as these by first measuring the parameters in question for the actual input, and then matching these measured values against the range of possibilities.

Using geometrical parameters for recognition purposes is practical in cases where the input picture involves relatively clearcut shapes (=simply connected regions) or point patterns which can be easily "extracted" from the background of the input as a whole. This approach breaks down, however, when the input is of a more complicated nature - for example, when it is a photograph which contains a large amount of detailed pictorial information. Practical problems involving such inputs arise in connection with the interpretation of aerial photographs for military intelligence or other pruposes. In dealing with such a photograph, conventional techniques of shape and pattern recognition are not directly applicable, since there is no natural distinction between figures (= shapes or patterns) and background in the photograph.

Considerable work has been done on the simplification of the information content of aerial photographs by, e.g., successive processes of averaging and differentiation. This approach does have the effect of extracting clearcut figures from the photograph. However, these figures do not in general provide important information about the content of the photograph. For example, if the photograph shows an urban area, the extracted figures will be highly complex curves, the exact shapes of which will be almost entirely unimportant.

In this paper, an approach to automatic recognition is described which is applicable to inputs having high information content. This approach was initially suggested by an analysis of methods of strategic aerial photographic interpretation. It is thus particularly applicable to automatic photointerpretation, and is perhaps most easily illustrated by examples drawn from this area; however, it is generally applicable to arbitrary high-information input situations.

When a human photointerpreter examines and reports on an aerial photograph, his report -- no matter how detailed -- is enormously lower in information content, in the mathematical sense, than the photographic image which he is describing. Compared to his description, the information contained in the image is very highly redundant. The interpreter eliminates this redundant, or "irrelevant", information, while preserving the important informational features of the image, by performing two basic kinds of operations which may be referred to as background texture analysis and conspicuous figure extraction.

In studying an aerial photograph, the human interpreter will generally make an immediate subdivision of the photograph into regions each of which has a substantially uniform "visual texture". These regions will roughly represent different types of terrain, such as bodies of water, uncultivated and cultivated land, and urban and suburban areas. The notion of visual texture may be thought of as depending on the average value of the image density over the given region and on the average spatial frequency with which contrasts of various degrees occur. Most of the fine detail present in the image will serve only to contribute to this "texture recognition".

In addition to perceiving the photograph to be made up of regions having various visual textures, the human interpreter will also generally see certain relatively conspicuous "figures" as "standing out" from the background which these regions comprise. The conspicuousness of a figure depends primarily on its size and on the degree to which it contrasts with the adjacent portions of the photograph. Conspicuous figures are thus likely to be high contrast discontinuities seen against the background of a low-contrast texture, or sharply defined boundaries between regions which have significantly different textures. Another important factor in figure conspicuousness is the simplicity of its shape (or its regularity, if it is a pattern rather than a shape). The simplest figures, such as those which involve straight lines and simple straight line combinations (parallels, perpendiculars, etc.), will tend to be "disproportionately" conspicuous.

The foregoing remarks may be illustrated by references to Figures 1 through 4. These figures are aerial photographs at assorted scales showing various types of terrain. Each of these photographs can be relatively easily subdivided into fairly well-defined "components" each having in some sense a uniform visual texture. Conspicuous figures of both of the kinds mentioned just above are also present. It will be noted in particular that certain straight line configurations are conspicuous as figures in spite of the fact that they involve lower levels of contrast than other, less simply shaped, configurations.

In order to automate these operations of background texture analysis and conspicuous figure extraction, it is necessary to formulate the concepts of visual texture and figure conspicuousness in objective, mathematical terms. As it turns out, both of these concepts can be formulated in terms of certain statistical properties of the photographic image. Preliminary simulation experiments indicate that the resulting statistical model does indeed lead to notions of texture and conspicuousness which correspond closely to the judgments of human observers.

As hinted earlier, the visual texture of a portion of a photographic image seems to be related to the distributions of densities and density contrasts in the given portion. For the present purposes, the contrast at a point may be defined as the absolute value of the derivative of the density in the direction of the density gradient.

One way of describing an arbitrary statistical distribution is by computing its moments of various orders. In this connection, the moments of higher order are of decreasing importance. Successive "approximations" to the description of a distribution can be obtained from the first few moments. The most useful information about a given distribution will generally be provided by the mean (= the first moment) and the variance (equivalent to the second moment).

In the present case the mean density and the mean contrast frequency (which may be thought of as the mean "level of detail") will provide the most useful statistical information. The variance of the density will usually be closely related to the mean contrast frequency, since a high density variance means a high degree of fluctuation in the density, which is essentially equivalent to a high frequency of contrasts. The variance of the contrast frequency corresponds to the degree of "evenness" with which contrasts are distributed, and thus provides more detailed information about the appearance of the image. Alternatively, more detailed information can be obtained by computing mean frequencies for each of several levels of contrasts -- that is, by giving separate consideration to, e.g., high, medium and low contrasts. Experimental work which has been described elsewhere[1] suggests that this type of detailed contrast frequency analysis provides enough information about the visual texture of the image to make possible the automatic identification of many basic terrain types.

The visual texture of any given portion of a photograph can be objectively described in terms of statistical parameters such as those just discussed. This texture analysis can then be used to generate an overall description of the photograph. This involves the recognition of regions of uniform texture within the photographic image. In this connection, it should be pointed out that the degree of uniformity of a given region depends strongly on the size of the portion over which the statistical parameters in question are averaged or measured. Reference to Figures 1 - 4 reveals that some uniform-appearing regions are statistically uniform only if the parameters are averaged over portions which are comparable in size to the region itself, while other regions remain uniform even when averaging is performed over very small portions. Textural uniformity is thus a function of what might be called the "level of coarseness" over which the texture is measured.

The extraction of conspicuous figures from the photographic image also involves the concept of visual texture, since the boundaries between adjacent areas having significantly different textures will define such figures. Examples of boundaries defined by "texture contrasts" of various types are shown in Figure 5. These boundaries may be roughly classified into two types, depending on whether the associated texture changes involve primarily differences in mean density or in mean contrast frequency.

It will be observed that for a boundary of either of these types to be conspicuous, a marked density change must occur at the boundary, and there can be no comparable density change nearby on at least one side of the boundary. This is an obvious condition for conspicuousness in a boundary which primarily involves a change in mean density and which bounds a region of appreciable size. On the other hand, suppose that a boundary is primarily associated with a conspicuous contrast frequency change. The images on the two sides of it must then have markedly different levels of detail. It follows that the contrasts present on one side must be either significantly lower or more sparsely scattered than those on the other side. There will thus be relatively high contrasts coming up to the boundary on this other side, but no comparable contrasts within an appreciable distance on the first side. The criterion for conspicuousness just formulated clearly applies also to conspicuous figures of the other type mentioned earlier, namely high-contrast density discontinuities in a region of low-contrast texture.

The above discussion must be made quantitative by specifying the degree to which the contrast associated with a boundary must exceed the nearby contrasts, and by making precise the notion of "nearby". These conspicuousness parameters should probably not be absolute constants, but should depend on the texture(s) of the portions of the image adjacent to the boundary. In initial experimental work, however, it has been found possible to obtain useful results using constant parameter values.

Two experimental numerical simulations of a simple mathematical model for boundary conspicuousness have been performed manually using the photograph strips shown in Figure 5. For simplicity, the measurements made were restricted to one line along each strip, as indicated in the Figure. Image densities along each of these lines were sampled at intervals of 1/32" using a densitometer having an aperture 1/32" in diameter. Before further processing, the density data were converted to a linear scale. These corrected data are shown in Tables 1 and 2.

For the purposes of these experiments, the following conspicuousness criterion was used: The differences between each density datum and the one three places behind it were computed. The values of these skipped differences are shown in the tables. Skipped, rather than adjacent, differences were used to compensate for the fact that the finite densitometer aperture tended to flatten out abrupt density changes when these were measured between adjacent density readings. A point was judged to be part of a conspicuous boundary if the skipped density difference at that point exceeded by 40 or more the skipped differences at each of the third, fourth, fifth and sixth preceding points or at each of the third, fourth, fifth and sixth succeeding points. Comparison with the more immediately preceding or succeeding points would not have given a correct measure of conspicuousness, since the skipped differences at these points are interleaved with that at the given point. The limitation of comparisons to six places away from the given point corresponds to a minimum "appreciable" region size of 6/32". The critical difference of 40 is an arbitrary value corresponding to the range of the arbitrary linear density scale.

A comparison of Figure 5 and the tables reveals that the above defined simple conspicuousness criterion correctly identified all of the "obvious" boundaries. These included boundaries between regions having a variety of average densities and average contrast frequencies. In particular, this criterion correctly located the beginning of the highly detailed region on Strip B, while defining only one "spurious boundary point" inside that region. Indeed, only three spurious boundary points (namely, Points Nos.82 and 162 on Strip A, and 152 on Strip B) were picked out by this criterion. These few exceptions resulted from the fact that the densitometer "saw" only a 1/32" wide strip along the scanned line. There will in general be a few conspicuous boundary points along such a thin strip which lose their conspicuousness relative to a wider strip such as those shown in Figure 5. These points can thus be eliminated by extending the criterion for conspicuous boundaries from one to two dimensions.

A definitive criterion for boundary conspicuousness must necessarily be two-dimensional, since conspicuousness is influenced by the shape of the boundary. This influence may be regarded as a lowering of the standard of conspicuousness in cases where the resulting boundary would have a particularly simple shape, as defined (say) by the fewness of changes in direction along the boundary.

The techniques described in this paper make it possible to automatically generate a simplified description of a given complex picture which corresponds, at least in general terms, to what a human observer might report in describing the picture. The key step of statistical analysis makes use of the very multiplicity of information which the picture contains as a basis for simplification. It is felt that this statistical approach or its equivalent will become increasingly important in the development of cognitive systems to handle highly complex pictorial inputs.

[1] In a paper entitled Automatic Recognition of Basic Terrain Types on Aerial Photographs, presented at the 1961 East Coast Conference on Aerospace and Navigational Electronics, and to be published shortly in Photogrammetric Engineering.

TABLE 1

DENSITY DATA - STRIP A

(Points marked * are conspicuous boundary points corresponding to M=40)

Datum number	Datum	Skipped difference
1	117	
2	92	
3	72	
*4	59	58
5	59	33
6	53	19
7	53	6
8	53	6
9	49	4
10	59	6
11	72	19
12	72	23
*13	139	80
*14	160	88
*15	163	91
16	139	0
17	135	25
18	139	24
19	143	4
20	139	4
21	127	12
22	139	4
23	172	33
24	160	33
25	160	21
26	150	22
27	160	0
28	160	0
29	163	13
30	165	5
31	154	6
32	150	13
33	172	7
34	181	27
35	139	11
36	139	33
37	143	38
38	175	36
39	172	33
40	172	29
41	143	32
42	154	18
43	122	50
44	154	11
45	143	11
46	122	0
*47	0	154
*48	45	98
*49	45	72
*50	59	59
51	68	23
52	77	32
53	70	11
54	70	2
55	83	6
56	79	9
57	79	9
58	83	0
59	59	20
60	41	38
61	41	42
62	41	18
63	41	0
64	45	4
65	45	4
66	45	4
67	53	8
68	81	36
69	68	23
70	59	6
71	75	6
72	117	49
*73	163	104
*74	160	85
75	154	37
76	122	41
77	139	21
78	112	42
79	99	23
80	157	18
81	163	51
*82	150	51
83	143	14
84	143	20
85	143	7
86	143	0
87	139	4
88	143	0
89	154	11
90	147	8
91	150	7
92	160	6
93	163	16
94	160	10
95	172	12
96	166	3
97	157	3
98	160	12
99	163	3
100	157	9
101	154	3
102	147	16
103	160	3
104	169	15
105	172	25
106	139	21
107	135	34
108	135	37
109	122	17
110	117	18
111	139	4
112	143	21

TABLE 1 Cont'd.

Datum number	Datum	Skipped difference	Datum number	Datum	Skipped difference
113	139	22	176	113	10
114	150	11	177	111	5
115	143	0	178	116	13
116	139	0	179	116	3
117	139	11	180	121	10
118	150	7	181	115	1
119	139	0	182	103	13
120	157	18	183	108	13
121	154	4	184	114	1
122	143	4	185	117	14
123	150	7	186	122	14
124	160	6	187	117	3
125	157	14	188	112	5
126	147	3	189	115	7
127	139	21	190	117	0
128	135	22	191	112	0
129	143	4	192	115	0
130	147	8	193	111	6
131	147	12	194	117	5
132	160	17	195	115	0
133	163	16	196	68	43
134	163	16	*197	45	72
135	160	0	198	70	45
136	157	6	199	77	9
137	160	3	200	75	30
138	175	15	201	72	2
139	154	3	202	77	0
140	181	21	203	83	8
141	157	18	204	83	11
142	163	9	205	127	50
143	154	27	*206	169	86
144	157	0	*207	166	83
145	143	20	208	169	42
146	139	15	209	169	0
147	147	10	210	169	3
148	143	0	211	169	0
149	139	0	212	160	9
150	139	8	213	106	63
151	143	0	*214	75	94
152	150	11	*215	77	83
153	147	8	216	75	31
154	117	26	217	81	6
155	117	33	218	87	10
*156	92	55	219	92	17
157	56	61	220	95	14
158	56	61	221	72	15
159	53	39	222	68	24
160	62	6	223	62	33
161	106	50	224	70	2
*162	107	54	225	65	3
163	113	51	226	81	19
164	112	6	227	111	41
165	103	4	*228	150	85
166	103	10	*229	154	73
167	99	13	230	157	46
168	101	2	231	154	4
169	115	12	232	160	6
170	117	18	233	160	3
171	117	16	234	160	6
172	106	9	235	163	3
173	103	14	236	166	6
174	106	11	237	157	3
175	103	3	238	154	9

TABLE 2

DENSITY DATA - STRIP B

(Points marked * are conspicuous boundary points corresponding to M=40)

Datum number	Datum	Skipped difference
1	163	
2	163	
3	181	
4	157	6
5	139	24
6	143	38
7	135	12
8	135	4
9	154	11
10	157	22
11	160	25
12	172	18
13	143	14
*14	41	119
*15	59	93
*16	77	66
17	83	42
18	85	26
19	78	1
20	56	27
21	45	40
22	53	25
23	59	3
24	59	14
25	59	6
26	68	9
27	62	3
28	68	9
29	45	23
30	45	17
31	83	15
32	85	40
33	88	43
34	83	0
35	77	8
36	77	11
37	77	6
38	85	8
*39	166	89
*40	157	80
*41	139	54
42	139	27
43	139	18
44	143	4
45	139	0
46	131	8
47	131	12
48	122	17
49	127	4
50	117	14
51	117	5
52	122	5
53	117	0
54	117	0
55	122	0
56	122	5
57	122	5
58	117	5
59	117	5
60	117	5
61	117	0
62	117	0
63	127	10
64	131	14
65	122	5
66	127	0
67	131	0
68	127	5
69	127	0
70	122	9
71	117	10
72	131	4
73	135	13
74	131	14
75	131	0
*76	45	90
*77	45	86
*78	49	82
79	56	12
80	62	17
81	85	36
82	85	29
83	77	15
84	92	7
85	99	14
86	95	18
87	111	19
88	117	18
89	117	22
90	112	1
91	112	5
92	112	5
93	111	1
94	117	5
95	127	15
96	122	11
97	122	5
98	117	10
99	88	34
100	83	39
101	65	52
102	65	23
103	65	18
104	83	18
105	83	18
106	85	20
107	83	0
108	83	0
109	83	2
110	88	5

TABLE 2 Cont'd.

Datum number	Datum	Skipped difference	Datum number	Datum	Skipped difference
111	87	4	143	83	29
112	77	6	144	41	76
113	70	18	145	35	82
114	78	9	146	83	0
115	59	18	147	92	51
116	53	17	148	127	92
117	45	33	149	157	74
118	41	18	150	122	30
119	41	12	151	92	35
120	41	4	*152	83	74
121	36	5	153	106	16
122	41	0	154	79	13
123	41	0	155	81	2
124	56	20	156	106	0
125	53	12	157	106	27
126	59	18	158	59	22
127	65	9	*159	49	57
128	85	32	160	117	11
*129	143	84	161	92	33
130	85	20	162	112	63
131	117	32	163	117	0
132	83	60	164	135	43
133	112	27	165	122	10
134	41	76	166	99	18
135	78	5	167	92	43
136	72	40	168	83	39
137	41	0	169	106	7
138	106	28	170	122	30
139	92	20	171	112	29
140	112	71	172	83	23
141	117	11	173	79	43
142	117	25	174	59	53

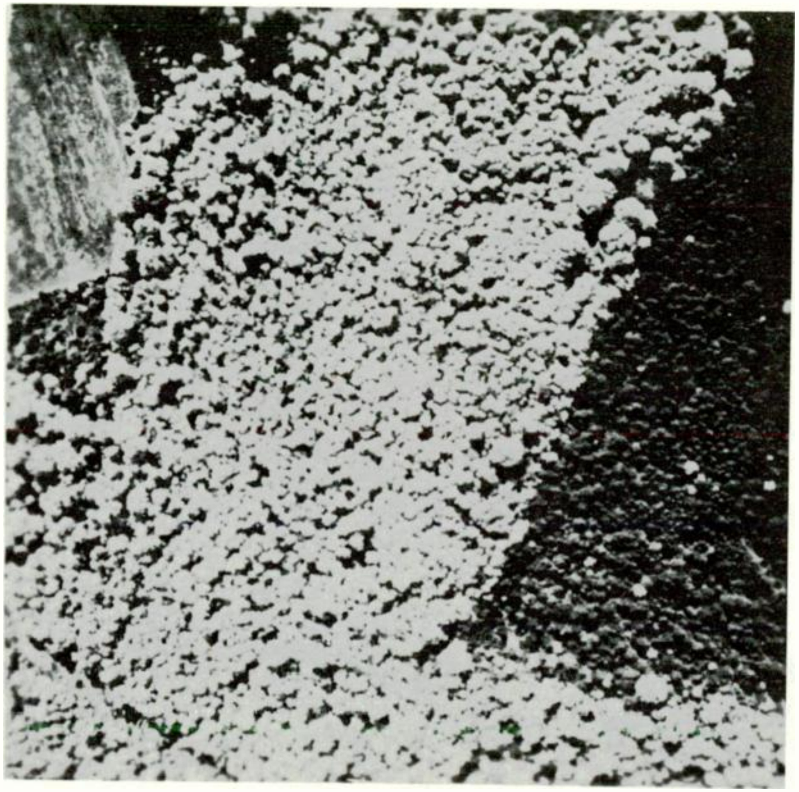


Fig. 1. 1:3300

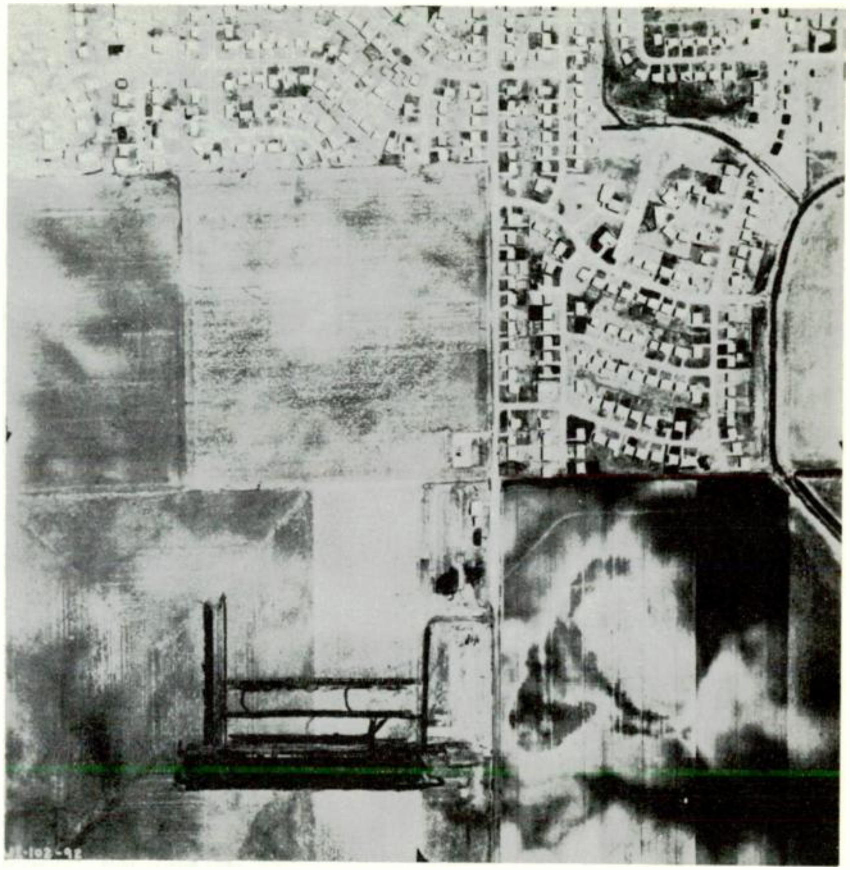


Fig. 2. 1:9600

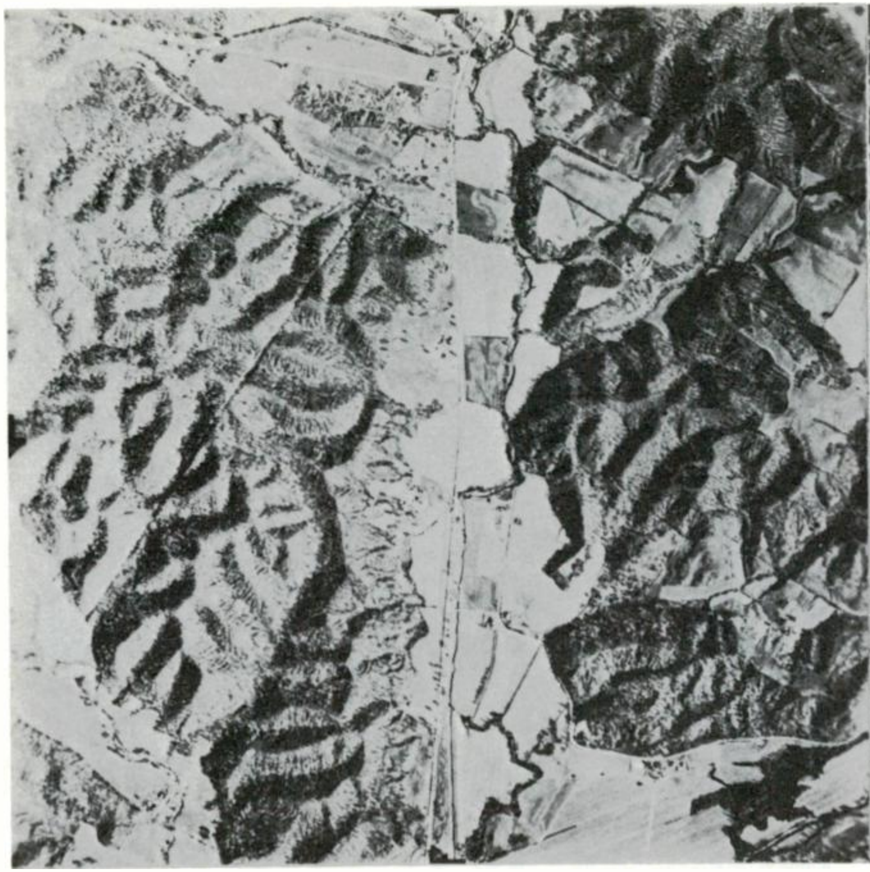


Fig. 3. 1:29000.

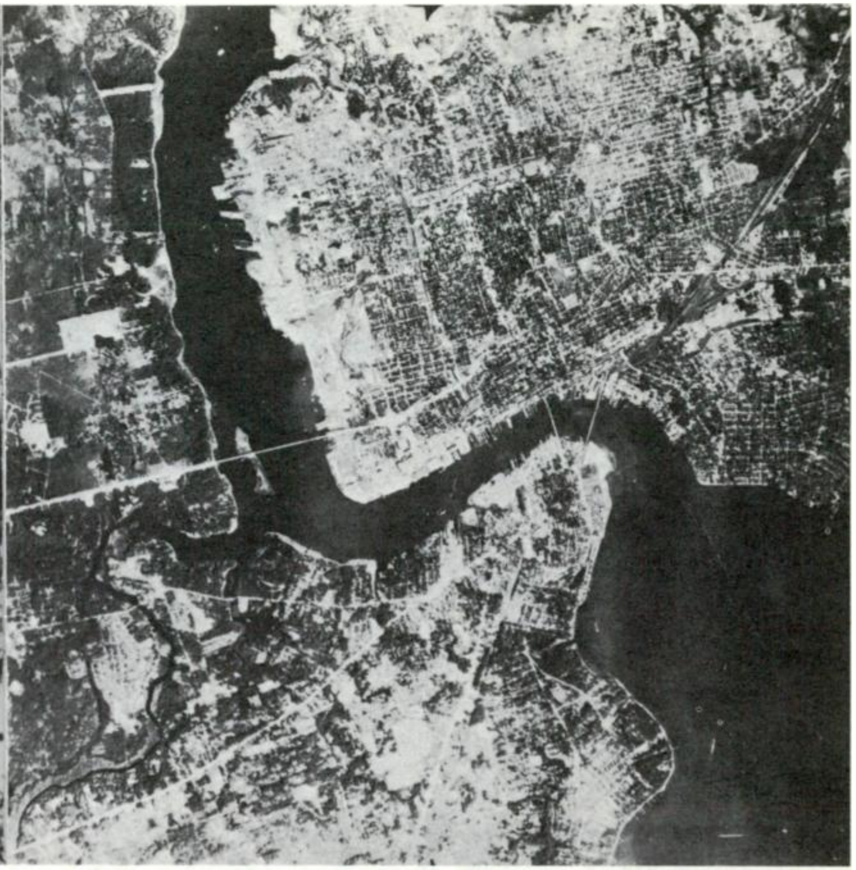


Fig. 4. 1:105000.

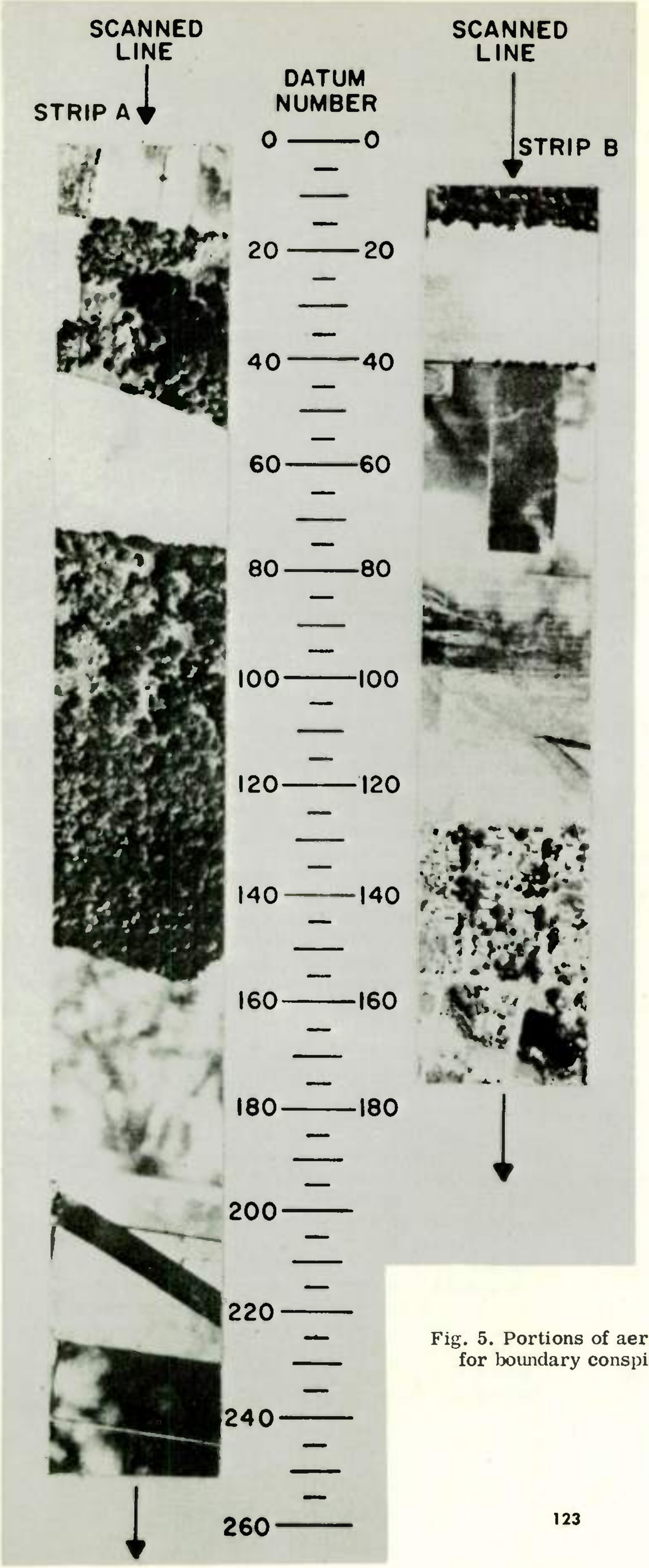


Fig. 5. Portions of aerial photograph used to test model for boundary conspicuousness.

ADAPTIVE DECISION ELEMENTS TO IMPROVE THE RELIABILITY OF REDUNDANT SYSTEMS

W. H. Pierce

Solid State Devices Department
Westinghouse Research Laboratories
and
Carnegie Institute of Technology
Pittsburgh, Pennsylvania

Summary

Redundant but unreliable binary or analog signals are used most effectively when each signal's error probability is used to compute a statistical estimate of the correct signal. When errors in the inputs are independent, the statistical estimates can be made by simple summing circuits., Adaptive circuits are proposed which give the most reliable inputs the largest weights. Reliability of each input is estimated by comparing it with: (1) an answer supplied externally during a check routine, or (2) the output of the decision element. Adaptive decision elements are most valuable when used throughout systems with quite unreliable signals.

Introduction

In 1953 von Neumann proposed a particular configuration of Sheffer-stroke organs* which would make a reliable system out of unreliable parts.[1] Since then, several different approaches have been proposed. They are:

1. A redundant configuration is postulated, and then shown to improve reliability. Examples are the majority organ and Sheffer-stroke organ configurations of von Neumann[1], and the relay networks of Moore and Shannon[2].

2. The coding theory developed for unreliable communications channels is applied to computing systems. Examples are Hamming codes[3], which may be used directly for memory and ensuring transfer of digits between registers. A partial application of these codes to combinatorial and sequential networks has been given by Armstrong[4].

3. The logical network may be designed so as to be insensitive to the unreliability expected. An example is the work of Verbeek[5], who has studied the synthesis of reliable logic from devices whose Boolean output is predictable only for some of the possible inputs.

4. Statistical decision theory may be systematically applied to determine the optimum use of redundant information, and then adaptive techniques used to implement the statistical decisions.

This paper is an exposition of the statistical decision theory approach to the use of redundancy, and the adaptive decision elements which are the inevitable product of the decision theory approach. These adaptive decision elements are important for two reasons: (1) they may be the most economical means of using redundant signals, and (2) they define the optimum use of redundant signals, and thereby serve as a standard for evaluating how efficiently other devices use redundant signals.

The decision elements themselves have the function of using all m redundant inputs so as to produce a more reliable single output. The pattern of using redundant devices throughout a system is implicit in the multiplexing scheme of von Neumann[1], and has been used by many workers subsequently[6,7,8,9]. After each operation by redundant computing elements, the m redundant signals enter a restoring organ composed of m decision elements. Each output of the restoring organ comes from a separate decision element; each decision element receives all m inputs. The restoring organ, and its placement in redundant systems, are illustrated in Figures 1 and 2.

The Relation Between Vote-taking and Decision Theory

The application of decision theory in optimizing the use of redundant information in electrical circuits can be approached in two ways: A simple electrical circuit could be proposed, and then its parameters optimized, or else the optimum use of the redundant information could be established, and then circuits designed to carry out the required functions. It turns out that the actual approach is immaterial, since both lead to the same solution. The solution is a simple circuit which takes a linear combination of the inputs, and then processes this sum to obtain the output. The processing is a threshold for binary circuits and an attenuation for

* A Sheffer-stroke organ has Boolean inputs A and B, and output $(\overline{AB})$.

analog circuits. This simple circuit is called a vote-taker; it is probably the most useful type of decision element for using redundant information. Its usefulness follows from the close relation between vote-taking and decision theory for independent errors in the inputs.

Suppose, for example, that the redundant circuits of an analog computer have computed the quantities $x_1, x_2, \dots x_m$, which are the correct signal y corrupted by independent, zero-mean Gaussian noises. Thus

$$x_i = y+n_i \qquad \text{for} \qquad i = 1,2, \dots m \tag{1}$$

where n_i is distributed $N(o, \sigma_i^2)$ and n_i is independent of n_j, $i \neq j$.

These unreliable, redundant input signals are to be used to make a maximum likelihood (most probable) estimate of y, the estimate to be denoted by $\hat{y}$. A straightforward calculation of the a posteriori probability of y, given the input vector $x = (x_1, x_2, \dots x_m)$ is given in proof 1 of the appendix. It shows that $p(y|x)$ has a Gaussian distribution with mean and variance

$$E(y) = \left(\sum_{i=o}^{m} x_i/\sigma_i^2\right) \Bigg/ \left(\sum_{i=o}^{m} 1/\sigma_i^2\right)$$

$$\sigma^2(y) = 1 \Bigg/ \left(\sum_{i=o}^{m} 1/\sigma_i^2\right)$$

where the a priori probability of y has been assumed to be $N(x_o, \sigma_o^2)$.

Consequently, the maximum likelihood estimate (and also least mean square estimate) of y is

$$\hat{y} = \frac{\sum\limits_{i=o}^{m} a_i x_i}{\sum\limits_{i=o}^{m} a_i} \tag{2}$$

where $$a_i = \frac{1}{\sigma_i^2} .$$

Thus the maximum likelihood estimate is just a linear combination of the inputs, and leads to the decision element of Figure 3, with vote weights a_i given by Equation 2.

The optimum values of the a_i for independent non-Gaussian distributions on the n_i may be readily derived for a mean square error criterion. Let σ_D^2 be the expected value of $(y-\hat{y})^2$. Then

$$\sigma_D^2 = E\left[y - \left(\sum_{i=o}^{m} a_i x_i\right) \Bigg/ \left(\sum_{i=o}^{m} a_i\right)\right]^2 .$$

Using Equation 1,

$$\sigma_D^2 = E\left[\left(\sum_{i=o}^{n} a_i n_i\right) \Bigg/ \left(\sum_{i=o}^{n} a_i\right)\right]^2 .$$

Now assume n_i is independent of n_j, $i \neq j$, and that $\sigma^2(n_i) = \sigma_i^2$. Then

$$\sigma_D^2 = \left(\sum_{i=o}^{m} a_i^2 \sigma_i^2\right) \Bigg/ \left(\sum_{j=o}^{m} a_j\right)^2 . \tag{3}$$

(In the special case when all $\sigma_i^2 = \sigma_1^2$, and all $a_i = 1$, then σ_D^2 becomes σ_1^2/m, so the rms noise goes down by a factor $1/\sqrt{m}$ for a redundancy of m.) The optimum vote weights occur where the partial derivatives are zero.

$$\frac{\partial \sigma_D^2}{\partial a_j} = \frac{2}{\left(\sum\limits_{i=o}^{m} a_i\right)^3} \left[\sum_{i=o}^{m} a_i (a_j \sigma_j^2 - a_i \sigma_i^2)\right]$$

Since the second derivative is positive (assuming all $a_i \geq o$), the solution

$$\frac{\partial \sigma_D^2}{\partial a_j} = o \implies a_j = \frac{K}{\sigma_j^2} \tag{4}$$

$$j = o, 1, \dots m$$

gives the minimum value of σ_D^2.

Equations 2 and 4 have shown two important aspects of vote-taking. Equation 2 showed that, for independent Gaussian noises in the inputs, the circuit of Figure 3 is the circuit which makes the maximum likelihood (and least mean square) estimate of y, provided a_i is inversely proportional to the variance of the noise in the i-th input. Equation 4 showed that, if the circuit of Figure 3 is to be optimized in a mean square error sense for any type of independent noises in the inputs, then a_i is also to be in versely proportional to the variance of the noise in the i-th input. In general, as uncertainty in the input goes up, the vote weight goes down.

A similar tie between decision theory and vote-taking exists for binary digits, say +1 and -1. Assume digits $x_1, x_2, \ldots x_m$ are available, and that they have independent error probabilities $\lambda_1, \lambda_2, \ldots \lambda_m$ respectively. It can be shown[6,10,11] that the more probably correct binary digit is

$$\text{signum}\left[\log \frac{1-p_o(\bar{y})}{p_o(\bar{y})} + \sum_{i=1}^{m} x_i \log \frac{1-\lambda_i}{\lambda_i}\right] \qquad (5)$$

$$\text{where signum } z = \begin{cases} +1 & \text{if } z > o, \\ -1 & \text{if } z < o. \end{cases}$$

$p_o(\bar{y})$ = a priori probability that y = -1 (i.e., the relative frequency that y = -1).

For convenience, let

$$s_i = \log \frac{1-\lambda_i}{\lambda_i} \qquad i = 1, 2, \ldots m$$

be called the sureness information of the i-th source, and let

$$s_o = \log \frac{1-p_o(\bar{y})}{p_o(\bar{y})}$$

be called the a priori sureness information that y = +1. Then the binary digit more likely to be correct* is

$$\text{signum}\left[s_o + \sum_{i=1}^{m} x_i s_i\right] . \qquad (6)$$

This expression suggests a vote taker, the binary version of which is shown in Figure 4. In the circuit of Figure 4, x_1 is multiplied by a_1, x_2 is multiplied by a_2, etc., and these values summed. If the sum is positive, the output is +1; if the sum is negative, the output is -1. Thus the output decision D is

$$D = \text{signum}\left[a_o + \sum_{i=1}^{m} a_i x_i\right]$$

Equations 5 and 6 assert that the more probable digit is obtained for the optimum vote weights

$$a_o = s_o = \log \frac{1-p_o(\bar{y})}{p_o(\bar{y})} \qquad (7)$$

(sureness information y = +1)

$$a_i = s_i = \log \frac{1-\lambda_i}{\lambda_i}, \; i = 1, 2, \ldots m$$

(sureness information of i-th input source).

These sureness informations are generally useful in the manipulation of probabilities based upon independent information.

Another application of these sureness informations is in pattern recognition of signals with independent information. For instance, in dichotimizing between patterns A and $\bar{A}$ from +1 and -1 inputs

$$\log \frac{p(y = A|x)}{p(y = \bar{A}|x)} = \log \frac{p_o(A)}{p_o(\bar{A})} \qquad (8)$$

$$+ \sum_{i=1}^{m} x_i \left[\log \frac{p(x_i = +1|y = A)}{p(x_i = -1|y = A)}\right]$$

assuming $p(x_i = +1/y = A) = p(x_i = -1|y = \bar{A})$.

The associated circuit, of course, is just Figure 4. The connection between decision theory and vote-taking by electrical circuits such as Figures 3 and 4 is believed to be new. However, logarithms of ratios of probabilities have appeared in the literature of statistics in a context similar to that of equation 8 (ref. 12), while the data-processing capabilities of the circuit of Figure 4 have been demonstrated by Mattson[13] and Widrow and Hoff[14].

*The statistical decision can be generalized into a Bayes decision. When the loss for correct outputs is zero, only an additional bias term is required, which is the log of the ratio of relative losses.

Types of Errors and Types of Reliability

One type of error in binary systems is caused by thermal or other noise which occurs randomly in time. Another type of error occurs randomly in space throughout the equipment, persisting from operation to operation; an example is catastrophic failure. Errors, therefore, can occur randomly in time or space. Repetition can control only time errors. Redundancy of equipment can control both time and space errors.

The same mathematics are used in computing system error probabilities from component error probabilities for both time and space errors. If the component error probability is in time, the system error probability will be in time. If the component error probability is in space, the system error probability is in space, i.e., the probability that the system works at all. For example, the maximum likelihood estimate of the analog variable y given by Equation 2 has a Gaussian error distribution according to Equation A3 (in the appendix). The actual error, however, could be constant in time with the given *a priori* Gaussian distribution, or the actual error could be white or colored noise with the given Gaussian distribution. Thus, although the distribution of the error is given by Equation A3, its actual characteristics depend upon the error mechanisms involved.

"Reliability" may mean that newly manufactured items have a high yield, or that a digital system has an accurate output, or that a system has a long lifetime before failure. In this paper, "reliability" is taken to mean all three, because the mathematical analysis is the same for all three, and because redundancy of equipment controls problems of yield, accuracy, and lifetime.

Error Probabilities of Decision Elements

The error in the analog decision element of Figure 3 will have a continuous probability distribution. When input errors are independent, Equation 3 has shown that the variance of the error is

$$\sigma_D^2 = \left(\sum_{i=o}^{m} a_i^2 \sigma_i^2 \right) \Big/ \left(\sum_{i=o}^{m} a_j \right)^2 \quad .$$

If the input errors are Gaussian as well as independent, the error in the output is also Gaussian with the above variance.

The error probability of the binary decision element of Figure 4, given a particular input x, can be obtained directly from the equations used to derive Equation 5, when the vote weights are the optimum vote weights given by Equation 5. The output D is selected so that

$$\log \frac{p(D \text{ is correct})}{p(D \text{ is incorrect})} = \left| \log \frac{p(y = +1|x)}{p(y = -1|x)} \right|$$

and consequently

$$\log \frac{p(D \text{ is correct})}{p(D \text{ is incorrect})} = \left| s_o + \sum_{i=1}^{m} x_i s_i \right| \quad .$$

The *a posteriori* error probability for the binary decision element does depend upon x, while the equivalent uncertainty in the analog decision element is independent of x. The error probability of the binary decision element, not given x, is the reliability parameter required for design. This probability, for general a_i, is the probability that the random variable v is less than zero, where (using y to represent the correct output)

$$v = y \cdot (\text{output of summer})$$

$$= a_o y + \sum_{i=1}^{m} a_i (x_i y)$$

$$\text{and where } (x_i y) = \begin{cases} +1 \text{ with probability } 1-\lambda_i \\ -1 \text{ with probability } \lambda_i \end{cases}$$

An upper bound for $p(v \leq o)$ suited to hand computation is[6,10,11]

$$\prod_{i=1}^{m} 2\sqrt{\lambda_i (1-\lambda_i)} \cosh \frac{a_i - \ln(1-\lambda_i)/\lambda_i}{2} \quad .$$

The bound clearly demonstrates the advantage of the proper vote-weights, for the cosh term has its minimum, for each i, when

$$a_i = \ln(1-\lambda_i)/\lambda_i .$$

If the optimum a_i are used, adding an input with error probability λ_j multiplies the bound by

$$2\sqrt{\lambda_j (1-\lambda_j)} .$$

Thus, for small λ_i and proper a_i, the bound goes roughly as 2^m times the product of the square roots of the λ_i. If the input has error probability of one-half, the bound will be increased unless that input is given a vote-weight of zero. If improper vote-weights are used, then there must be an increase in redundancy over that required with proper vote-weights if the output reliability is to be the same.

Adaption Procedures to Optimize Vote-Weights

The optimum vote weights for both analog and digital vote-takers depend upon the error probabilities of the input signals. Consequently, optimal or near optimal performance using the redundant inputs can be expected from decision elements which adaptively adjust the vote weights to near the optimum values of log $(1-\lambda_i)/(\lambda_i)$ for binary inputs, or $1/\sigma_i^2$ for analog signals. The following methods for adjusting the vote weights for analog inputs are suggested. They parallel similar methods for digital inputs[6,11].

Adaption Method I. The estimated error variance of the i-th input, $\hat{\sigma}_i^2$, is obtained from the circuit which produces x_i. This straighforward open loop adaption requires no analysis.

Adaption Method II-A. $\hat{\sigma}_i^2$ is obtained periodically every T seconds from a comparison of $x_i(t)$ and an externally supplied correct answer $y(t)$ during the last adaption period. The mean value of the error can be subtracted out by appropriate bias in a_o. This method would normally be applied intermittantly during check routines established for the purpose of adjusting the vote weights.

Adaption Method II-B. $\hat{\sigma}_i^2$ is obtained every T seconds from a comparison of $x_i(t)$ and the output of the decision element, either treating the output of the decision element as if it were always correct, or as if the noise not correlated with $x_i(t)$ is zero. The performance of the decision element with random noise in the inputs is a complicated statistical process. An inequality analysis of this process has been obtained for binary decision elements (6, summarized in 10), but not for analog decision elements. In the binary analysis, randomness is effectively eliminated by careful application of the weak law of large numbers, and bounds upon the expected behavior are obtained from a sensitivity analysis of the effect of vote weights upon output reliability. Calculations for a reasonable example show that, at modern computing speeds, the reliability data of a few seconds are accurate enough to prevent a serious misadjustment of vote weights for hundreds of centuries. This analysis justifies feeding back the output in other binary adaption methods which are well suited to implementation but poorly suited to mathematical analysis.

Adaption Method III. This method uses Widrow's surface searching techniques[16], where the output error variance is the parameter to be minimized by a search of the performance surface given by output error variance as a function of the vote weights. The method can be extended to binary decision elements, but the equilibrium vote weights are sub-optimum. The analog adaption procedure requires an external version of the correct signal, but the binary version may replace the external correct signal with the output of the decision element after the threshold to adjust the sum entering the threshold.

Adaption Methods IV-A and IV-B. Let y be the correct analog output, and $\hat{y}$ be the actual output. Method IV-A applies $(x_i-y)^2$ to a linear low pass filter to obtain $\hat{\sigma}_i^2$. Method IV-B applies $(x_i-\hat{y})^2$ to a linear low pass filter to obtain $\hat{\sigma}_i^2$.

Adaption Methods V-A and V-B. The analog version applies $(x_i-y)^2$ [Method V-A] or $(x_i-\hat{y})^2$ [Method V-B] to a linear low pass filter. If the output of the filter exceeds θ, $a_i = o$. Otherwise, $a_i = 1$. Proof 2 of the appendix shows that inputs which are either good [with $\sigma_i^2 \leq A$] or bad [with $\sigma_i^2 \geq B$] can be distinguished in Method V-B whenever σ_D^2 is sufficiently low that $B-A > 2\sqrt{B\sigma_D^2}$. The method improves reliability by turning off bad inputs, and it can actually use its own output to find the bad inputs.

Summary and Conclusions

Here is a brief summary of the results which have been presented:

1. The optimum use of redundant computing circuits is attained by the insertion of decision elements which compute a statistical decision.

2. When errors are independent, a simple vote-taker can make the optimum decision from binary or analog inputs.

3. Adaption is necessary to implement the optimum decision, since the pertinent error probabilities will have to be measured. The adaption may be accurate or crude; many different adaption circuits may be used.

4. Analyses have established that the binary vote-taker can maintain its reliability by feeding back its own output to judge the reliability of its inputs, and that catastrophic failures can be detected in an analog vote taker by feeding back its own output.

5. Previous work dating back to von Neumann has established a pattern of using decision elements so as to tolerate errors in logic circuits, vote-takers, and interconnections.

The use of even simple adaption can greatly increase the reliability of a system. Consider a binary system with 100 stages, each of which has survival probability e^{-t}. Assume that at failure the stages become stuck (randomly) in one of the binary digits. Errorless decision elements are used after each stage, some with majority-rule decision functions and some which have an adaptive circuit which requires only one good input (such as Method II-A). The survival probabilities of nonredundant and redundant systems are plotted on logarithmic scales in Figure 5. Note that the use of adaption greatly extends the median lifetime (intersection with dotted line), and is roughly equivalent to having more redundancy.

My conclusion is that vote-takers are excellent devices for use in redundant systems. They are efficient partly because each vote-taker receives all the redundant information, and partly because their performance may be optimized by adaption.

Appendix

1. Derivation of formula for $p(y|x)$ for x an input vector

$$x = (x_1, x_2, \dots x_m),$$

where y = correct output
$x_i = y + n_j$

and n_i and n_j are independent Gaussian noises, $i \neq j$, with variances σ_i^2, σ_j^2. For convenience let x denote the vector whose i-th component is x_i. Using the independence of n_i and n_j, $j \neq i$,

$$p(x=x'|y=y') = \prod_{i=1}^{m} (2\pi\sigma_i^2)^{-1/2} \exp \frac{(x_i'-y')^2}{2\sigma_i^2} . \quad \text{(A1)}$$

Now

$$\begin{aligned} p(x=x',y=y') &= p(x=x'|y=y')p(y=y') \\ &= p(y=y'|x=x')p(x=x') \end{aligned}$$

so

$$p(y=y'|x=x') = \frac{p(x=x'|y=y')p(y=y')}{p(x=x')} \quad \text{(A2)}$$

(Bayes' law)

Assume the *a priori* distribution on y, namely $p(y=y')$, is $N(x_o, \sigma_o^2)$. Then, by using (A2) in (A3) and completing the square of the y' polynomial in the exponent, (A4) is readily obtained:

$$p(y=y'|x=x') = C \exp \frac{\left[y' - \left(\sum_{i=o}^{m} \frac{x_i'}{\sigma_i^2}\right) \Big/ \left(\sum_{i=o}^{m} \frac{1}{\sigma_i^2}\right)\right]^2}{\left[2\left(\sum_{i=o}^{m} \frac{1}{\sigma_i^2}\right)\right]} . \quad \text{(A3)}$$

Since C is independent of y', the above density is Gaussian with mean and variance

$$E(y) = \left(\sum_{i=o}^{m} x_i/\sigma_i^2\right) \Big/ \left(\sum_{i=o}^{m} 1/\sigma_i^2\right)$$

$$\sigma^2(y) = 1 \Big/ \left(\sum_{i=o}^{m} 1/\sigma_i^2\right) \quad \text{(A4)}$$

2. A sufficient condition for distinguishing good and bad analog inputs in Adaption Method V-B.

$$\hat{\sigma}_i^2 = E\left[x_i - \hat{y}\right]^2 + k_i$$

$$= E\left[y + n_i - \frac{\sum a_j(y+n_j)}{\sum a_k}\right]^2 + k_i , \quad \text{(A5)}$$

where k_i is zero mean noise.

By the independence of n_i and n_j, $i \neq j$, (A5) reduces to, using (6),

$$\hat{\sigma}_i^2 = \sigma_i^2 - 2\frac{a_i\sigma_i^2}{\sum a_k} + \sigma_D^2 + k_i . \quad \text{(A6)}$$

Now by Equation 3

$$\sigma_D^2 = \frac{a_i^2\sigma_i^2 + \sum_{j \neq i} a_j^2\sigma_j^2}{\left(\sum a_k\right)^2} \geq \frac{a_i^2\sigma_i^2}{\left(\sum a_k\right)^2} \quad \text{(A7)}$$

since the a_i, being variance estimates are non-negative. Taking square roots of Equation A7 therefore establishes that

$$0 \leq \frac{a_i}{\sum a_k} \leq \sqrt{\frac{\sigma_D^2}{\sigma_i^2}} \quad \text{(A8)}$$

Using Equation A8 in Equation A6 gives

$$-|k_i| + \sigma_i^2 - 2\sqrt{\sigma_i^2 \sigma_D^2} + \sigma_D^2 \leq \hat{\sigma}_i^2 \leq \sigma_i^2 + \sigma_D^2 + |k_i| \quad (A9)$$

In the limit as the averaging time approaches infinity, $|k_i| < \epsilon$ with probability 1 for any $\epsilon > o$.

Now, for any $\epsilon > o$ and $\delta > o$, the averaging time may be chosen sufficiently large so that $|k_i| < \epsilon$ with probability at least $1-\delta$. Therefore, with arbitrarily high probability, Equation A9 becomes

$$-\epsilon + \sigma_i^2 - 2\sqrt{a_i^2 \sigma_D^2} + \sigma_D^2 \leq \hat{\sigma}_i^2 \leq \sigma_i^2 + \sigma_D^2 + \epsilon \quad (A10)$$

Consequently, the criterion that $\hat{\sigma}^2_{\text{good inputs}} < \hat{\sigma}^2_{\text{bad inputs}}$ becomes

$$A + \sigma_D^2 + \epsilon < B + \sigma_D^2 - \epsilon - 2\sqrt{B\sigma_D^2}$$

or

$$B - A > 2B\sqrt{\sigma_D^2}$$

since ϵ is arbitrarily small.

Bibliography

1. J. von Neumann, "Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components", Automata Studies, Ed. C. E. Shannon and J. McCarthy, Princeton University Press, Princeton, N. J., 1956.

2. E. F. Moore and C. E. Shannon, "Reliable Circuits Using Less Reliable Relays", Journal of the Franklin Institute, 262, September and October 1956, pp. 198-208, 281-297.

3. R. W. Hamming, "Error Detecting and Error Correcting Codes", Bell System Technical Journal, Vol. 29, p. 147, 1950.

4. Bell System Technical Journal, March 1961, D. B. Armstrong, "A General Method for Applying Error Correction to Synchronous Digital Systems".

5. L. A. M. Verbeek, "Reliable Computation with Unreliable Circuitry", MIT Tech. Research Lab of Elect. (1960).

6. W. H. Pierce, "Improving Reliability of Digital Systems by Redundancy and Adaption", Stanford Electronics Laboratories Technical Report 1552-3, July 17, 1961. (Also available as a Ph.D. thesis from University Microfilms, Ann Arbor, Mich.)

7. R. Miller, Majority Logic Analysis, Hermes Electronics Co., Cambridge, Mass., Publication M 895, August 15, 1960.

8. R. Wasserman, W. G. Brown, and J. Tierney, "Improvement of Electronic Computer Reliability through the Use of Redundancy", Proceedings of the National Electronics Conference, Vol. XVII, Chicago, Ill., October 1961, pp. 341-359.

9. W. C. Mann, "Systematically Introduced Redundancy in Logical Systems", 1961 IRE International Convention Record, 9, Pt. 2, March 1961, pp. 241-263.

10. W. H. Pierce, "A Proposed System of Redundancy to Improve the Reliability of Digital Computer", 29, Stanford Laboratories Technical Report, TR 1552-1, July 29, 1960.

11. W. H. Pierce, "Adaptive Vote-Takers Improve the Use of Redundancy", Proceedings of the Symposium on Redundancy Techniques for Computing Systems, Feb. 6-7, Washington, D.C.

12. I. J. Good, Probability and the Weighting of Evidence, Charles Griffin and Co., Ltd., London, 1950, p. 71.

13. R. L. Mattson, "A Self-Organizing Logical System", 1959 Eastern Joint Computer Conference Convention Record, Institute of Radio Engineers, New York, N. Y.

14. Bernard Widrow and Marian Hoff, "Adaptive Switching Circuits," 1960 IRE Convention Record.

15. Bernard Widrow, "An Adaptive 'Adaline' Neuron Using Chemical Memistors," Stanford Electronics Laboratories, Technical Report 1553-2, October 17, 1960.

16. Bernard Widrow, "Adaptive Sampled-Data Systems - a Statistical Theory of Adaption," 1959 IRE Wescon Record, .Part 4.

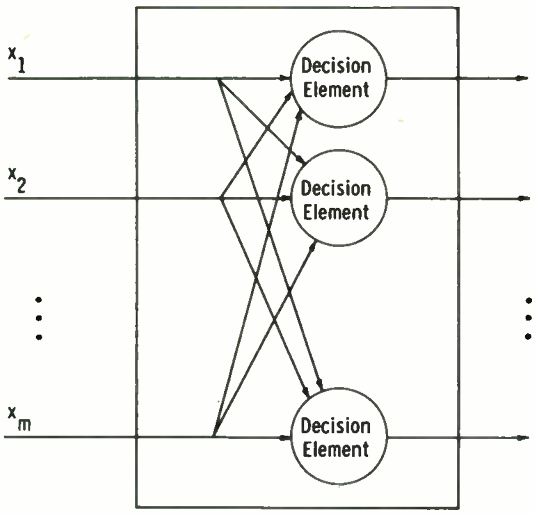
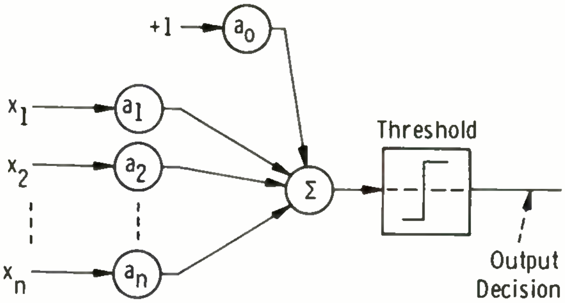


Fig. 1—An m-input, m-output restoring organ composed of m decision elements



A decision element or vote-taker for binary inputs of +1 or -1

Fig. 4.

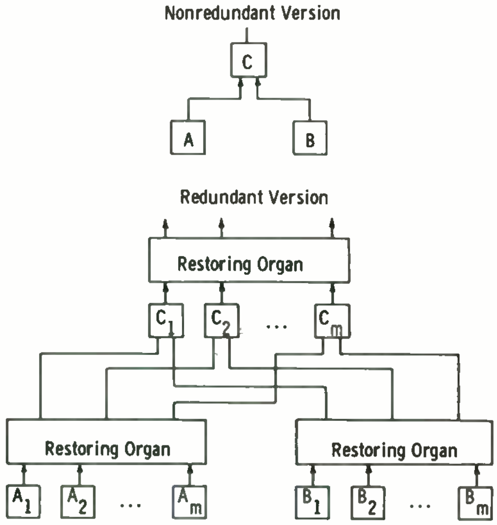
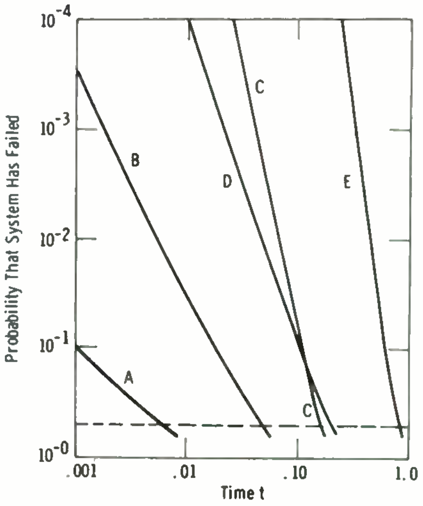


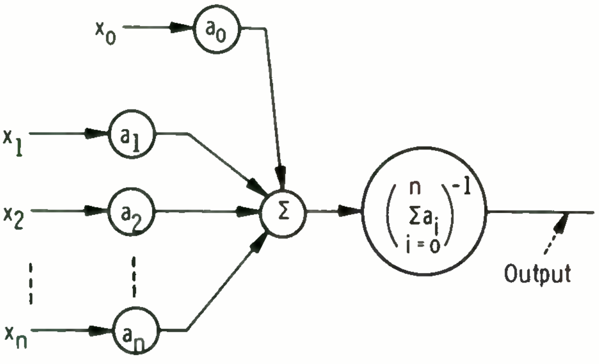
Fig. 2— Placement of restoring organs in a tree logic network



Reliability as a function of time for a system with 100 stages, each with survival probability e^{-t}.

Curve A. n = 1 (no redundancy)
Curve B. n = 3, majority-rule decision elements
Curve C. n = 9, majority-rule decision elements
Curve D. n = 3, perfectly adaptive decision elements
Curve E. n = 9, perfectly adaptive decision elements

Fig. 5.



An analog vote-taker which takes a linear combination of its inputs to form its output

Fig. 3.

CONFLEX I
—A CONDITIONED REFLEX SYSTEM—

Malcolm R. Uffelman

SCOPE
Incorporated
Falls Church, Virginia

Summary

The functional organization of a conditioned-reflex system (CR System) is presented. This general system is then analyzed by means of probability theory and design equations are derived. Finally the logical design and operational capabilities of a special purpose digital system, called CONFLEX I, which implements the concept is described.

Introduction

The general problem considered is the design of a system for the classification of minimally constrained stimuli. Given a system with an input retina containing N_S binary input cells, it is possible to describe 2^{N_S} black-white patterns. Thus, the minimum number of constraints are:

1. The stimuli are two-dimensional figures.

2. The stimuli are black-white figures.

Now, randomly select N_C groups of N_e patterns each from the ensemble of 2^{N_S} possible patterns. We define that these groups constitute N_C classes of N_e stimuli each.

For the purpose of the analysis presented, the following standard experiment is defined. The system is conditioned by having each of the $N_C N_e$ stimuli presented once. As each stimulus is presented, the desired output (one of the N_C possible outputs) is reinforced. After this procedure is completed, the system is tested by having each stimulus presented and the system's output noted.

The System

Components

The CR system is made up of five basic components. They are:

1. Sensory Cells (S-Cells). The S-cells are transducers which couple the system to the stimulus environment. An S-cell's output (s_j) is proportional to the intensity of its input. Its output range is $0 \le s_j \le +1$.

2. Discrimination Cells (D-Cells). The D-cells are majority decision gates. Each D-cell has A affirmative inputs and N negative inputs.

The output of a D-cell, d_j, is defined by:

$$d_j = +1 \text{ if } \sum_{i=1}^{A} a_{ij} - \sum_{k=1}^{N} n_{kj} > 0, \qquad (1)$$

$$d_j = 0 \text{ if } \sum_{i=1}^{A} a_{ij} - \sum_{k=1}^{N} n_{kj} = 0, \text{ and} \qquad (2)$$

$$d_j = -1 \text{ if } \sum_{i=1}^{A} a_{ij} - \sum_{k=1}^{N} n_{kj} < 0 \qquad (3)$$

Where: a_{ij} is $i^{\underline{th}}$ affirmative input of the $j^{\underline{th}}$ D-cell

n_{kj} is the $k^{\underline{th}}$ negative input of the $j^{\underline{th}}$ D-cell

3. Memory Cells (M-Cells). The M-cells are variable gain, single input-single output units. An M-cell's gain can be varied in positive or negative unit steps. The output of an M-cell is the product of its input and its gain at the time the input is applied.

4. Summing Units. Each summing unit accepts N_D inputs and produces an output which is equal to the algebraic sum of its inputs.

5. The Comparator. The Comparator accepts N_c inputs and produces a unit signal on one of N_c binary outputs. The output selected indicates which of the N_c inputs was algebraically largest.

System Organization

Fig. 1 shows the functional organization of a CR system. The S-field contains N_s S-cells, the D-field contains N_D D-cells, and each M-field contains N_D M-cells. The connections between the S-cells and the D-cells are shown to be randomly selected. A moment's thought will show why this is so. Each D-cell requires N_i inputs (N_i = A+N) and these must be selected in some fashion from the N_s S-cells. Thus, there are

$$\binom{N_s}{N_i} = \frac{N_s!}{N_i!(N_s-N_i)!} \qquad (4)$$

ways in which the D-cells input could be selected. Or, in other words, there are possible $\binom{N_s}{N_i}$ different D-cell functions. N_D D-cells are required. As will be shown, N_D is generally less than $\binom{N_s}{N_i}$. Because the stimuli are, by definition, randomly assigned to different classes, we have no reason to prefer one group of N_i inputs over any other. Thus, we can do no better than to randomly select N_D groups of N_i inputs from the ensemble of $\binom{N_s}{N_i}$ possible groups. Note that for a given model, once the connections are selected they remain fixed for that model.

The output of each D-cell is connected to one M-cell in each M-field. There are N_c M-fields, one for each stimulus class. The outputs of the M-cells in a given M-field are connected to a summing unit. When a stimulus is presented for testing purposes, each M-cell produces an output which is the product of its D-cell input and the value of its gain at that time. Thus, each summing unit produces an output:

$$\sigma_i = \sum_{j=1}^{N_D} m_{ji} d_j \qquad (5)$$

where: σ_i is the sum for the ith M-field.

m_{ji} is the value of the gain of the jth M-cell in the ith M-field.

d_j is the value of the output ($\pm$1,0,-1) of the jth D-cell.

Equation (5) is simply a discrete form of the well known correlation function.

The comparator then indicates which M-field produced the largest output due to the test stimulus (i.e. which M-field produced the highest correlation). This identifies the most probable class identity of the test stimuli.

The M-cell values are determined during the conditioning phase. As each of the N_e stimuli in a given class is presented, each D-cell produces an output. These outputs are transferred to the M-field which has been selected by the operator to represent the given class. The gain of each M-cell is changed by the amount of the associated D-cell's output. That is, if D-cell #4 produces a +1 output, then the gain of M-cell #4 in the selected M-field will be increased by +1. After conditioning, the value of an M-cell is:

$$m_{ji} = \sum_{k=1}^{N_e} d_k(ij) \qquad (6)$$

where: $d_k(ij)$ is summed over the stimuli in the jth class only.

m_{ji} is the value of the ith M-cell in the jth M-field.

Analysis

General Analog Case

Each stimulus may be considered a vector, $\bar{S}_a$, with N_s components, each of which is either 0 or +1. The response of the D-cell is uniquely determined for each $\bar{S}_a$. This set of values may also be considered a vector, $\bar{D}_j$. Where $\bar{D}_j$ has N_D components, each of which may be +1, 0, or -1. Thus, each stimulus vector uniquely defines a D vector. The values of M-cells in a given M-field may also be represented by a vector having N_D components. However, for an M vector each component may have a value $-N_e \le m_{ji} \le +N_e$. If we consider that the operation required by equation (5) is also the definition of the Dot Product of two vectors in a linear vector space covered by a rectangular coordinate system of N_D dimensions, it is obvious that

$$\sigma_j = \bar{D}_a \cdot \bar{M}_j \qquad (7)$$

where: $\overline{D}_a$ is the D vector caused by stimulus $\overline{S}_a$.

$$\overline{M}_j = \sum_{k=1}^{N_e} \overline{D}_k(j) \qquad (8)$$

where: $\overline{D}_k$ is summed over the stimuli for the jth class only.

Now consider a CR system having only two M-fields, and therefore able to divide the stimulus environment in two classes. For a randomly selected test stimulus there are two possibilities. The D vector produced, $\overline{D}_u$, belongs to either $\overline{M}_1$ or $\overline{M}_2$. That is, $\overline{D}_u$ is a component vector in the sum $\overline{M}_1$ or in the sum $\overline{M}_2$. Let us define that $\overline{D}_u$ belongs to $\overline{M}_1$. Thus we may write

$$\sigma_1 = \overline{D}_u \cdot \overline{M}_1 = \overline{D}_u \cdot \overline{D}_u + \overline{D}_u \cdot (\overline{M}_1 - \overline{D}_u) \qquad (9)$$

but $$\overline{D}_u \cdot \overline{D}_u = \sum_{j=1}^{N_D} d(u)_j d(u)_j \qquad (10)$$

and $$\overline{D}_u \cdot (\overline{M}_1 - \overline{D}_u) = \sum_{j=1}^{N_D} d(u)_j (m(1)_j - d(u)_j) \qquad (11)$$

where $d(u)_j$ is the value of the jth component of $\overline{D}_u$

$m(1)_j$ is the value of the jth M-cell in $\overline{M}_1$

Obviously equation (11) may also be represented by

$$\overline{D}_u \cdot (\overline{M}_1 - \overline{D}_u) = \sum_{k=1}^{N_e - 1} \sum_{j=1}^{N_D} d(u)_j \; d(1)_{jk} \qquad (12)$$

where: $d(1)_{jk}$ is the value of the jth D-cell output for the kth stimulus in class 1.

Note that the summation is over the range 1 to N_e-1 because $d(u)_j$ is not included in the sum. The mean value of $\overline{D}_u \cdot \overline{M}_1$ may be expressed:[1]

$$E(\sigma_1) = E(\overline{D}_u \cdot \overline{D}_u) + E(\overline{D}_u \cdot (\overline{M}_1 - \overline{D}_u)) \qquad (13)$$

Consider $E(\overline{D}_u \cdot \overline{D}_u)$. From above

$$E(\overline{D}_u \cdot \overline{D}_u) = E \sum_{j=1}^{N_D} d(u)_j d(u)_j \qquad (14)$$

$$= \sum_{j=1}^{N_D} E(d(u)_j d(u)_j) \qquad (15)$$

The value of $E(d(u)_j d(u)_j)$ may be evaluated by

$$E(d(u)_j d(u)_j) = +1(P(d(u)_j = +1))$$

$$- 1(P(d(u)_j = -1)) + 0(P(d(u)_j = 0)) \qquad (16)$$

By using the D-cell's logical rules, it can be shown that

$$P(d_j = +1) = \qquad (17)$$

$$\sum_{a=n+1}^{A} \sum_{n=0}^{A-1} \binom{A}{a} \binom{N}{n} (\Psi)^{a+n} (1-\Psi)^{A+N-a-n}$$

$$P(d_j = -1) = \qquad (18)$$

$$\sum_{n=a+1}^{N} \sum_{a=0}^{N-1} \binom{A}{a} \binom{N}{n} (\Psi)^{a+n} (1-\Psi)^{A+N-a-n}$$

$$P(d_j = 0) = \qquad (19)$$

$$\sum_{a=n}^{*} \sum_{n=0}^{*} \binom{A}{a} \binom{N}{n} (\Psi)^{a+n} (1-\Psi)^{A+N-a-n}$$

*A or N, whichever is smallest.

where: $\Psi = \gamma/N_s$, the relative stimulus size

γ = number of S-cells activated by the stimulus

If we set A equal to N it can be seen that:

$$P(d_j=+1) = P(d_j=-1) \qquad (20)$$

$$P(d_j=0) = \sum_{b=0}^{B} \binom{B}{b}^2 (\Psi)^{2b}(1-\Psi)^{2B-2b} \qquad (21)$$

where: B is equal to A(A=N).

Because $P(d_j=+1) = P(d_j=-1)$, we use p to represent the value of either one. Also, $P(d_j=0)$ will be represented by p_0:

$$p_0 = 1 - 2p \qquad (22)$$

Fig. 2 shows a plot of p_0 as a function of Ψ. Returning to equation (16), we see that it becomes:

$$E(d(u)_j d(u)_j) = 2p \qquad (23)$$

$$\text{and} \quad E(\overline{D}_u \cdot \overline{D}_u) = 2pN_D \qquad (24)$$

From equations (12), (20), and (21) we find that

$$E(\overline{D}_u \cdot (\overline{M}_1 - \overline{D}_u)) = 0 \qquad (25)$$

Thus,

$$E(\sigma_1) = 2pN_D \qquad (26)$$

By employing the well-known variance relation, Var $x = E(x)^2 - E^2(x)$, and the above results, we find that

$$\text{Var}(\sigma_1) = N_D(2p-4p^2) + 4p^2N_D(N_e-1) \qquad (27)$$

Now consider the case $\overline{D}_u \cdot \overline{M}_2$ where $\overline{D}_u$ is not a component of $\overline{M}_2$. By application of the same methods used above, we find

$$E(\sigma_2) = 0 \qquad (28)$$

and

$$\text{Var}(\sigma_2) = 4p^2N_DN_E \qquad (29)$$

The situation is now analogous to monitoring two output channels. Let both channels respond when an input is presented. We define that one produces a "noise" or random valued output, Var(σ_2); the other produces a signal E(σ_1) plus a noise, Var(σ_1). The problem is to decide which output channel carries the signal. The accuracy with which the proper channel can be selected is governed by the ratio of the expected signal value and the deviation of the noise. Thus, we define Γ, the signal-to-noise ratio:

$$\Gamma = \sqrt{\frac{E(\sigma_1)}{\text{Var}(\sigma_1)+\text{Var}(\sigma_2)}} \qquad (30)$$

Substituting equations (26), (27), and (29) we find

$$\Gamma = \sqrt{\frac{N_D}{2N_e + \frac{1-4p}{2p}}} \qquad (31)$$

Consider the term $\frac{1-4p}{2p} = \frac{2p_0-1}{1-p_0}$ (32)

Fig. 3 is a plot of $\frac{2p_0-1}{1-p_0}$. Note that if p_0 is less than or equal to 0.5, the value of $\frac{2p_0-1}{1-p_0}$ is zero or less. From Fig. 2 we see that p_0 is less than 0.5 for Ψ between 0.1 and 0.9 if B equals 4 or more. Therefore, in general, it is reasonable to assume

$$2N_e \gg \frac{1-4p}{2p} \qquad (33)$$

Thus, equation (31) becomes

$$\Gamma = \sqrt{\frac{N_D}{2N_e}} \qquad (34)$$

If we assume that N_e is sufficiently large, the probability of correctly selecting the proper class for a test stimuli, P_c, is given by the normalized gaussian form:

$$P_c = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\Gamma} \exp - \frac{x^2}{2}\, dx \qquad (35)$$

In general, N_c is greater than two, in which case it can be shown that

$$P_c \simeq \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\Gamma} \exp - \frac{x^2}{x}\, dx \right]^{N_c - 1} \qquad (36)$$

Clipped Memory Case

Assume that the number of inputs to the D-cells, N_i, is extremely large and that as a result p_0 is zero for practical purposes. If N_e is large we may assume that the distribution of values in the M-cells is gaussian. With these assumptions let us investigate the consequences of clipping the memory values according to the rules:

If $m_j \geq 0$ replace it with +1

If $m_j < 0$ replace it with -1

Consider first the case of $\overline{D}_u \cdot \overline{M}_2$, $\overline{D}_u$ not a component of M_2. The probability that $d_j m_j = +1$ is 0.5 and the probability $d_j m_j = -1$ is 0.5.

Clearly:

$$E(\sigma_2) = 0 \qquad (37)$$

and

$$\mathrm{Var}(\sigma_2) = N_D \qquad (38)$$

In the case of $\overline{D}_u \cdot \overline{M}_1$ we must consider the conditional probabilities involved. After the stimulus is presented we know that either $d_j = +1$ or that $d_j = -1$. Due to the symmetry, one of these results need be considered. We chose the result $d_j = +1$. Now we must evaluate the probabilities that $m_j = +1$ <u>given</u> $d_j = +1$ and that $m_j = -1$ <u>given</u> that $d_j = +1$. The knowledge that $d_j = +1$ causes a probability bias of amount $1/\sqrt{2\pi N_e}$; $\sqrt{N_e}$ being the deviation of the M-cell value. Thus:

$$P(m_j = +1 | d_j = +1) = 1/2 + 1/\sqrt{2\pi N_e} \qquad (39)$$

$$P(m_j = -1 | d_j = +1) = 1/2 - 1/\sqrt{2\pi N_e} \qquad (40)$$

By the method used above:

$$E(\sigma_1) = N_D \sqrt{\frac{2}{\pi N_e}} \qquad (41)$$

and

$$\mathrm{Var}(\sigma_1) = N_D - \frac{2N_D}{\pi N_e} \qquad (42)$$

which leads to a value:

$$\Gamma = \sqrt{\frac{N_D}{\pi N_e}} \text{ (assuming } N_e >> 1) \qquad (43)$$

Compare this result with equation (34). Clipping the memory requires an increase in D-cells of only $\pi/2$ times the number of D-cells needed in the general analog case. However, it greatly reduces the storage capacity required of the memory.

Summary of Other Analytical Work

Space does not permit a full development of the mathematical theory of CONFLEX I. However, a few salient points should be noted.

The above analysis demonstrates the rote learning ability of CONFLEX I. We have also demonstrated analytically that the system can generalize. This generalization is based purely on stimulus overlap on the S-field. However, if a suitable conditioning procedure is used the system can correctly classify <u>new</u> stimuli with extremely high accuracy.

In addition, because of the symmetric nature of the D-cell's logical function, there is no tendency on the part of the system to become biased in favor of one output due to frequency of stimulus occurrence. In the above analysis, we assume N_e stimuli in each class. However, it is easily shown that the factor controlling P_c is N_cN_e. That is, for the two-class system analysed, we could place $2N_e$-X stimuli in one class and X stimuli in the other class without affecting P_c.

If the number of inputs to the D-cells is reasonably large (in the order of 10 or more) and the stimuli are between size $\psi = 0.1$ and $\Psi = 0.9$ there is negligible size bias. That is, within the specified range the affect of stimulus size on P_c is quite small.

Finally, it has been shown that the constraint (2) may be removed. That is, the stimuli need not be restricted to black-white patterns but may be shaded, black-grey-white, patterns.

CONFLEX I

CONFLEX I was designed to serve as an experimental CR system. For this reason, it was decided that the design should provide for several thousand D-cells and multiple M-fields (classes). If the parallel design suggested by the functional organization of Fig. 1 were followed, the cost of the system would be very large indeed. Therefore, the decision was made to base the design on digital techniques, utilizing serial operations to a maximum. Fig. 4 shows a simplified block diagram of CONFLEX I. Notice that the system contains only one D-cell circuit. To generate the large number of D-cells required, the inputs to this single circuit are sequentially switched by the memory clock. This means that the D-field "exists" in time, not in space.

The Sensory Field and D-Cell

The S-field is composed of 400 photo-resistive elements arranged in a 20 x 20 matrix. One-half of the photo-resistors are wired to give affirmative outputs and the other half negative outputs. Each photo-resistor is controlled by a sampling gate. Each gate has two inputs--a row and a column input as shown in Fig. 5. Two maximal-length sequence generators (M-S-G's) are used to produce the sampling singals.[2] Fig. 5 shows a simplified diagram of this scheme applied to a 4 x 4 matrix. The shaded cells are being sampled by the M-S-G state's shown. As shift pulses from the memory clock are applied, the M-S-G's step through their various states. Each state causes a different group of photo-resistors to be sampled.

All of the outputs from the photo-resistors are connected to a transistor-summing amplifier which serves as the D-cell input. Only the sampled photo-resistor's outputs contribute to the amplifier's output. Thus, for a given pattern projected on the S-field, the amplifier's output is a function of the photo-resistors being sampled. The amplifier's output is sent to a threshold circuit which performs the logical functions used to define the D-cell. This circuit has two output lines, one which becomes active for a +1 output and the other for a -1 output.

The Memory

A rotating magnetic disc memory is employed in CONFLEX I. The memory, a BD-500 produced by the Laboratory for Electronics, Inc., has a storage capacity of approximately 500,000 bits. In CONFLEX I the storage is arranged into eight tracks of temporary memory, one clock track, two programming tracks, 48 general storage tracks, and 10 spare tracks. Each track contains 5100 bit locations. The disc speed is 3600 rpm.

The Arithmetic Unit

The A.U. contains two matrices, a flip-flop, and a bi-directional shift register. First, we shall consider the two matrices. For problems of the size we intend to study, the likelihood of the value stored in any M-cell exceeding ±127 is extremely close to zero. Thus, at no time will the A.U. ever be required to perform addition or subtraction on an operand of magnitude greater than 127. With this range in mind, a two-digit base 16 number was chosen for the internal code. However, because of the upper magnitude limit of 127, one-half of the 16^1 digit is required. Thus, for internal use, a 3-bit word and a 4-bit word are sufficient. Notice that this choice of radix, allows maximum utilization of the seven bits:

$$127 = (7\times16^1) + (15\times16^0) = 111 \quad 1111 \tag{44}$$

When computing the value of a Dot Product, the largest value of each new entry is 127; however, the partially accumulated sum can be greater. This will result in an overflow from the two-digit adder. We have shown that the probability of the Dot Product exceeding 8191 in magnitude is extremely small. The A.U. has a flip-flop and a five-stage, bi-directional shift register to handle this overflow from the two-digit adder. The technique used is based on residue algebra and was developed by SCOPE Incorporated under another contract.[3]

The Comparator

The comparator can store two 13-bit words plus two sign-bits, and the source address (i.e., M-field number) of each. The two words are compared and the largest word and its source address are retained.

System Operation

Let us first consider the conditioning phase. Initially, the memory is cleared. When a stimulus is presented, the operator signals the system by means of a push button and the system clock is started. The system clock causes the M-sequence generators, starting in a preset state, to shift through their sequences. As each M-S-G assumes its next state the D-cell produces an output. In synchronism with the D-cell output, a 7-bit word plus a sign bit is read from temporary memory and sent to the A.U. This word is the value of the gain for the M-cell assigned to that temporary memory location. The M-cell value and the output of the D-cell are added together in the A.U. and the new M-cell value is returned to the assigned temporary memory location. Thus, it requires one revolution of the memory disc to update all the M-cells for the class being conditioned. At the completion of the required revolution, the M-S-G's are reset to the proper preset state and the system clock is stopped. When another stimulus is presented, the operator again signals the system and the above process is repeated.

In order to minimize the cost of CONFLEX I, only one set of 7-bit plus sign registers (called temporary memory) can communicate directly with the A.U. Thus it is required that all of the stimuli for a given class be presented in sequence. After all the stimuli for a class have been presented, the contents of the N_D temporary memory words are transferred to final memory. The N_D words may be transferred in one of three modes: (1) The full word may be transferred, in which case the system is in the analog mode, (2) The words may be clipped to 2-bits (sign and magnitude) by the same rules which define a D-cell. In this case the system is in the partially clipped mode. (3) The words may be clipped to one bit (+1 or -1) as described above. Here the system is in the clipped mode. After a class has been presented and transferred to final memory, the temporary memory is cleared and the system is ready for a new class.

Now consider the test phase. A test stimulus is presented and the system is notified by the operator and the system clock starts. This causes the M-S-G's to step through their sequences. As each D-cell output is produced, the appropriate M-cell value from M-field #1 is read from final memory. These two values are multiplied together and sent to the A.U. for accumulation. Because d_j is either +1, -1, or 0, the multiplication is only an inhibit and sign modification operation. After one revolution of the memory disc, the Dot Product, σ_1, for the first M-field is formed. This value is sent to the comparator and stored, and the M-S-G's are reset. The entire operation is then automatically repeated for M-field #2 and the new Dot Product, σ_2, is transferred to the comparator. The comparator selects the largest and retains its value and M-field number. The process is again repeated for M-field #3 and the comparator again retains the largest. Thus, during the test phase it requires one revolution of the memory disc for each class or M-field to be tested. After all of the M-fields are checked, the system signals the operator, via the indicators, which M-field produced the largest Dot Product.

System Capabilities

CONFLEX I can be programmed by a simple switch setting to have 500, 1000, 2000, 3000, or 5000 D-cells. The system has six M-fields in the analog mode, 24 M-fields in the partially clipped mode, and 48 M-fields in the clipped mode. It requires 16 msec to process one stimulus in the conditioning phase and 16 msec per M-field to process a stimulus in the test phase. When used in the clipped mode, CONFLEX I has a P_c of about 0.99 for 120 stimuli assigned to each of 48 classes. In other words, CONFLEX I can place some 5,760 stimuli into 48 classes with an expected accuracy of 99%.

Conclusions

The theory developed on this project has shown that a relatively powerful, adaptive pattern recognition system can be built from basic logical functions arranged in a simple manner. The use of disjunct M-fields for each class permits a simple method for expanding the class size of a basic system. The CR system is insensitive to a wide range of stimulus size. Because of the majority logic D-cells the CR system has no tendency to become biased in favor of one M-field due to unequal values of N_e.

With respect to mechanization of the theory, CONFLEX I offers a design using standard components and conventional digital techniques.

List of Most Used Symbols

d_j	Output of $j^{\underline{th}}$ D-cell
m_{ji}	Value of the $j^{\underline{th}}$ M-cell in the $i^{\underline{th}}$ M-field
N_c	Number of classes
N_D	Number of D-cells
N_e	Number of stimuli per class
p	Probability d_j = +1 (-1)
p_o	Probability d_j = 0
P_c	Probability of correct response
Γ	Signal-to-noise ratio
σ_j	Value of the $j^{\underline{th}}$ Dot Product
ψ	Relative stimulus size

Acknowledgements

The author takes pleasure in acknowledging the help and direction given by E. C. Driese, J. S. Gerig, and W. L. Holford in the development of the mathematical theory presented. It is also a pleasure to acknowledge the work of R. E. Young and T. G. Toth in the logical design of CONFLEX I.

This work was performed under an Air Force contract and is a part of the advanced research effort of the Electromagnetic Warfare and Communication Laboratory of the Aeronautical Systems Division. The author wishes to express his appreciation to Capt. O. E. McFolin, Lt. H. C. Falk, and Lt. R. L. North, all of ASD, for their support and direction in this work.

References

1. E. Parzen, "Modern Probability Theory and Its Applications;" John Wiley & Sons, Inc., New York, 1960.

2. B. Elspas, "The Theory of Antonomus Linear Sequential Networks," IRE TRANSACTIONS ON CIRCUIT THEORY, CT-6, pp 45-60, 1959.

3. E. Driese, G. Glen, and R. Young, Jr., "Computer Applications of Residue Class Notations;" ASD Technical Report 61-189, Sept. 1961.

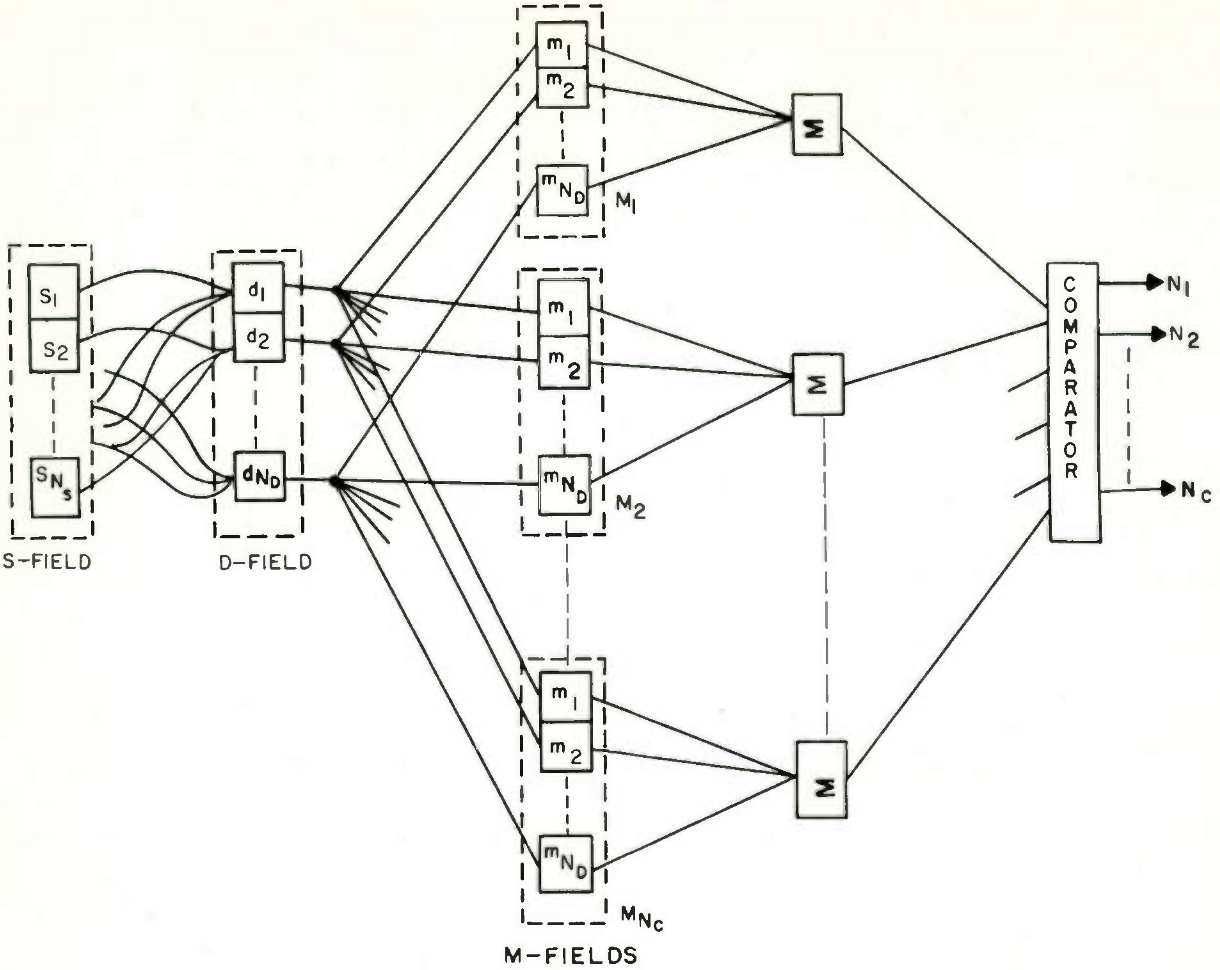
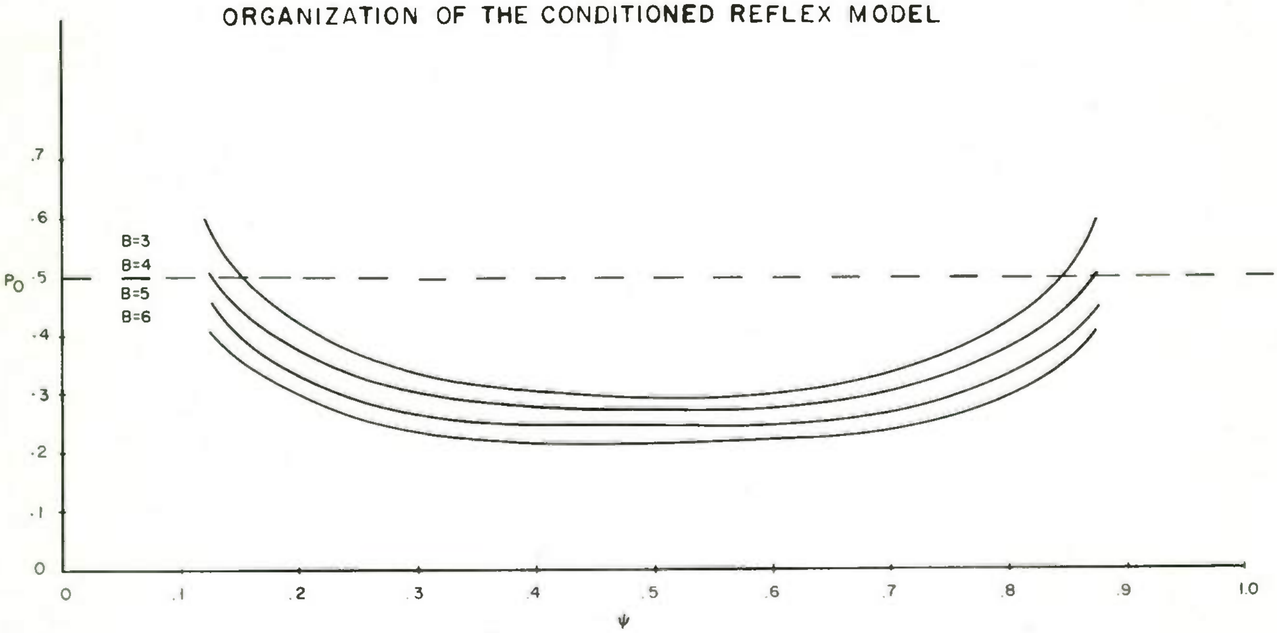


Fig. 1.

ORGANIZATION OF THE CONDITIONED REFLEX MODEL



P_0 AS A FUNCTION OF ψ

Fig. 2.

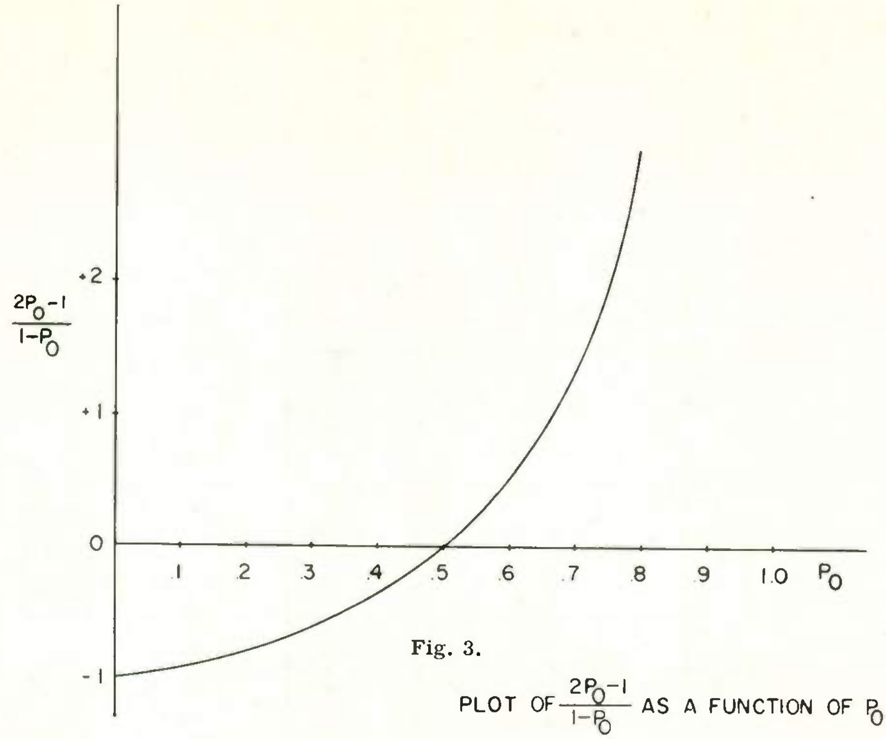
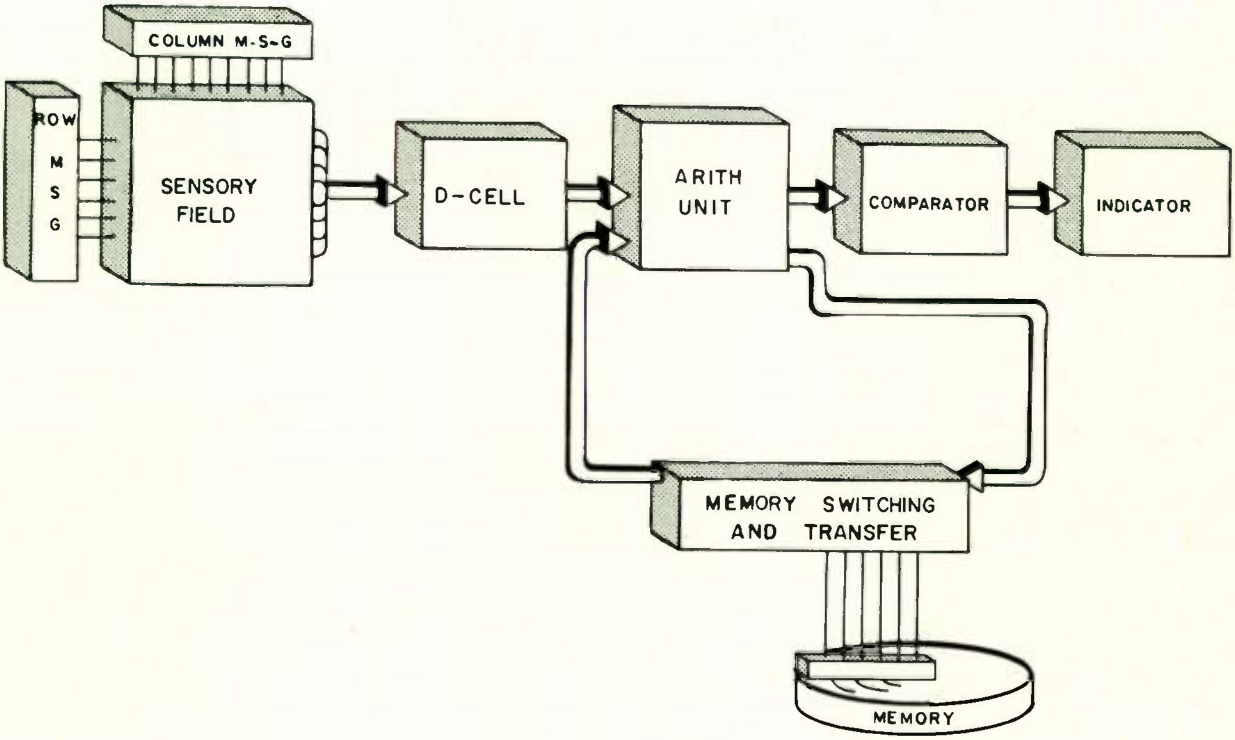


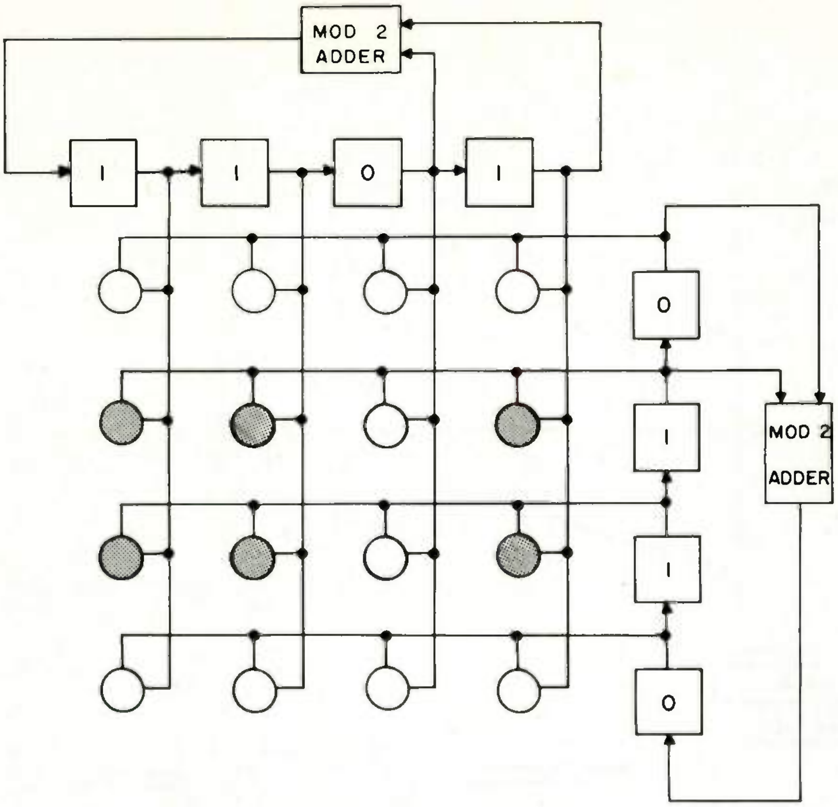
Fig. 3.

PLOT OF $\frac{2P_0-1}{1-P_0}$ AS A FUNCTION OF P_0



CONFLEX I

Fig. 4.



S – FIELD

Fig. 5.

AN EVALUATION OF RECENT DEVELOPMENTS IN THE FIELD OF LEARNING MACHINES

Oliver G. Selfridge
Lincoln Laboratory*, Massachusetts Institute of Technology
Lexington, Massachusetts

When it was suggested that I contribute a paper to this session, I had in mind that I would discuss and try and put into some kind of technological context the other papers of the session. Much of my own work of recent years has been in the field of learning machines, and artificial intelligence. There are some of us who are interested in seeing machines behave intelligently, and some of us who are only interested in having the machine simulate theories about how real brains work. I suppose that the former must predominate here, and I belong to that class myself.

It is therefore a reasonable question to ask how we shall recognize intelligent behavior in a machine when we manage to find some. I'm not sure that I can answer that except by saying that I should try to use the same standards that I use in people; but I start out by being prejudiced that people, my friends at least, are intelligent and that machines are not, even the ones I'm friendly to. There are a very few computer programs that have behavior which, even if not bright, cannot be called stupid; the famous checkers program by Arthur Samuel of IBM is one. But I do not call them intelligent yet. The learning that they do, while interesting and flexible, does not seem to me to show ineluctable evidence of being of the caliber that can be called intelligent. However, it is a real question.

Unfortunately, in the present session, that question does not even arise. I shall discuss them from the point of view of artificial intelligence, but none of the three other papers purports to discuss or propose subjects that have very much to do with artificial intelligence, so perhaps the kindest thing to do for the sake of consistency would be to retitle the session post facto. It may be argued that I am being over strict here and applying high standards of relevance not usually applied. After all, it may be said, even if these works do not themsleves discuss or propose techniques of artificial intelligence, perhaps the techniques they do discuss may be useful in artificial intelligence. That, of course, I do not doubt. Automatic Recognition Techniques[1], the first paper, may well have presented ideas useful in devices to classify and recognize pictorial inputs; that is, useful as part of pattern recognition schemes, which are almost certainly going to be necessary in any artificially intelligent machine. But when I attend a session on computer systems, I do not expect to find a paper on the manufacturing processes of ferrite cores, though they are certainly useful in most computer systems today.

Dr. Rosenfeld of course has no illusions and no pretensions on this score. In his abstract there was no suggestion that he purported that his work was connected with artificial intelligence except that he said, "This technique was suggested by an analysis of human recognition processes."

Dr. Rosenfeld's aim, as I understand from his abstract, is to deal chiefly with aerial photographs, but I intend to discuss his work as it applies to more general kinds of visual input.

For the kind of problems Dr. Rosenfeld is talking about, his techniques are almost certainly going to be useful. But I think that to some slight extent he overstates the generality of that applicability.

His technique is based, as he says, on "two fundamental operations, each of which serves to reduce the irrelevant information content of the input while preserving its major informational features." Here he is making a very arbitrary judgment about what is and what is not relevant. In some cases he is right, but often he will be wrong. Of the two operations, the first measures texture and classifies different regions according to it, and the second extracts simple geometric figures - "those involving simple straight line combinations (parallels, perpendiculars, etc.), circles, and so forth."

There seems to me to be no doubt he is describing important features of pictorial inputs, but they are not universally important, nor always even applicable. Dr. Rosenfeld defines texture as depending on the local first and second order distribution functions of the input amplitude; that is, on the mean and variance of the input amplitude locally. This will take care of many interesting cases. Fig. 1 is from a paper by Bela Julesz[2] of

*Operated with support from the U. S. Army, Navy and Air Force.

the Bell Telephone Laboratories in the recent special issue of the Proceedings of the Professional Group on Information Theory. The difference between the left half and the right half of the picture is that the mean amplitudes are different. (One must compute means and variances over not too small pieces of the picture.) In Fig. 2 each part of the figure has the same mean, but the variances are different. And we can tell that there *is* a distinction between the two parts, also, although it is not as clear as in the first case.

The notion of visual texture is certainly applicable to more cases than merely aerial photographs. Fig. 3 shows an example which is probably susceptible. The complicated crowd at the top of the picture would have, I presume, more or less uniform texture. Fig. 4 from the latest issue of *The Farm* can also be classified on the basis of Dr. Rosenfeld's texture. But such photographs seem to me to be rare on the basis of a few minutes hunting through magazines.

The same general comment ought to be made from the extraction of the simple geometric figures. Fig. 5 from the current issue of *Datamation* shows circles that can be extracted, but only because the photograph was taken full face; from any other points of view circles look like ellipses. The figure of the bullfighter and the crowd (Fig. 3) was taken from a recent issue of the magazine *Playboy*. That magazine has many examples, which I am not going to show as slides, of photographs with abundant texture and a happy absence of straight lines and simple geometric figures. It is on the basis of this evidence that I dispute Dr. Rosenfeld's assertion that "most important among these figures will be the simplest ones, e.g., those involving straight line combinations...circles and so forth."

But one should realize that there are other obvious limitations, as Dr. Rosenfeld well acknowledges. Fig. 6 has a consistent first and second order distribution all over the picture, though it is conceivable that there are some edge effects that *may* be detectable. But not as boundaries between areas distinguishable by his techniques. And if we add a noise level of a curious kind, as in Fig. 7, the square still stands out, though not so clearly as before.

Thus I am suggesting that the techniques of Dr. Rosenfeld are only a couple of extra tools for our work bench. And all the tools we have cannot approach our own visual processing system. Fig. 8 is blown up from an advertisement in Aviation Week, I believe. It is immediately obvious what the larger letters are, although most of them are absolutely unreliable. But look at the smaller ones Not only can we read the town, but the state as well. (It may help to defocus slightly.)

It ought to be said more often that real life neurons are, as far as we can discover, almost fantastically reliable in the ordinary sense. I don't know where they got their reputation for unreliability, but it is unfounded. And there are many creatures who thrive happily with *single* neurons handling many of the important functions of their bodies.

This theme, reliability through redundancy, is in great favor these days, and the paper by Professor Pierce[3] follows well in the footsteps of Von Neumann. The trend these days is somewhat away from the majority-of-three technique which others have analyzed in detail.

It is interesting in science and technology how the rise in popularity of some descriptive term causes it to be applied very widely and where it would not have been applied before. Some years ago in Lincoln Laboratory was developed a communication system called Rake[4] which handled the reception of a radio signal which had been smeared out in time by ionospheric multipath. The contributions from the various relative delays had to be undelayed, so to speak, and recombined to form the true signal. Fig. 9 shows the general scheme of the Rake receiver. But before the contributions are recombined they first have to be weighted by, guess what!, their reliability. Exactly as with Professor Pierce.

The reliability of the signal at each of the delays is measured by the factors a_i, and these factors are computed from the average correlation of the signal with both reference signals at that delay. This procedure is obviously identical with the procedure that Professor Pierce numbers IIB. (I might add that there are other communication problems that have been solved and analyzed in the same way.)

I mention this both because it is interestingly relevant and because my friend, Dr. Robert Price, who is the co-inventor with Dr. Paul Green, of Rake, tells me that he had not realized that Rake was an *adaptive* system till many years after it had been built. And are not automatic gain circuits adaptive, too?

I think I must disagree when Dr. Pierce says that "An adaptive.vote-taker is a simple type of self-organizing system," as he does in his summary. An adaptive vote-taker is surely a simple type of adaptive system, but I don't see where the self-organization comes in. The Rake system is a maximum likelihood detector, or perhaps I should say decision system, and it can be called *adaptive* if you like, though it may be stretching things a little. But its organization is always absolutely unchanged. Even so with the adaptive vote-taker; *its* organization and processing are always absolutely unchanging.

What such systems can do is the following: given a set of more or less independent variables each contributing to some decision, they can assign stable weights to the variables according to a maximum likelihood rule.

Dr. Pierce claims that "the use of statistical decision theory for building reliability is new..." Of course it is very far from new - Rake is an example among many others. His Bayes analysis has been carried out in a large number of other places, and is a standard part of elementary courses in communication theory. His Fig. 4 (shown in Fig. 10), which does not, by the way, look like an electrical circuit at all, goes back not merely to 1959, but is substantially the same as the neurons in the diagrams of McCulloch and Pitts[5] nearly 20 years ago.

I find myself slightly disturbed by the positive nature of Dr. Pierce's conclusion that his analysis has proved that "vote-takers are excellent devices for use in redundant systems." Since he has not in fact tried them, why is he so sure that actual systems will have the requisite statistical properties? And later, he says: "The use of even simple binary adaption can greatly increase the reliability... Consider a system with 100 stages... Errorless decision elements are used after each step." Well, if I were building the system, I would build it entirely of those errorless decision elements.

The third paper, by Dr. Malcolm Uffelman[6] of Scope Incorporated, starts out with "The general problem considered is the design of a system for the classification of minimally constrained stimuli." Most of the paper is in fact about that general problem. But I wonder why he calls it a "conditioned reflex" system. Let me first of all take a look at conditioned reflexes as they are known to the people who deal with them, who are called psychologists.

The conditioned reflex was of course first intensely studied by Pavlov[7], who set the level of Russian psychology ever since. You are no doubt all aware of the primitive experiments. A dog is shown a piece of meat and at the sight his mouth fills with saliva in anticipation. If he is shown a piece of meat and simultaneously a bell rings or a metronome ticks while he is salivating, then, after several such trials, the sound of the bell is sufficient by itself to cause salivation. The sight of the meat we call the unconditioned stimulus, and the sound of the bell the conditioned stimulus, and the salivation the response.

I hope that my description of it made it sound like a fairly simple phenomenon, because it is indeed a simple and very reliable experiment as I described it. However, if we examine the circumstances surrounding it, and wonder about its reasonable and logical extensions, it becomes very obvious that it is a simple expression of a very complicated and difficult underlying mechanism. For example, Pavlov showed that the conditioned stimulus must precede the unconditioned stimulus in order to establish a stable conditioned response, and in general the order of the events seems very important to the organism. Now why should that be?

There is another large question which has bothered psychologists and still does. Suppose that we have conditioned the dog to salivate to the meat on hearing a note on the piano, say Middle C. To what other stimuli will the dog salivate? Will the dog salivate to the next note D? Suppose we play Middle C on a bassoon? and so on. Suppose we choose a note three octaves away? Suppose we give him some stimulus which isn't even an auditory one? In general the responses vary, and we are more likely to see responses with the first examples I gave than with the later. The phenomenon here is called stimulus generalization. The opposite phenomenon also occurs, called stimulus differentiation. If the dog salivates both to Middle C and Middle D, say, we can train the dog to differentiate between them by reinforcing one of them (by actually giving the dog the meat) and never reinforcing the other.

Now in what respects can we say that Conflex I exhibits anything like a conditioned reflex? Not in very many, I am afraid, Nothing depends on the order of presentation. Nothing corresponds to extinction, for instance. Extinction is the name given to the phenomenon where the response to the conditioned stimulus gradually dies down when it never receives the meat.

Dr. Uffelman has made here a very curious choice to use Random Stimuli in random classes. Why on earth random? Are they supposed to be more useful because they are random? At least none of the enormous body of work or conditioned reflexes has ever used random stimuli. We may not easily understand or analyze the binding rules for classes of similar stimuli for man and beast, but we may all be very sure that they are not random rules.

Conflex I exhibits only generalization and differentiation. From the paper Dr. Uffelman presented, he claimed no more. He says also that generalization is based purely on stimulus overlap. That is, Conflex I averages all the signals in one class so that those units always present have a large vote in favor of that class. One question here is what are the units of the signals. They must be the cells in the D-field, which correspond to some combination of the actual receptor units.

Let us re-phrase that statement while treating

Dr. Uffelman's diagram (Fig. 1) as an unreliable binary process. The inputs to the summing elements are given heavier or lighter vote weights by the weighting factors m_{ij} here according as the inputs are more or less reliable. In other words Conflex I is reállly an adaptive system of the kind discussed by Dr. Pierce.

Naturally there are differences. I should like to be able to compare their results directly, but I cannot. For one thing, Dr. Uffelman's unreliability comes from and with the signal, rather than from malfunctioning of circuit elements.

In Dr. Uffelman's conclusions he states "The theory developed...has shown that a relatively powerful, adaptive pattern recognition system can be built." Relative to what? Certainly not relative to anything that shows conditioned reflexes. And certainly not relative to other pattern recognition techniques which go beyond mere overlap.

It is true that his scheme permits "a simple method for expanding the class size of a basic system," by means of the "M-fields." Now in the ordinary run of the mill random net study the M-fields themselves do not form a separate layer at all but are merely the weighting of the connections. An analysis Marvin Minsky and I[8] did two years ago at the London International Information Theory Symposium applies to Dr. Uffelman's scheme.

The other comments we made in that paper hold, too, but they reflect more widely on whether one can extend the techniques (that can make a machine like Conflex I work) to perform tasks that are harder.

For there are tasks, even in this context, that are harder. Instead of assigning the stimuli to the classes at random, supposing they matched to some real property by a fairly sophisticated one. For example, suppose that one class consists of pictures of fraternal twins and another of identical twins. Where are the random fields sensitive to this difference?

There is thus an enormous gradation of classification or pattern recognition problems. Conflex I can learn to recognize and respond to a black blob on a white background. It can certainly not separate fraternal from identical twins. I believe that the threshold of its capabilities is not very far along the road from the first to the second. Sadly, I have only vague ideas about how to proceed faster or much further down the road.

As I said, the implications for artificial intelligence of Dr. Uffelman's paper and Dr. Pierce's paper are roughly the same. We can optimize a system with many variables, so that each pulls his appropriate weight. But only if their effect is, say, additive or independent.

If they have to be combined in some complicated non-linear way, we have very little to go on. The trouble is that most of the interesting problems seem to involve just such cases. Although many of the people with random nets will tell you their nets can generalize beyond simple overlap technique or simple linear maximization, they have not yet shown any examples.

Dr. Rosenfeld has described his work as a tool which may be useful for artificial intelligence. Both the other papers may be useful tools, too. One thing I miss very much is the experiments which can tell us whether those tools are in fact useful; and to what extent and in what way. Perhaps the authors can soon evaluate this work in the best way, by seeing how valuable they are to the advancing front of science and technology.

REFERENCES

1. Rosenfeld, A., "Automatic Recognition Techniques Applicable to High-Information Pictorial Inputs," Proc. IRE, March 1962, to be published.
2. Julesz, B., "Visual Pattern Discrimination," IRE Trans. on Information Theory, Vol. IT-8, No. 2, p. 84.
3. Pierce, W. H., "Adaptive Decision Elements to Improve the Reliability of Redundant Systems," Proc. IRE, March 1962, to be published.
4. Price, R. and Green, P. E., "A Communication Technique for Multipath Channels," Proc. IRE 46, p. 555.
5. McCulloch, W. S. and Pitts, W., "A Logical Calculus of Ideas Imminent in Nervous Activity," Bull. Math. Biophys., Vol. 5, pp. 115-137, 1943.
6. Uffelman, M. R., "Conflex I - A Conditioned Reflex System," Proc. IRE, March 1962, to be published.
7. Pavlov, I. V., CONDITIONED REFLEXES, 1927.
8. Minsky, M. L. and Selfridge, O. G., "Learning in Random Nets," Reprint from INFORMATION THEORY, Fourth London Symposium published by Butterworths, 88 Kingsway, London, W.C.2.

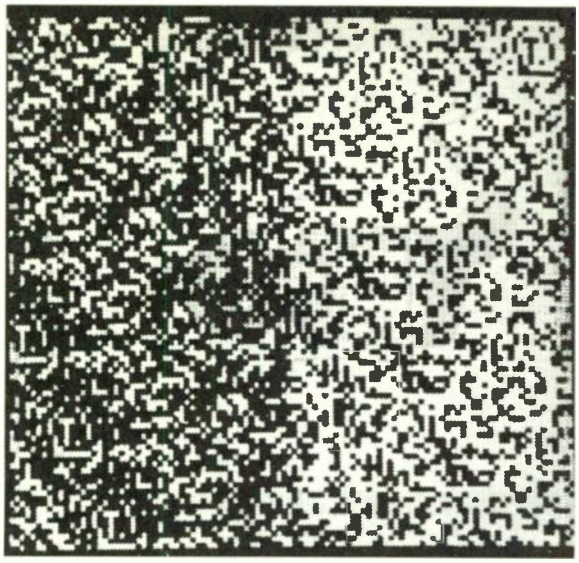


Fig. 1.

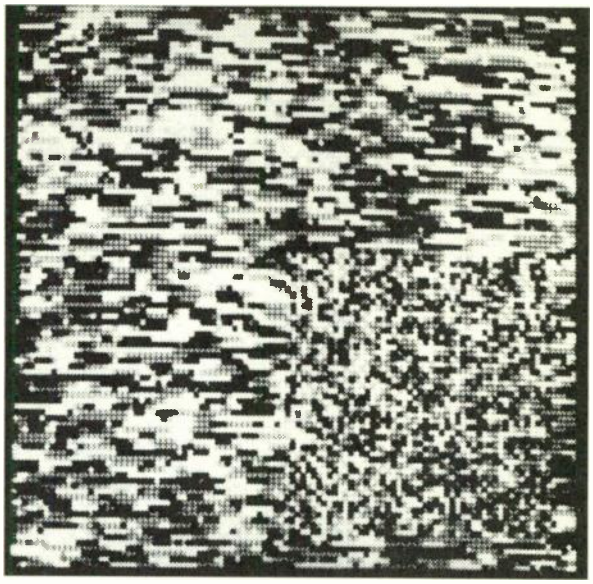


Fig. 2.

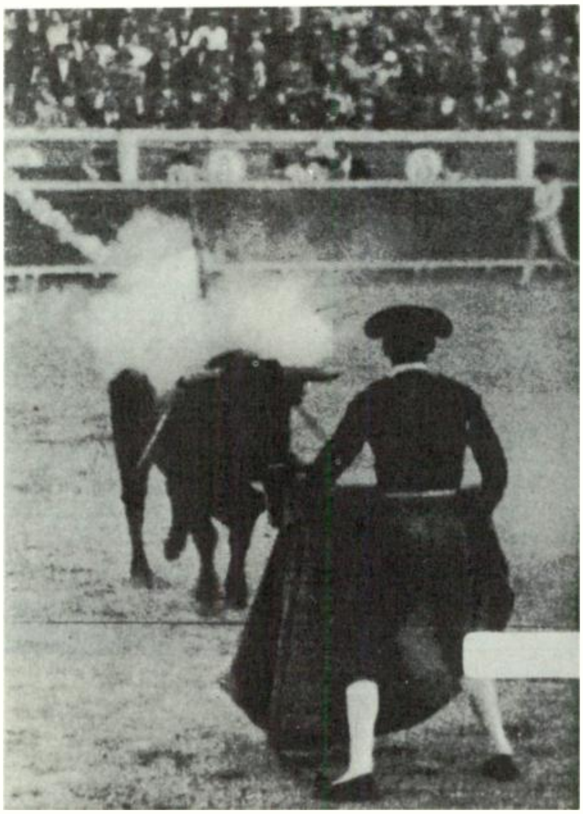


Fig. 3.

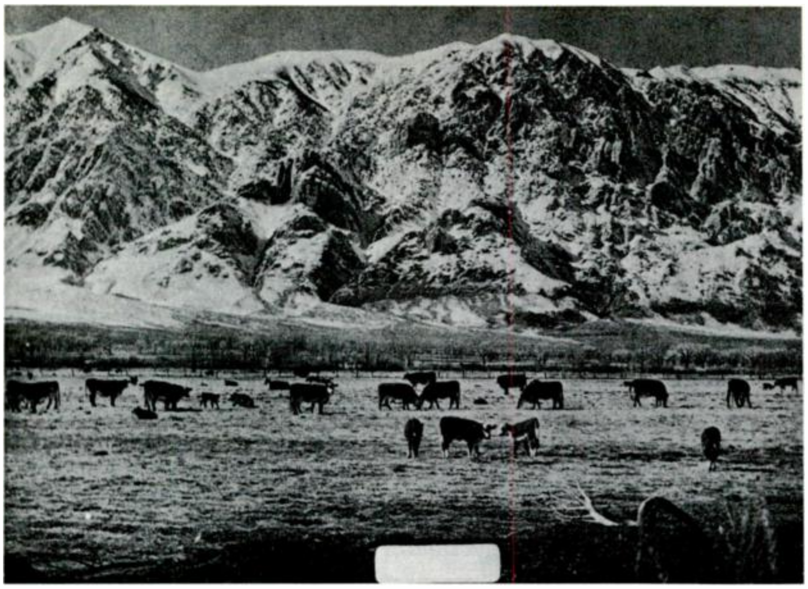


Fig. 4.

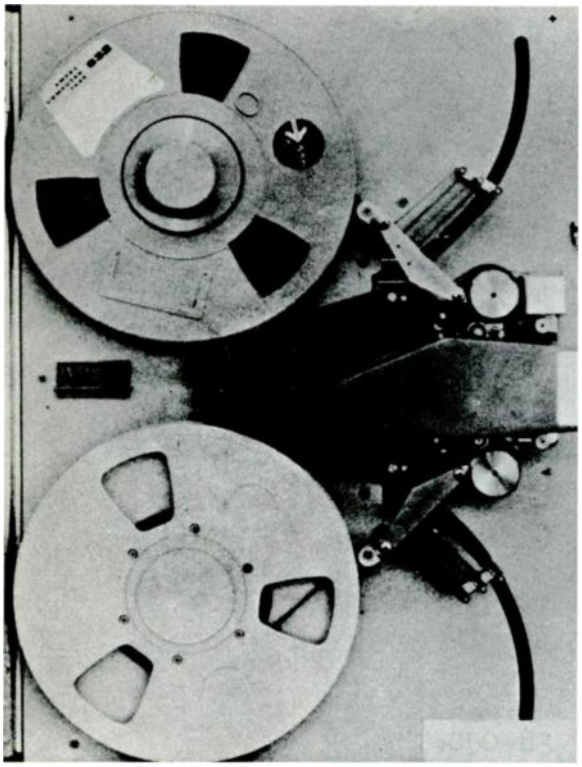


Fig. 5.

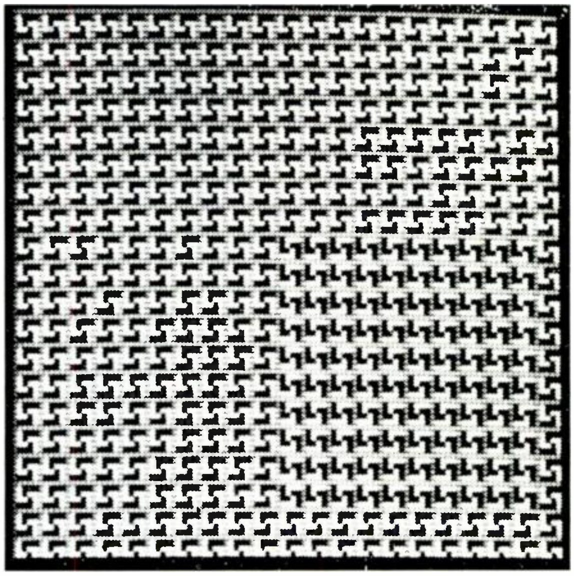


Fig. 6.

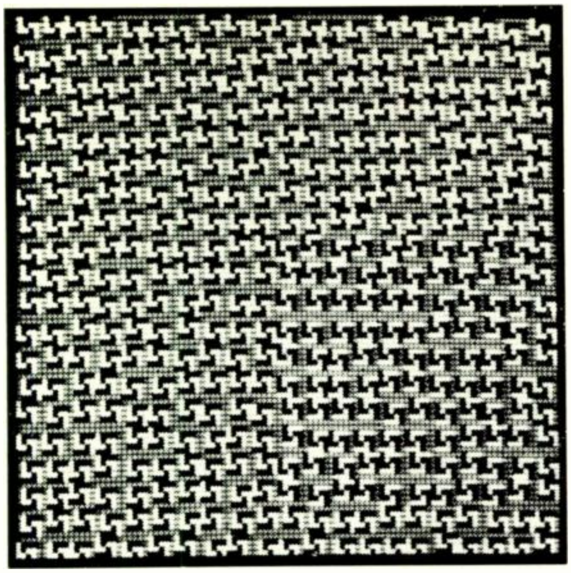


Fig. 7.

Fig. 8.

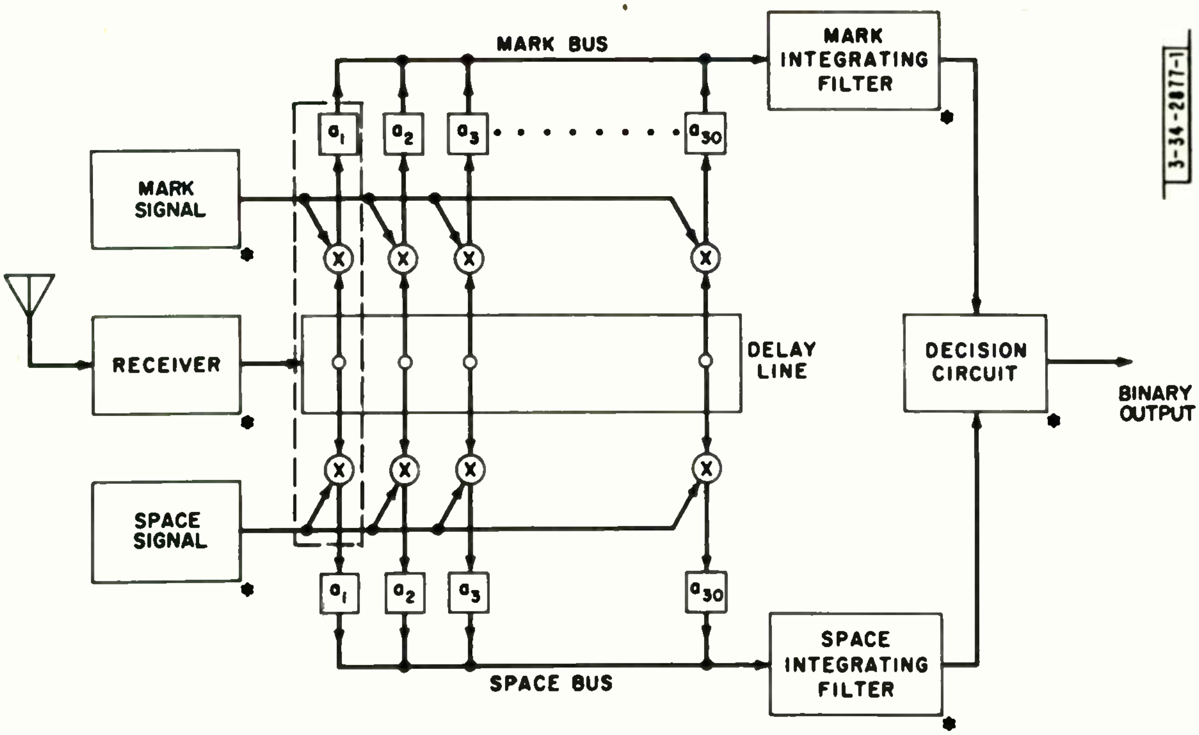
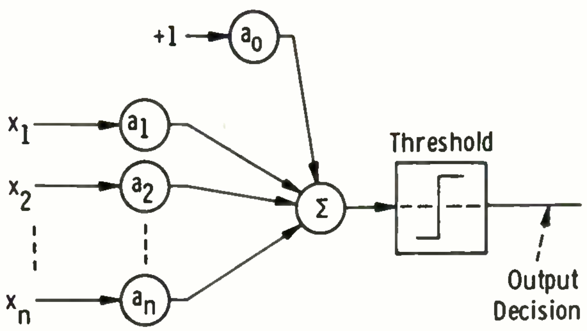
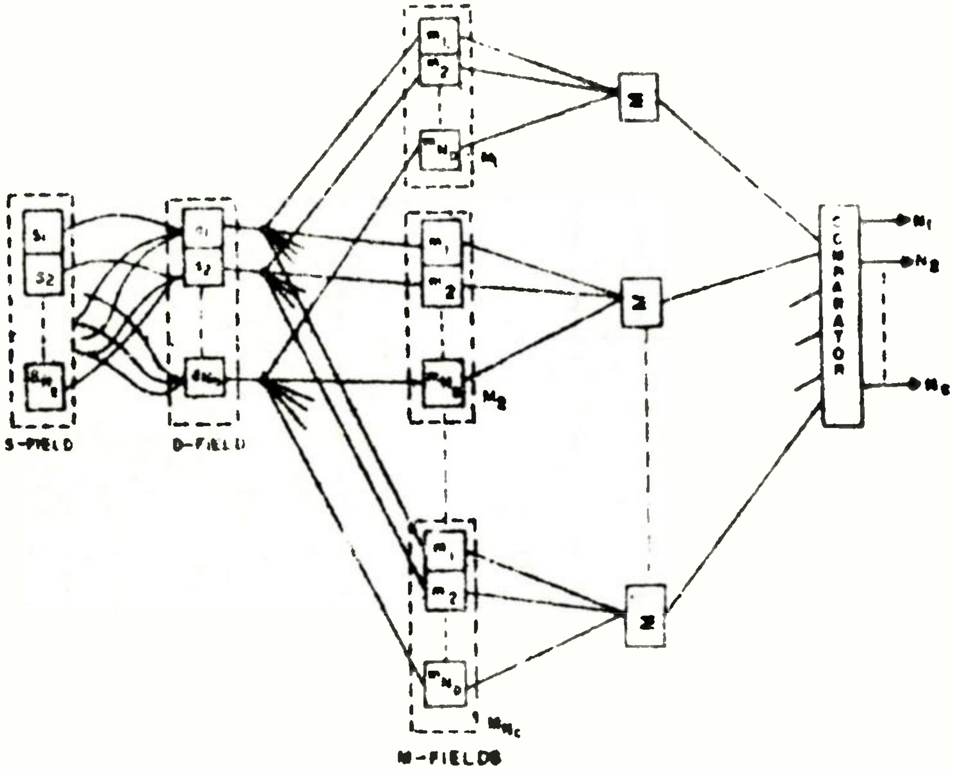


Fig. 9.



A decision element or vote-taker for binary inputs of +1 or -1

Fig. 10.



ORGANIZATION OF THE CONDITIONED REFLEX MODEL

Fig. 11.

PANEL: STATUS SESSION ON INFORMATION THEORY

Chairman: L. A. Zadeh
University of California
Berkeley, Calif.

Panel Members:

P. Elias
Massachusetts Institute of Technology
Cambridge, Mass.

R. Fano
Massachusetts Institute of Technology
Cambridge, Mass.

S. Golomb
Jet Propulsion Lab.
California Institute of Technology
Pasadena, Calif.

D. Slepian
Bell Telephone Laboratories, Inc.
Murray Hill, N. J.

Abstract

This session was designed to serve as a forum for the presentation and critical analysis of recent developments in selected areas of information theory. The discussion was conducted by a group of five experts and was centered on the three status papers presented at this session.

*APPLICABILITY OF CODING THEORY TO PHYSICAL CHANNELS

P. Elias
Department of Electrical Engineering
Massachusetts Institute of Technology
Cambridge, Mass.

*INSTRUMENTATION OF ENCODING AND DECODING

R. Fano
Department of Electrical Engineering and Lincoln Laboratory
Massachusetts Institute of Technology
Cambridge, Mass.

*NOISE AND WAVEFORMS

D. Slepian
Bell Telephone Laboratories, Inc.
Murray Hill, N. J.

*The text of this paper was not available at the time of publication.

THE "WHIRLING DERVISH" A SIMULATION STUDY IN LEARNING AND RECOGNITION SYSTEMS

Allen Hoffman
Philco Corporation
Blue Bell, Pa.

Summary--An optical correlation device is described. Its function is to generate correlation data between two dimensional patterns and random line templates. The output of the device is paper tape containing the correlation data and sufficient flagging and sync. information to facilitate digital computer processing. The use of this device in conjunction with a digital computer is discussed as a simulation procedure for the evaluation and design of recognition systems.

I. INTRODUCTION

Basically the Dervish is an optical correlation device. It was constructed from a 35 mm T.V. Color Cinescanner. Figure 1 shows the cinescanner with its associated film loop. The function of the device is to measure the degree of correlation between an inserted pattern and internally stored criteria. The criteria take the form of two dimensional patterns on a continuous film loop. The loop revolves about a bracket utilizing the normal cinescanner film transport mechanism. The loop makes one complete revolution or cycle for each pattern inserted. Early viewers of this weird gyration dubbed the device "The Whirling Dervish" and the name has been with the device ever since. Now that the mystery has been cleared up, let us take a look at the pattern analysis game.

I will confine my remarks to optical systems, although this is not meant as a restriction. This approach allows for simple descriptions and avoids the necessity for a discussion of sensors. When analyzing a pattern, the statement "Show to the device" will be used. If the reader wishes he may envision an optical imaging procedure, however the digestive process by which the device receives information to be analyzed is not important at this point.

If we wish to design a system capable of recognizing the difference between a Capital A and a Capital B in the optical sense, we have only to use the negative of either pattern as a template, and measure the degree of correlation between the patterns and the template. This deterministic test would be sufficient provided:

1) We had sufficient apriori knowledge about the form of the patterns to be analyzed. (In this case, we must know in advance that the problem is separate A and B.)

2) The patterns were always the same and always registered.

All is well, provided we are satisfied to only analyze stylized print or read coded bank checks. Even such devices as Post Office Mail Readers can be simply constructed. All one needs is sufficient knowledge of the form of the letters to be read in order to design the templates. The templates may be very complex, depending on the nature of the material to be matched. Furthermore the matching procedure may involve elaborate weighting systems. Most devices in this class are nevertheless template matchers and computer generated recognition templates are common place at present. Such problems as variation in size, orientation, and spacing of letters are generally solved external to the reader by special circuitry. Adjustments are made in the sensory system, and the stored criteria to correct for these problems before the recognition procedure begins.

If, however, we do not have sufficient knowledge about the form of the patterns to be analyzed, how do we design the templates? Obviously the deterministic recognition system fails.

What is required, is a system that does not merely discriminate between one pattern and another, but rather analyzes classes of patterns and is capable of extracting the properties of a class that are common to all its members.

Such devices are generally titled "Learning Machines." They "Learn" in the sense that they accumulate information about the properties of a particular pattern class after having been shown examples of that class. These examples are usually referred to as "Training" examples.

One of the early devices based on this principle was the "Rosenblatt Perceptron." It could be trained by repeated exposure to training examples to recognize the letters of the English alphabet, in fixed position and font. By means of the "Perceptron" and by computer simulation, Rosenblatt demonstrated that learning can occur in a linear summation network under very general conditions of repeated presentation of input and desired output patterns. The theory of the Perceptron will not be presented here. It suffices to say that; it was one of early successful attempts to introduce the learning concept into a cognative device.

More advanced versions of the simple perceptron model have been postulated, and some have actually been built. The literature contains a great many examples of simulated perceptrons, Trial and Error learning schemes, Conditional Reflex Networks, Artificial Neurons, Self-Organizing Devices and a host of other learning and recognition systems.

Although they vary widely in principle, many of the systems share one common property. They gather statistical data concerning the training examples by means of cross-correlation techniques. The training patterns or signals derived from them are compared to predetermined criteria. In the case of special purpose recognizers the criteria can be carefully designed. In the general case random criteria are usually used. The significance of random criteria, and a better understanding of the learning machine principle, will best be had if we consider a specific example. Figure 2 shows some of the items required for a discussion of this specific example.

Consider the following two classes of patterns, hand written A's and B's with much variation both in form and orientation. Define these as class 1 and class 2. In the case of our optical system consider the training examples as opaque patterns on clear slides. Also consider a set of opaque random patterns, here labeled R_1, R_2, etc. The correlation procedure involves superimposing each example with each of the random criteria and observing whether or not the transmitted light exceeded a preset threshold. If there is sufficient correlation to exceed the threshold we say that the random unit fired. If not we say the unit did not fire. For example, suppose we show all 100 examples of class A to random unit R_1 and it fires 80 times. Now, show all 100 examples of class 2 to unit R_1 and suppose it fires 56 times. The number of firings for each class is in effect the learned data. If we, now, show to the unit R_1 an unknown pattern and it causes R_1 to fire, most of us would agree that the odds are 80/56 that the unknown pattern belongs to class 1 rather than to class 2. This is certainly not a completely convincing argument. It does not take too much imagination to conceive of an A that could pass for a B and elude detection by this one test. If however we use the statistics associated with many statistically independent tests to evaluate our probability, the argument becomes more convincing. The more members of each class we have in the training sequence the more significance we can attach to the training data, likewise the more tests we make or the more random masks we use in the procedure the stronger the result.

Note at this point that there was nothing in the system that dictated what the two classes had to be. They could just as easily have been crossroads versus geographical boundaries taken from aerial photographs.

Also note the method of analyzing the correlation data. The training data is the summed firing of the random units and this data is used for a straight probabilistic scoring. It follows the philosophy of Dr. Gamba of the Univ. of Genoa. It is by no means the only way to process the correlation data, nor is it advertized to be an optimum procedure.

The "Whirling Dervish" was conceived and built to generate correlation data. The Dervish program has as a goal the simulation of many learning and cognative systems using as common input, this raw correlation data. By careful analysis of the simulated systems, direct comparisons can be made in terms of error rates.

II. CONSTRUCTION OF THE DERVISH

The Dervish (as was noted previously) was constructed from a 35 mm Cinescanner. The use of this piece of equipment as a starting point offered certain distinct advantages. First it contained all the required mechanisms for handling and moving 35 mm film. Secondly, it contained all the required optics for superimposing images. Lastly, it contained three color channels complete with photomultipliers and associated circuitry.

A block diagram of the Dervish is shown in Figure 3. The random criteria are shown here on a continuous loop labeled "Loop A." These random patterns were generated from random noise in a two color display. That is, red random lines on a blue field.

The training examples are shown on a strip, labeled "B". A_1 - corresponds to Class A, example number 1. A_2 - Class A, example number 2 and so on, until we reach A_n or end of class. This is followed by the next class of patterns to be used to train the device.

The procedure is to optically correlate each training example against all of the random masks. Thus we generate R_1A_1 as the correlation signal for mask number 1 and example A_1. This is followed by R_2A_1, R_3A_1, etc., until we reach $R_{100}A_1$, or the last mask. The number 100 is not a restriction, but merely an arbitrary choice for the number of masks in this example. We have now advanced the loop through one complete cycle. The second training example is now advanced into place and the procedure continued. All this occurs at a rate of 24 frames per second. The random mask loop also contains one green frame, that is used to provide a signal for our start-stop logic circuitry.

The device contains three phototube circuits. Dichroic mirrors are provided to separate the red, blue, and green components of the transmitted light signal. Red light corresponding to the degree of match between the patterns and the random lines. Blue corresponding to the match between pattern and areas surrounding the lines, and green light indicating a completed cycle. The correlation data may be processed in a variety of ways (several of which are shown in Fig. 3). Consider output (one) as a typical signal. It is a measure of the correlation between the random lines on the masks and the class examples. The output is fed to a preset threshold circuit. If a particular correlation signal exceeds this threshold, we send a "YES" answer out, in the form of a pulse. If the correlation signal does not exceed this threshold we indicate this by the absence of a pulse (or a zero output). The threshold is set so that the probability of firing for an arbitrary pattern with a 50% transmission is one half.

The "YES" or "NO" output signals from the threshold circuit are recorded with the proper flagging and sync information in the form of punches on five level paper tape. This paper tape serves as the input to our various Computer Simulation Programs.

Figure 4 shows the order in which the bits of information are stored on the paper tape. The upper diagram illustrates the mode in which information is stored. The circles represent potential storage positions. A punch at one of these positions indicates a "YES", or fire condition, for the correlation test. The absence of a punch indicates a "NO". The 5th position on the tape is used for flagging operations. It is always punched except when "end of look" is to be marked.

The lower diagram shows the groupings on the paper tape. Here we have the correlation data (or test responses) for the first example of class A. This group contains one position for each test, or mask. This is followed by a group of positions that are used for identification, sync and flagging. It contains codes to indicate end look, end class, end sequency, and end of unknowns. Incidently, the procedure for correlating unknown patterns is the same as for training examples. The results are grouped in the same manner immediately following the data from the last training class.

The computer programs are designed to test for the presence of these flags at various points in the data processing routines. They process the training data first and then investigate the unknowns.

The buffer store, shown here between the threshold circuitry and the teletype punch, serves a dual purpose. First it inserts the punch in the fifth position on the tape and flags that position at the conclusion of each look and secondly, it provides the Mark and Space pulses required for the teletype operation. The first four information bits are fed into the "A" register in a serial fashion at a synchronous rate of 24 frames per second. Register "A" fills in 167 milliseconds. The four bits are then transferred in parallel into four positions in the "B" register. The "B" register contains 7 positions, as shown here. The teletype will accept five bits 22 milliseconds long preceded by a 22 milliseconds space pulse and followed by a mark pulse which must be at least 33 milliseconds long. Six advance pulses spaced 22 milliseconds apart followed by a 35 millisecond spaced pulse (adding up to 167 milliseconds) empties the "B" register satisfying both the synchronous requirements of the teletype punch and the synchronous rate at which new information is available. The only problem remaining is obtaining the required advance and transfer pulses to drive the buffer registers.

Both the teletype motor and the cinescanner drive motor are of the synchronous variety and are synchronized with the 60~line. A timing disk (Figure 5) is geared to the cinescanner such that it makes one revolution in 1/6 of a second. Thus 167 milliseconds in time correspond to 360° of rotation of the disc. Slots are cut in the disc, allowing us to obtain light pulses through the disc to photodetectors. In this manner, we generate the proper time relationships for our system. The pulses thus obtained are processed, shaped and used to control the shift-register drivers. The outer circle contains the 7 slots required to advance the "B" register. Note the asynchronous spacing. The inner ring contains 4 evenly spaced slots required to advance the "A" register, and a transfer pulse slot is available to link the two sections of the buffer.

Figure 6 shows several of the random masks used in the system. Note that random lines are used. Random points or randomly dispersed opaque areas (or blotches) could also have been used. Intuition steers us toward random lines as a starting point when dealing with lined patterns. Which procedure is best, is not really known. The Dervish allows us to try many different forms of test criteria and experimental results will indicate the best approaches.

These particular patterns were generated in the following manner. First, bandwidth limited random noise was used to drive both axes of an oscilloscope. This caused a random tour of the spot on the face of the CRT. The display was photographed using polaroid slide film. The slides thus obtained were used in the T.V. signal center to generate a two color display on the face of a shadow-mask color TV set. Figure 6 contains 35 mm photographs taken directly from the face of the color set. The masks are processed into a continuous loop for use in the device.

Let us consider the degree of randomness in the masks. It is admitted that the best set of masks (and in fact the only valid set for probablistic inferences) is one in which all members exhibit statistical independence. It is further admitted that no effort has been made to measure the interdependence in the sets used in the system. The only defense for this method of generating masks is the ease and speed with which it can be performed. Rather than generate 100 orthogonal patterns (which is difficult and time consuming) the decision was made to generate 1000 masks from random spot tours and pick as criteria the 100 that yield the highest degree of discrimination in a particular problem. The generation of tens of thousands of masks is entirely feasible. The Dervish can be used as a mask selection device. Given a specific problem, in terms of prescribed classes, we can now generate a suitable set of criteria from our large stock of potential tests that will make the required separation. Mask selection is one of the most important features of this system.

To date, two methods of raw correlation data processing have been considered. They are the "Probabilistic scoring philosophy" (as previously described) and followed by Prof. Gamba, and the "Adaptive Training Algoritum" indicative of the "Reward and Punish" Perceptron type devices (developed by Rosenblatt). Computer programs have been written to perform both tasks.

Program 1 sums the number of firings for each mask and converts the number of firings to a logarithmic quantity. In the prerecognition section of the program the sums are logged and weighted. This information is then used as training data for the processing of unknown data. When shown an unknown pattern, the program computes the probability that it belongs to each of the classes used in the training routine.

Program 2 is designed to perform a dichotomy decision by creating a mask matrix, each element of which corresponds to a particular mask in the previously described loop. The firing data associated with each class and mask is compared to the elements of this matrix. Assuming an initial value of unity for each position in the matrix we proceed as follows. First compare the responses of each mask with its matrix position. If the mask fired reward the position by incrementing its value (in accordance with a particular truth table). If the mask did not fire punish it by decreasing its value. This procedure is continued for all training examples. The examples of each class are considered alternately. It was arbitrarily decided that the mask matrix to be constructed will cause the members of class I to produce scores less than or equal to a threshold value, and class II to generate scores above the threshold. This process is continued repeatedly until all examples are classified when compared to the matrix. The mask matrix is now trained and may be used to separate out and classify unknowns.

This procedure was found to be extremely useful in the weighting of masks and yields and optimum set of masks from our stored collection for a given problem. This is provided the training correction is continued until all training examples are separated.

It may be of interest to note at this point some of the preliminary results of the use of these two programs.

Handwritten characters were used to train the device. The letters A, D, and C were considered. Thirty examples of each were used for training. After training, all ninety of these examples were rerun through the system for recognition. There were 8 errors using probablistic scoring.

Using unknowns (or letters not used in the training sequence) the probabalistic scoring technique was compared to the Perceptron technique. The error rates were 14% for Perceptron and 20% for Probabalistic scoring. Although this indicates a slight advantage to the use of Perceptron type processing, neither of these error rates is very impressive. Improvements are anticipated when larger problems are run. Problems using several hundred training examples and 1000 masks are planned.

Many other simulation programs are planned, including layered type systems in which the training is performed in terms of features instead of letters. In addition non-alphanumeric data will be considered. Attempts will be made to analyze aerial photographs, nuclear traces, and general two dimensional patterns.

Any phenomena that can be transducer to two dimensional patterns are candidates for analysis using the "Whirling Dervish" as a first level correlator and suitable computer programs.

In conclusion, I would like to make the following acknowledgements.

The "Dervish" was conceived, designed and built at the Philco Research Laboratories in Blue Bell, Pennsylvania. The computer programs were written with the assistance of the Philco Research Engineering Staff. In addition I would like to thank Mr. Steve Moulton for sanctioning the presenting of this paper on behalf of the Philco Corporation.

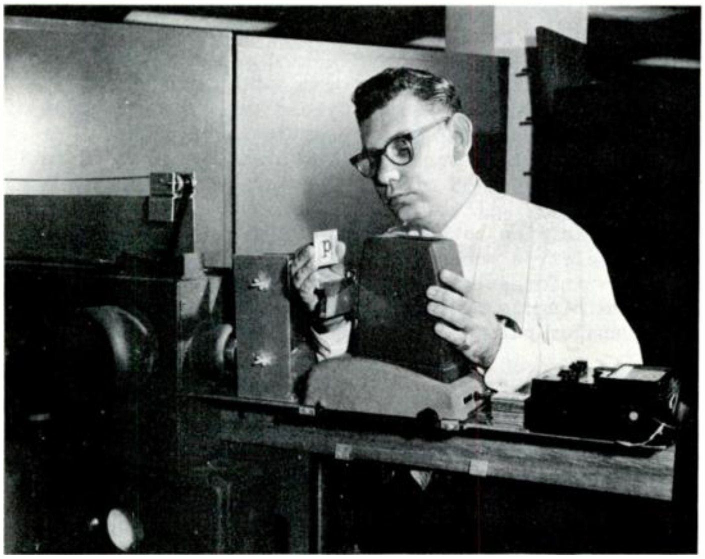


Fig. 1. Photograph of "Dervish." Constructed from 35 mm Cinescanner, with projector replacing flying spot scanner on front end.

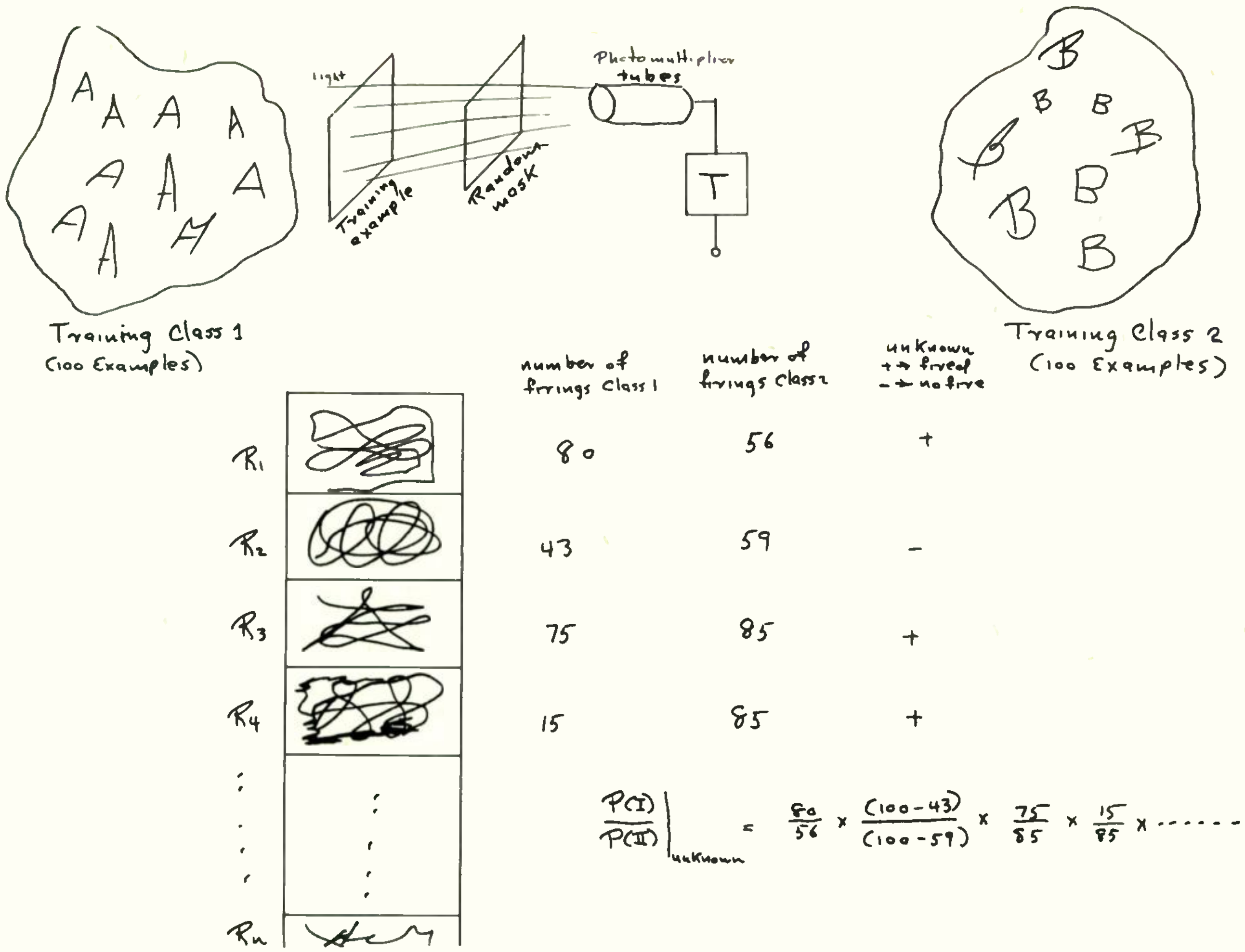

A A A A A A A A A A
light
Photomultiplier tubes
T
Training example
Random mask
B B B B B B B B
Training Class 1 (100 Examples)
Training Class 2 (100 Examples)
number of firings Class 1
number of firings class 2
unknown + → fired − → no fire
R1 80 56 +
R2 43 59 −
R3 75 85 +
R4 15 85 +
Rn
$\left.\frac{P(I)}{P(II)}\right|_{unknown} = \frac{80}{56} \times \frac{(100-43)}{(100-59)} \times \frac{75}{85} \times \frac{15}{85} \times \cdots\cdots$

Fig. 2. Example of Probabilistic Scoring. Numbers are chosen arbitrarily for discussion in text of paper.

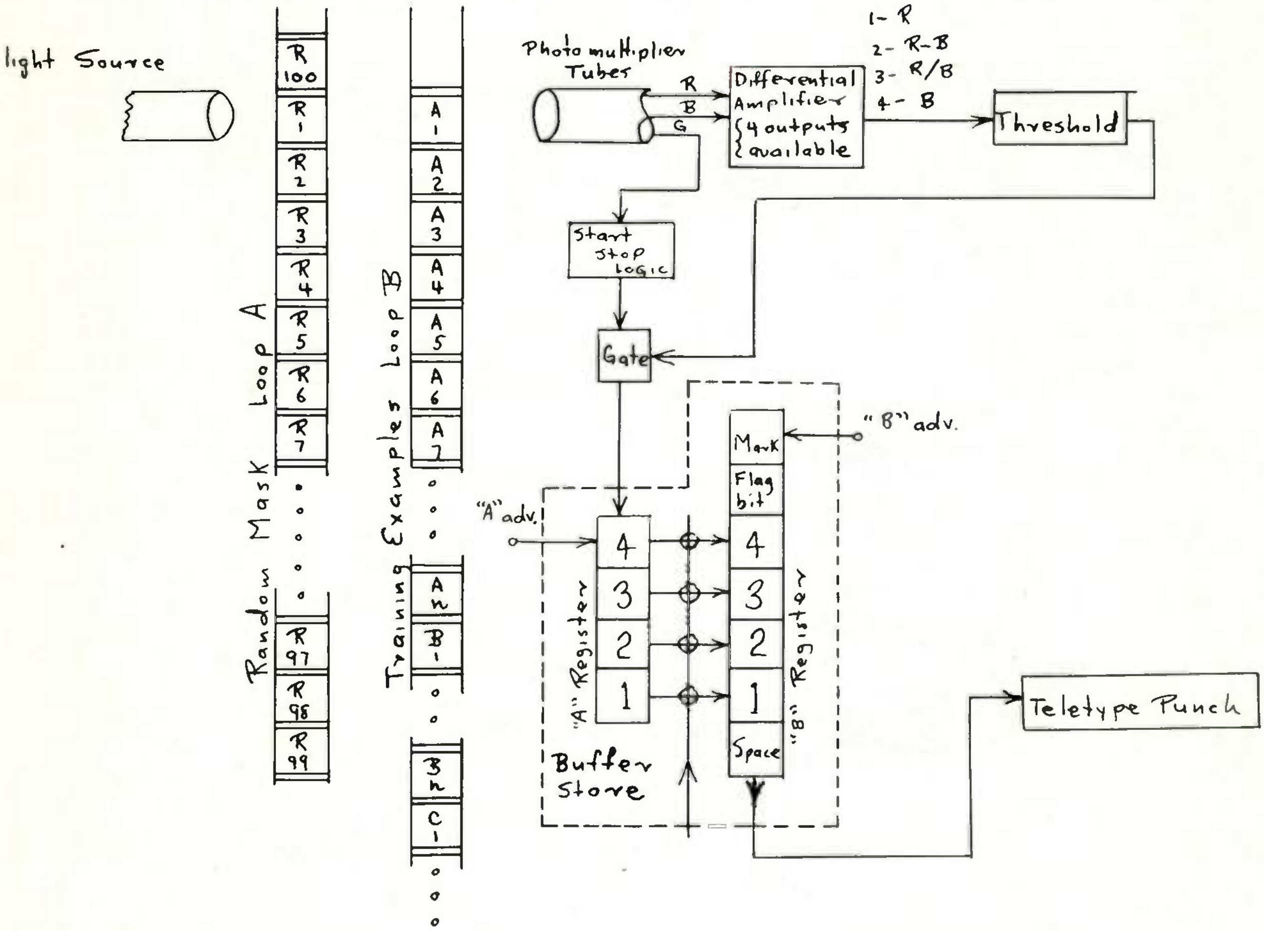


Fig. 3. Block diagram of "Whirling Dervish" optical correlator.

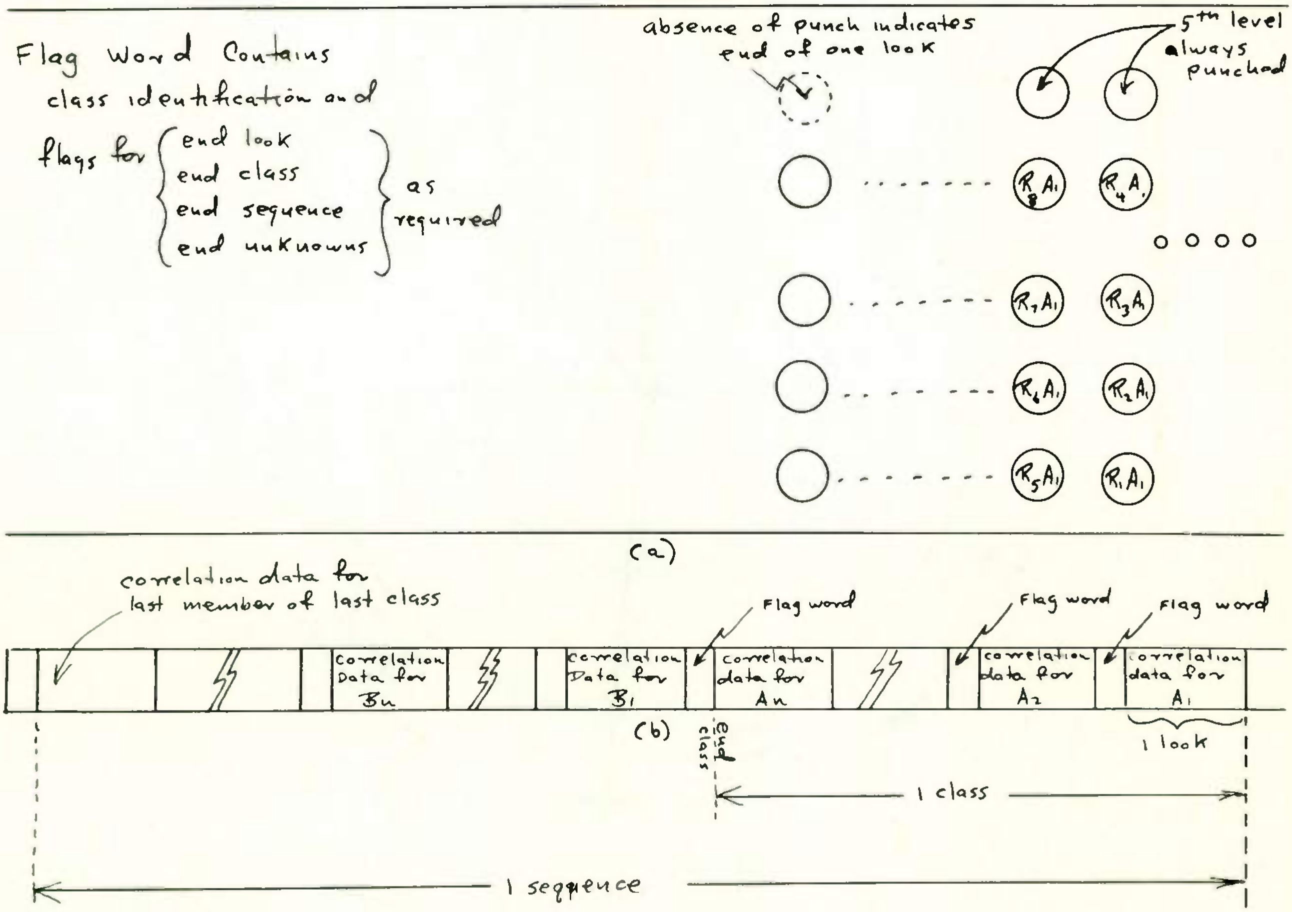


Fig. 4. Paper tape format. (a) Position of individual information bits. (b) Groupings on tape.

360° → 167 milliseconds

inner ring → advances register A

outer ring → advances register B

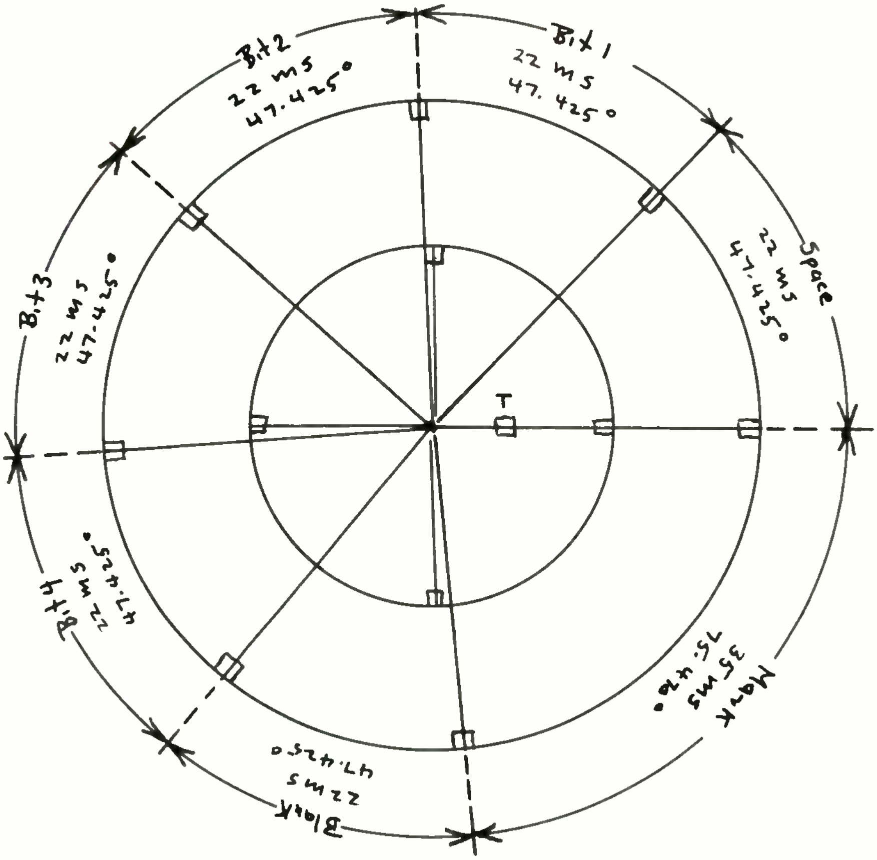


Fig. 5. Timing disc.

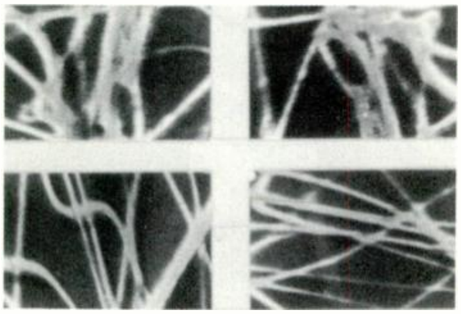


Fig. 6. Random line patterns generated from random noise.

A METHOD OF DIGITAL SIGNALLING IN THE PRESENCE OF ADDITIVE GAUSSIAN AND IMPULSIVE NOISE

Ludwik Kurz
Department of Electrical Engineering
New York University
New York 53, New York

Summary

This paper considers the basic problem of transmitting digital information through a noisy channel with minimum probability of error in finite time. The transmitted signals are average-power limited, and the interference is assumed to be a linear combination of additive Gaussian and impulsive noises. The Gaussian noise is specified by its corresponding demodulated power density spectrum (either flat or increasing with increasing frequency), and the impulsive noise by the statistics of its amplitudes (large-variance, zero-mean) and the times of occurrence of the impulses (Poisson's distribution is assumed but other distributions can be handled in a similar manner). The noise components are assumed to be statistically independent of each other and the signals. The theory of efficient and equidistant codes[6,8,9] is extended to the case of Gaussian and impulsive noise. Basically, it is shown that subdividing the decision interval (0,T) and coding each subinterval using optimum waveforms is a more efficient method of combatting the impulsive noise than using optimum waveforms in the full decision interval. Namely, the decision interval (0,T) is divided into u equal duration subintervals. Each subinterval is coded for best performance with no impulsive noise present. A non-linear detector suppresses the contributions of subintervals with impulsive noise -i.e., the decision at the receiver is essentially based on the subintervals with no impulsive noise. To find the optimum value for u and an estimate of the error probability, the concept of an ideal discarding detector is introduced. It is shown that considerable improvement in the reliability of the system can be achieved by decreasing the transmission rate. For white Gaussian noise, the performance of a system using the repeated equidistant code is optimum from the point of view of reliability and simplicity of the resulting detector. Performance results are given for different codes for white Gaussian noise, and Gaussian noise with a power spectrum that increases with increasing frequency. A simple receiver circuit for decoding equidistant and repeated equidistant codes is suggested.

Introduction

The purpose of this paper is to extend the theory developed in a previous paper*[9] to a more realistic case when a communication link is disturbed by additive Gaussian and impulsive noise[10, 11,12]. Both components of noise are assumed to be statistically independent of each other and the signals.

Wherever possible, the mathematical notation is the same as in the previous paper (I). Whenever changes are made or new additional symbols are introduced, a clear explanation accompanies them. It is assumed also that the reader is thoroughly familiar with (I).

The following model is used to describe the behavior of the impulsive noise. The impulsive noise will be assumed to consist of very narrow pulses (approximating impulses) occurring at random with the frequency of occurrence of the noise pulses governed by Poisson's distribution** and the amplitudes of the noise pulses governed by a large-variance and zero-mean but otherwise unspecified probability density function. As in (I), the Gaussian noise is specified by its noise power spectrum either flat or increasing with increasing frequency. The latter model of the Gaussian noise power spectrum acts as a weighting function which confines the signal to an available band of frequencies. Other basic assumptions are the same as in (I).

The first part of the paper considers in detail the optimization procedure for binary or the one-bit case. This case is of practical importance if the communication link connects two relay stations with limited decoding ability.[3] The main purpose of this part of the paper, however, is to present techniques for combatting interference composed of both additive Gaussian noise and impulsive noise.

The remainder of the paper is devoted to the extension of the binary or one-bit case to the multi-signal system or m-bit case. Several methods of coding are discussed and the associated optimum detectors are found. A modified version of a simplified detector for equidistant codes is given. The performances of various systems, using different methods of coding are compared for the cases in which the interference is a linear combination of white or non-white Gaussian and impulsive noises.

Optimum Coding and the Associated Detector for the Case of One Bit of Information Transmitted Through the Channel

Consider a single link communication system in which the interference is additive and given by

$$n(t) = n_1(t) + n_2(t) \ , \qquad (1)$$

* Hereafter, this reference will be denoted by the symbol (I).

** The theory developed in this paper is valid for other distributions of the frequency of occurrence of the noise pulses.

where $n_1(t)$ is Gaussian with a known power spectrum, $\Phi_{n_1}(j\omega)$, and $n_2(t)$ is impulsive noise represented by

$$n_2(t) = \sum_k \eta_k \, \delta(t-t_k)* \qquad (2)$$

The amplitudes of the impulsive noise component have a probability density function $p(\eta_k)$ with zero and large variance, σ_η^2; the function $p(\eta_k)$ is otherwise unspecified. The number of noise impulses, n, in the duration T will be assumed to be governed by Poisson's distribution

$$P(n) = \frac{(\nu T)^n}{n!} e^{-\nu T} , \qquad (3)$$

where ν represents the long time average of the number of pulses occurring per unit time. In addition, $n_1(t)$ and $n_2(t)$ are considered to be statistically independent. The first transmission system discussed will be binary in which the two possible signals, $s_1(t)$ and $s_2(t)$, occur with equal a priori probability and are subject to an average power limitation. The error probability, P_e, will be minimized.

Consider the total signal duration, T, subdivided into u equal subintervals each of duration $\frac{T}{u}$. In each subinterval the waveform is identical. This signal structure is chosen because the impulsive interference, $n_2(t)$, is as likely to occur in one subinterval as any other. Thus, the waveform in each of the subintervals have the same probability of being destroyed by the impulsive noise. On the other hand, the waveforms should have equal capability for combatting the Gaussian noise.

Let the two signals per decision subinterval be $s_A(t)$ and $s_B(t)$ respectively. Let $p(y_j/s_A)$ be the conditional probability density function that, having sent $s_A(t)$ in the j^{th} subinterval, $y_j(t)$ will be received. $p(y_j/s_B)$ is defined similarly for the signal $s_B(t)$. The expressions for these conditional probability density functions are

$$p(y_j/s_A) = P_o \, p_o(y_j/s_A) + P_1 \, p_1(y_j/s_A) \qquad (4)$$

and

$$p(y_j/s_B) = P_o \, p_o(y_j/s_B) + P_1 \, p_1(y_j/s_B) , \qquad (5)$$

where P_o is the probability that the jth subinterval contains zero pulses of impulsive noise and P_1 is the probability that the jth subinterval contains one or more pulses of impulsive noise. Using equation (3)

$$P_o = e^{-\frac{\nu T}{u}} , \qquad (6)'$$

and

$$P_1 = (1-e^{-\frac{\nu T}{u}}) . \qquad (7)$$

The other symbols in equations (4) and (5) have the following meaning: $p_o(y_j/s_A)$ is the conditional probability density function that, having sent $s_A(t)$ in the jth subintervale, $y_j(t)$ will be received which came from the sum of $s_A(t)$ and Gaussian noise alone; $p_1(y_j/s_A)$ is the conditional probability density function that, having sent $s_A(t)$ in the jth subinterval, $y_j(t)$ will be received which came from the sum of $s_A(t)$, Gaussian and impulsive noises. The probability density functions $p_o(y_j/s_B)$ and $p_1(y_j/s_B)$ have the same meaning for signal $s_B(t)$ as $p_o(y_j/s_A)$ and $p_1(y_j/s_A)$ for signal $s_A(t)$. Equations (4) and (5) are a mathematical expression of the fact that $p_o(y_j/s_A)$, $p_1(y_j/s_A)$, $p_o(y_j/s_B)$, and $p_1(y_j/s_B)$ describe mutually exclusive events.

Since all the subintervals of the decision interval (0,T) are affected by Gaussian noise independently,* using the same reasoning as in (I), the decision rule becomes: accept $s_1(t)$ if

$$\prod_{j=1}^{u} \frac{p(s_A/y_j)}{p(s_B/y_j)} \geq 1 \qquad (8)$$

accept $s_2(t)$ otherwise. $p(s_A/y_j)$ is the a posteriori probability that, having received a signal $y_j(t)$ in the jth subinterval, it came from signal $s_A(t)$ and noise. $p(s_B/y_j)$ is defined similarly for the signal $s_B(t)$. As in (I), $s_A(t)$, $s_B(t)$, and $y_j(t)$ may be expressed by

$$s_A(t) = \sum_{k=1}^{N} a_{Ak} \Psi_k(t) , \qquad (9)$$

$$s_B(t) = \sum_{k=1}^{N} a_{Bk} \Psi_k(t) , \qquad (10)$$

and

$$y_j(t) - \sum_{k=1}^{N} y_{jk} \Psi_k(t) , \qquad (11)$$

* The unit impulses may be just narrow pulses.

* $\Phi_{n_1}(j\omega) = A^2 + B^2\omega^2$

where the set of functions $\{\psi_k(t)\}$ is the orthonormal set generated by the integral equation

$$\tau_k^2 \,\psi_k(t) = \int_0^{\frac{T}{u}} K(x-t)\,\psi_k(x)dx, \text{ for } 0 \le x \le \frac{T}{u}, \tag{12}$$

and K(x) is the inverse Fourier transform of the Gaussian noise power density spectrum, $\Phi_{n_1}(j\omega)$.

Following the procedure of (I), the decision rule (8) reduces to

$$\sum_{j=1}^{u} \frac{2ay_{j1}}{\tau_1^2} + \sum_{j=1}^{u} \log\left\{1 + \frac{1}{C}\exp\frac{1}{2}\,\frac{(y_{j1}-a)^2}{\tau_1^2}\right\}$$

$$-\sum_{j=1}^{u} \log\left\{1 + \frac{1}{C}\exp\frac{1}{2}\,\frac{(y_{j1}+a)^2}{\tau_1^2}\right\} \ge 0^{*}, \tag{13}$$

where

$$C = \frac{P_o}{P_1\tau_1\alpha\sqrt{2\pi}}, \tag{14}$$

$a^2 = S^2\frac{T}{u}$, and α is a small positive constant of the order $\frac{1}{\sigma_n}$, or we assume that in subintervals filled with impulsive noise, the distribution of the received y_{j1} is essentially uniform and is governed almost completely by the impulsive noise, namely, since the variance of the amplitudes of the impulsive noise is large,

$$p(y_j/s_A) \doteq p_1(y_j/s_B) \doteq p_1(0) = \alpha. \tag{15}$$

The waveforms for $s_A(t)$ and $s_B(t)$ are then

$$s_A(T) = -s_B(T) = a\,\psi_1(t), \tag{16}$$

where $\psi_1(t)$ is the eigenfunction of the integral equation (12) corresponding to the lowest eigenvalue τ_1^2.

For the case when no impulsive noise is present ($C \longrightarrow \infty$), the inequality (13) reduces to

$$\sum_{j=1}^{u} \frac{2ay_{j1}}{\tau_1^2} \ge 0, \tag{17}$$

which results in the same linear detector as in (I) with the optimum choice of u equal to one. On the other hand, if the impulsive noise is present, the linear detector must be replaced by a detector with a non-linear element, and the optimum choice of u is $u \ge 1$.

Consider the structure and behavior of the non-linear element of the detector. This element must have the transfer characteristic of the form

$$G(y_1) = \frac{2ay_1}{\tau_1^2} + \log\left\{1 + \frac{1}{C}\exp\left[\frac{1}{2}\,\frac{(y_1-a)^2}{\tau_1^2}\right]\right\}$$

$$- \log\left\{1 + \frac{1}{C}\exp\left[\frac{1}{2}\,\frac{(y_1+a)^2}{\tau_1^2}\right]\right\} \tag{18}$$

$$\text{for } y_1 \ge 0$$

and

$$G(-y_1) = -G(y_1) \quad \text{for } y_1 < 0 \tag{19}$$

The transfer characteristics, $G(y_1)$, are plotted in Figure 1 for a fixed value of τ_1^2 and two different values of C and a; the associated linear transfer characteristics ($C \longrightarrow \infty$) are also shown. The non-linear transfer characteristic approaches the linear characteristic for small values of cross-correlator output, y_1, and completely rejects very large cross-correlator outputs. The fall off of the transfer characteristic is much sharper for large values of C and small values of a (which corresponds, respectively to heavy impulsive noise and low signal power) and occurs for larger values of the cross-correlator output. In this case, the occurrence of impulsive noise in a subinterval corrupts it so completely that the detector effectively rejects the contribution of such a subinterval to the final decision because the uncertainty in this case is very high. **For large values of a (large signal power)** the region of fall off on the transfer characteristic starts earlier and is not as sharp as in the case of low values of a (low signal power). This is to be expected, because for a substantial range of values of the cross-correlator output, the signal is not sufficiently corrupted by the impulsive noise to be completely discarded.

The actual optimum detector for the single-bit case, with both the Gaussian and impulsive noises disturbing the communication link, will be essentially a cross-correlator operating over a time interval of duration $\frac{T}{u}$. The cross-correlator produces in its output y_{j1} which is the result of cross-correlating $y_j(t)$ with $s_A(t)$ for all j=1, 2, ... , u. The cross-correlator is followed by

* See reference [7] or [8] for details.

a non-linear element with the transfer characteristic given by equations (18) and (19). At the output of the non-linear element is a voltage storing device (for instance, a capacitor) which stores outputs of the non-linear element for all j=1, 2, ..., u. If at the end of the full decision interval (0,T), the stored quantity is positive or zero $s_1(t)$ is accepted; otherwise, $s_2(t)$ is accepted.

The Error Probability for the One Bit Case (Binary Communication)

If the variance of the amplitudes of the impulsive noise is large, as was assumed originally, the error probability per jth subinterval is

$$P_e^{(j)} = \frac{1}{2} P_1 + P_o P_{oe}^{(j)} , \qquad (20)$$

P_1 and P_o are given by equations (6) and (7), and $P_{oe}^{(j)}$ is the error probability if only Gaussian noise is present as found in (I). Although it is quite simple to express the error probability jth subinterval of duration $\frac{T}{u}$, it is difficult to find an expression for the error probability in a full decision interval (0,T) due to the complexity of the decision rule as expressed by the inequality (13). If we were able to find this expression, we would have to minimize it with respect to u to find the optimum scheme of detection. One can see that such an optimum value of u exists from the following physical considerations. The minimum permissible u is one; under these conditions P_e is given by equation (20) yielding a high value for the error probability, especially if ν is high. On the other hand, if we make u very large to have the non-linear element of the detector effectively reject the subintervals with impulsive noise and base the decision on the subintervals affected by Gaussian noise alone, the error probability will be high due to the high value of $P_{oe}^{(j)}$. Thus, between the two extreme values of u there must be an optimum value which will minimize the error probability.

The Ideal Discarding Detector for the One Bit Case, Its Associated Optimum Value of u, and the Minimum Error Probability

Since we are unable to find an expression for the error probability, we will replace, for the purpose of analysis, the actual detector with a non-linear element by an ideal discarding detector. The ideal detector would know exactly in which subinterval impulsive noise occurs, discard this subinterval completely and base its decision on the subintervals disturbed by the Gaussian noise only. Thus, for the ideal discarding detector, the decision will be based on (u-n) subintervals, where n is the number of subintervals discarded in a decision interval (0,T).

The error probability, P_e, for a communication link using the ideal discarding detector will be a good estimate of the error probability in an actual system, especially, if σ_γ^2 is large and τ_1^2 small. The optimum value of u as found for the ideal-discarding-detector model, will be used in the actual communication system.

Following the same procedure as in (I), the decision rule for the ideal discarding detector is

$$\sum_{j=1}^{u-n} \frac{2ay_{j1}}{\tau_1^2} \geq 0 \quad . \qquad (21)$$

The y_{j1}'s in equation (21) are independent and Gaussian with variances τ_1^2 and means a for the signal $s_1(t)$ and variances τ_1^2 and means -a for the signal $s_2(t)$. Using the procedure of (I), the error probability for n out of u subintervals discarded is

$$P_e(u-n) = \frac{1}{2} [1 - \Phi(\frac{\xi}{\sqrt{2}})] , \qquad (22)$$

where

$$\xi^2 = \frac{(u-n)\, S^2 T}{u\, \tau_1^2} , \qquad (23)$$

and $\Phi(x)$ is the tabulated error function

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz \quad . \qquad (24)$$

Since the occurrence of noise pulses is random in time and the frequency of their occurrence is governed by Poisson's distribution, the probability that (u-n) subintervals of the u available in any decision interval (0,T) will be empty (without impulsive noise) is governed by the binomial distribution*, namely,

$$P(u-n) = C_n^u \, e^{-\frac{\nu T}{u}(u-n)} [1 - e^{-\frac{\nu T}{u}}]^n . \qquad (25)$$

The error probability for a system using the ideal discarding detector in the one-bit case can be expressed by

$$P_e = \sum_{n=0}^{u} P_e(u-n)\, P(u-n) , \qquad (26)$$

*This is actually a case of Bernoulli trials[2].

where $P_e(u-n)$ and $P(u-n)$ are given by equations (22) and (25) respectively.

In general, one can find the optimum value of u and the corresponding error probability by calculating, using equation (26) for a given set of parameters and different values of u, the value for P_e. The value of u which will result in the minimum in the error probability will be accepted as optimum.

In the case of practical importance if the value of the separation function, ξ^2, is large, one can obtain a simplified expression for the error probability. If ξ^2 is large, equation (22) can be replaced by the asymptotic expression

$$P_e(u-n) = \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{\xi^2}{2}}}{\xi} . \qquad (27)$$

The numerator of equation (27) is a much faster varying function of ξ than the denominator; therefore, in the neighborhood of a minimum in the error probability in the denominator of equation (27) may be replaced by a constant, say ξ_o.

Substituting from equations (23), (25), and (27) in equation (26), we obtain

$$P_e = \frac{1}{\sqrt{2\pi}\,\xi_o} \sum_{n=o}^{u} C_n^u\, Q(u,n)$$

where $Q(u,n)$ =

$$\exp-(u-n)\left[\frac{\nu T}{u} + \frac{S^2 T}{2u\tau_1^2}\right] \left[1-\exp-\left(\frac{\nu T}{u}\right)\right]^n, \qquad (28)$$

Applying Newton's binomial expansion theorem to equation (28), the final expression for the error probability reduces to

$$P_e = \frac{1}{\sqrt{2\pi}\,\xi_o} \left\{1-e^{-\frac{\nu T}{u}} \left[1-e^{-\frac{S^2 T}{2u\tau_1^2}}\right]\right\}^{u*} . \qquad (29)$$

In the minimization process ξ_o is kept constant with the value of n in it replaced by the long time average of the number of noise pulses in the interval of time (O,T), namely,

$$\bar{n} \doteq \nu T^* \quad . \qquad (30)$$

In a practical situation one will prefer to plot P_e as a function of u for fixed values of the parameters and one will find from the graph the optimum value of u (the integer which corresponds to the smallest value of the error probability). Figs. 2 and 3 show the behavior of the error probability as a function of u for a system using the ideal discarding detector and for different sets of parameters ν, T, S^2 and two different Gaussian noise power spectra - rapidly rising with rising frequency (Fig. 2) and slowly rising with rising frequency (Fig. 3). As one can note from these figures, for small values of u the error probability is controlled by the impulsive noise and for large values of u by the Gaussian noise as predicted from physical considerations. Increasing the decision time (lowering the rate of transmission of information) is a much more effective method to combat the impulsive noise than an increase in the average power of the transmitter. In general, the choice of optimum u is critical (all the curves display a sharp minimum).

Discussion of Some Methods of Extension of the Theory Developed for the One-Bit Case to the m-Bit Case

The methods of extension of the theory developed for the one-bit case to the m-bit case will be dictated in part by the type of coding used if the impulsive noise were not present. As one recalls from (I), basically, there are two classes of coding one can use in connection with Gaussian noise disturbing the communication link: orthogonal digit coding and binary digit coding. These two classes of coding suggest two different methods of extension of the theory to the case when impulsive noise is present.

At first, consider the decision interval (O,T) divided into u equal subintervals of duration $\frac{T}{u}$. Let us code each subinterval, using orthogonal digit coding, so that each subinterval contains all the information of the signal and is coded best for no impulsive noise present**. The optimum receiver will essentially discard the subinterval with impulsive noise and base its decision on the subintervals containing the Gaussian noise and signal only. Since each subinterval contains all the information of the transmitted signal, one expects this method to yield reliable performance for large values of T. It is true that for large values of T there are more pulses of noise present in the decision interval but, on the other hand,

* The procedure for finding the optimum value of u to minimize this error probability is given in Appendix A of reference 7 or Appendix 2 of reference 8.

* Approximately equal sign in this case implies nearest integer.

**For instance, for a Gaussian noise spectrum increasing rapidly with increasing frequency, this will require minimax coding considering that the decision subinterval of duration $\frac{T}{u}$ is short [See (I)].

we can also use higher values of u to obtain optimum performance. Thus, decreasing the rate of transmission of information should improve considerably the reliability of the communication system using this method of coding.

Let us now consider the use of the equidistant code in the decision interval (0,T). In this interval there will be $u = (2^m-1)$ binary digits. Out of the u digits $\lambda = (2^{m-1})$ digits will differ from those of any other signal of the transmitted set $\{s_i(t)\}$, $i = 1, 2, \ldots, 2^m$. We will use the redundancy in binary digits to improve the performance of the system if the impulsive noise is present. The impulsive noise will cause some digits to be rejected, but since the redundancy is high, one can base the decision on the remaining digits. For a fixed value of m, u and λ are also fixed; thus, for small values of T the Gaussian noise will cause the error probability to become high ($\frac{T}{u}$ is small or $\tau_1{}^2$ is large). On the other hand, for large values of T the impulsive noise will cause a rise in the error probability because too many digits will be discarded. Therefore, one expects an optimum value for the decision interval (0,T) to exist, which will minimize the error probability for fixed values of parameters m, S^2, γ, and $\Phi_{n1}(j\omega)$. For large values of the decision interval (0,T) we will expect the system using full information per subinterval coding to perform better than the system using equidistant coding.

One can combine the positive qualities of both methods of coding to combat the impulsive noise for large values of the decision time by subdividing the interval (0,T) into r equal subintervals of duration $\frac{T}{r}$ and coding each subinterval using the equidistant code. In this case, the pulses of noise do not destroy the complete information contained in a given subinterval but only one or more digits per subinterval. Thus, one expects a system using the repeated equidistant coding to perform best because it seems that in this case the impulsive noise destroys the least amount of available information. One cannot make this statement definite because the equidistant code is the least efficient code if no impulsive noise is present, and its performance deteriorates rapidly with decrease in the decision time for increasing with increasing frequency Gaussian noise power spectra [see (I)].

Before proceeding with the mathematical formulation of the ideas presented in this section, let us mention the fact that the ideas applied to the equidistant code apply also to the nearly equi distant codes. For m=1 all the methods of coding reduce to the one-bit case discussed previously.

The Optimum Detector for the Method of Coding Using the Full Information per Subinterval and an Estimate of the Associated Error Probability

Let the set of signals $\{s_i(t)\}$, $0 \leq t \leq T$, $i = 1, 2, \ldots, 2^m$, be transmitted with equal a priori probability; the optimum receiver will accept the signal corresponding to the largest a posteriori probability $p(s_i/y)$. Consider each decision interval (0,T) subdivided into u equal duration subintervals. Let the signal per jth subinterval of signal $s_i(t)$ be $s_{Ai}(t)$. Using the same reasoning as in the one-bit case, the form of $s_{Ai}(t)$ must be best for transmission with no impulsive noise present. Since the duration of a subinterval, $\frac{T}{u}$, is usually small, this will dictate the use of the minimax code if the Gaussian noise power spectrum is increasing with increasing frequency*.

Modifying the procedure of (I), and using the notation of (I) and section 2, the decision rule reduces for any pair of signals $s_i(t)$ and $s_\mu(t)$ to: accept $s_i(t)$ if

$$\frac{\sum\limits_{j=1}^{u} \log\left\{\alpha P_1 + \frac{P_o}{(2\pi)^{\frac{m}{2}} \prod\limits_{k=1}^{m} \tau_k} \exp\left[-\frac{1}{2}\sum\limits_{k=1}^{m} \frac{(y_{jk}-a_{ik})^2}{\tau_k{}^2}\right]\right\}}{\sum\limits_{j=1}^{u} \log\left\{\alpha P_1 + \frac{P_o}{(2\pi)^{\frac{m}{2}} \prod\limits_{k=1}^{m} \tau_k} \exp\left[-\frac{1}{2}\sum\limits_{k=1}^{m} \frac{(y_{jk}-a_{\mu k})^2}{\tau_k{}^2}\right]\right\}} \geq 0 \tag{31}$$

accept $s_\mu(t)$ otherwise.

For the case of no impulsive noise present and u = 1 equation (31) reduces to the result of the Gaussian noise alone disturbing the communication link as obtained in (I). If the impulsive noise is present, the optimum receiver should be capable of performing the operations indicated by equation (31). A simple modification of the linear detector in the form of addition of a nonlinear element, as in the one-bit case, will not suffice. The quantities y_{jk} can be obtained at the receiver by simple cross-correlation between $y_j(t)$ and $\{\Psi_k(t)\}$, $k = 1, 2, \ldots, m$; all the other quantities in equation (31) are known at the receiver. If α is very small (large variance of the amplitudes of the impulsive noise), the optimum receiver approaches in performance a receiver using the ideal discarding detector.

As in the one-bit case, a receiver using the ideal discarding detector will be used to estimate the error probability and to find the value of u which will minimize this probability. For large values of the separation function the estimate of the error probability can be expressed as**

* If the Gaussian noise is white of $\frac{T}{u}$ is large, other methods of coding per subinterval may be used, but the theory will be similar.

**See Appendix B of reference 7 or Appendix 3 of reference 8.

$$P_e \doteq \frac{1}{\sqrt{2\pi}} \sum_{\lambda=1}^{m} \frac{\gamma_\lambda}{\xi_{o\lambda}} \left\{ 1-e^{-\frac{\nu T}{u}\left[1-e^{-\frac{\lambda S^2 T}{2u \sum_{k=1}^{m} \tau_k^{\,2}}}\right]} \right\}^{u^*}, \tag{32}$$

where

$$\xi_{o\lambda}^{\,2} = \frac{(u-\bar{n})\ \lambda S^2 T}{u \sum_{k=1}^{m} \tau_k^{\,2}}, \tag{33}$$

$$\text{for } \lambda = 1, 2, \ldots, m$$

and $\bar{n}$ is defined by equation (30). The optimum value of u will correspond to the value which will minimize the estimate of the error probability expressed by equation (32). The complexity of the optimum receiver will be a major reason for not using this method of coding in a practical situation.

The Optimum Detector for the Equidistant Code and an Estimate of the Associated Error Probability

As it was shown in (I), for the equidistant code with no impulsive noise present the optimum receiver must form

$$\beta_i = \sum_{k=1}^{u} y_k\, a_{ik} \tag{34}$$

for all $i = 1,2,3, \ldots, 2^m$ and accept signal $s_i(t)$ corresponding to the largest β_i. Thus, the decision is based on comparing various detected levels β_i. The decision rule will remain the same if all the β_i are increased by the same amount in each digit position. If we add to each digit position ay_k, the equation (34) will become

$$\beta_i' = \sum_{k=1}^{u} (a_{ik}+a)\ y_k \tag{35}$$

Equation (35) demonstrates that it is sufficient to sum y_k only for positive pulses of the received signal (there are λ such pulses in each signal). For the identity signal no detection voltage is needed, since if all β_i's are negative, the identity signal is accepted as the received signal.

The decision rule-accept the signal corresponding to the largest β_i'-is easy to instrument as was shown by S S.L. Chang, et al[1]. One can now modify this simplified detector to combat not only Gaussian but also impulsive noise. For each y_k of equation (35), instead of taking its full contribution to β_i', only its contribution, after being sent through a non-linear element with transfer characteristics of the form shown in Fig. 1, will be taken. Thus, the extension of the detection scheme when only Gaussian noise is present to the case in which both Gaussian and impulsive noises are present is similar to the scheme developed for the one-bit case.

Figure 4 shows a diagram of the essential components of a detector for a three-bit system. The (7,3) Slepian's group code used in this case is given in Table 1 of (I). The video signal is first passed through a non-linear device, the transfer characteristics being of the type shown in Figure 1. The output of the non-linear device is applied to the movable center arm of a seven-contact step switch which moves in synchronism with the signal pulses. The contacts of the switch are connected through the printing coils to a modified glow tube which has 8 symmetrical positions. The eighth position, corresponding to the identity signal, is connected through a printing relay to ground. The cathode of the glow tube is connected by means of a current limiting resistor, R_c, to a negative pulse source. Under normal conditions the cathode of the glow tube is near zero potential, and the eight anode voltages are small enough so the tube cannot fire. At the end of each decision interval (0,T), a negative pulse is applied to the cathode of the glow tube through the limiting resistor. The negative pulse permits conduction between the cathode and the anode of the glow tube having the greatest positive voltage with respect to ground. Thus, the proper printing relay is activated. Thereafter, all the integrators are discharged by means not shown and the detector is ready to receive the next signal. Since the interval of storage is relatively small, the integrator amplifiers need not meet high performance standards. A single stage amplifier will usually be adequate.

Again, as in the one-bit case, the ideal-discarding-detector model will be used to find an estimate of the error probability. For large values of the separation function, following a procedure similar to the one shown in Appendix B of reference 7, estimate of the error probability can be expressed as

$$P_e \doteq uP_D = \frac{u}{\sqrt{2\pi}\ \xi_{oD}} \left\{ 1-e^{-\frac{\nu T}{u}\left[1-e^{-\frac{S^2 T}{2u\tau_1^{\,2}}}\right]} \right\}^{\lambda}, \tag{36}$$

where $u = 2^m-1$, $\lambda = 2^{m-1}$,

* The notation in this equation is the same as for the minimax code in (I).

$$\xi_{oD}^{2} = \frac{(\lambda - \bar{n})S^{2}T}{u\tau_{1}^{2}} , \qquad (37)$$

and $\bar{n}$ is defined by equation (30).

For fixed values of m, S^2, ν, and a specified Gaussian noise power spectrum, $\Phi_{n1}(j\omega)$, the method of finding the optimum duration of the decision interval (0,T), which will minimize P_e of equation (36), is given in Appendix C of reference 7 or Appendix 4 of reference 8. It is simpler to plot P_e as a function of T and find from the plot the optimum value of T.

The Optimum Detector for the Repeated Equidistant Code and an Estimate of the Associated Error Probability

As was discussed previously, for large values of the decision time the equidistant code becomes inefficient. To improve the performance of the communication system, we can now use the repeated equidistant code.

The decision interval (0,T) is subdivided into r equal subintervals of duration $\frac{T}{r}$. Each subinterval is coded using the equidistant code. The same performance may be obtained if we code the whole decision interval (0,T) using the equidistant code but each binary digit of the code is repeated r times in the subinterval of duration $\frac{T}{u}$.

Following the same reasoning as for the one-bit case, for fixed values of parameters m, T, S^2, ν, and a specified Gaussian noise power spectrum, there must exist an optimum value of r which will minimize the error probability.

If the ideal-discarding-detector model is used to find an estimate of the error probability and if the error probability is small, equation (36) will be modified to yield

$$P_e \doteq \frac{u}{\sqrt{2\pi}\,\xi_{or}} \left\{ 1 - e^{-\frac{\nu T}{ur}} \left[1 - e^{-\frac{S^2 T}{2ur\rho_1^2}}\right] \right\}^{\lambda r} , \qquad (38)$$

where $u = (2^m - 1)$, $\lambda = (2^{m-1})$,

$$\xi_{or}^{2} = \frac{(\lambda r - \bar{n})\, S^{2}T}{ur\rho_{1}^{2}} , \qquad (39)$$

and ρ_1^2 is the lowest eigenvalue of the integral equation (12) with $\frac{T}{u}$ replaced by $\frac{T}{ur}$.

If the optimum value of r is large, it may be advisable to use smaller than optimum values of r. Large values of r (small values of $\frac{T}{ur}$) will cause a large increase in the error probability due to bandwidth limitation of the communication system*. Using less than optimum values of r will result in a smaller increase in the error probability due to the bandwidth limitation of the system, or a net decrease in the error probability may be achieved.

The actual detector for the repeated equidistant codes can be easily instrumented by a slight modification of the detector shown in Fig. 4. Each subinterval of duration $\frac{T}{r}$ is coded using the equidistant code. The detector operates for each subinterval as in the equidistant code case, only the negative pulse is now introduced at the cathode of the glow tube at the end of the full decision interval (0,T). The integrators are discharged at the end of the full decision interval (0,T).

Comparison of Performance Using Various Methods of Coding and Detection

An example of performance using various methods of coding is shown in Figs. 5 and 6. Both figures show plots of the minimum error probability of a system using an ideal discarding detector as a function of normalized decision time** for specified Gaussian noise power spectra, average transmitter power, number of bits transmitted, and average number of noise pulses per unit time.

The equidistant code curves display the expected minimum in the error probability for both white (Fig. 5) and increasing with increasing frequency (Fig. 6) Gaussian noise power spectrum. For the white Gaussian noise the curve of the error probability displays a large region of flatness in the neighborhood of the minimum.

A considerable improvement in the reliability of the system can be achieved by decreasing the rate of transmission of information (large values of T) for both types of noise power spectra and full information per subinterval or repeated equidistant coding. For the white Gaussian noise spectrum, the performance of the system using the repeated equidistant code is best from the point of view of reliability and simplicity of the associated optimum detector. No definite statement can be made about the reliability of a system disturbed by the Gaussian noise described by an increasing with increasing frequency power spectrum. For moderate rates of transmission of information, the repeated equidistant code seems to perform best. From a practical point of view, one would use the repeated equidistant code considering the relative simplicity of the associated optimum detector.

For further comparison, the curves of the Slepian's[4] best (6,2) and (10,3) group codes with digit-by-digit detection are also shown. The per digit error is found using equation (20). It is interesting to note that these curves display a

* See (I).

**Note that the information rate is m/T.

minimum in the error probability. The occurrence of the minimum is to be expected because for small values of the decision time too many digits are in error due to the additive Gaussian noise and for large values of the decision time too many digits are in error due to the impulsive noise.

In general, the performance of Slepian's best group codes with digit-by-digit detection is much worse than for the other schemes of coding and detection. This result is explained by the fact that a large amount of information is destroyed in the process of making a digit-by-digit decision instead of a decision per code group.

Acknowledgements

The author wishes to thank Dr. S.S.L. Chang for his suggestions and guidance.

The research reported in this paper was performed at New York University and was supported in part by the National Science Foundation Science Faculty Fellowship and in part by the Communication Theory Branch, AFCRL, Office of Aerospace Research (USAF), Laurence G. Hanscom Field, Bedford, Massachusetts under contract No. AF 19(604)-6168. It is based on parts of a thesis accepted by the faculty of the Graduate Division of the College of Engineering of New York University in partial fulfillment of the requirements for the degree of Doctor of Engineering Science.

References

1. Chang, S.S.L., Harris, B., Hauptschein,A., Hoffman, D., Morgan, K.C., Schwartz, L.S., "Evaluation and Optimization of Digital Communication Systems", New York University, Fourth Scientific Report, Contract No. AF 19 (604)-1964, January (1958).

2. Feller, W., "Probability Theory and Its Applications", I, pp. 104-106, New York, John Wiley, 1950.

3. Metzner, J.J., "Binary Relay Communication and Decision Feedback", IRE National Convention Record, Part 4, pp. 112-122, (1959).

4. Slepian, D., "A Class of Binary Signalling Alphabets", Bell Syst. Tech. J., 35, pp. 203-234, January, (1956).

5. Woodward, P.M., "Probability and Information Theory with Applications to Radar", New York, McGraw-Hill Book Co., Inc., 1953.

6. Kurz, L., "An Optimization Procedure for a Single-Link Unidirectional Digital Communication System in the Presence of Additive Gaussian Noise and for Detection Independent of Fading", New York University, Third Scientific Report, Contract No. AF 19(604)-6168, September, (1960).

7. Kurz, L., "An Optimization Procedure for a Single-Link Unidirectional Digital Communication System in the Presence of Additive Gaussian and Impulsive Noise and for Detection Independent of Fading", New York University, Fourth Scientific Report, Contract No. AF 19 (604)-6168, September, (1960).

8. Kurz, L., "An Optimization Procedure for a Single Link Unidirectional Digital Communication System in the Presence of Additive Gaussian and Impulsive Noise and for Detection Independent of Fading", An Eng. Sc. D. Thesis, College of Engineering, New York University, Spring, (1961).

9. Kurz, L., "A Method of Digital Signalling in the Presence of Additive Gaussian Noise", IRE Transactions on Information Theory, vol. IT-7, pp. 215-223, October, (1961).

10. Mertz, P., "A Model of Impulsive Noise for Data Transmission", IRE International Convention Record, Part 5, pp. 247-260, March 21-24, (1960).

11. Mertz, P., "Model of Error Burst Structure in Data Transmission" NEC, Data Transmission Session T123, pp. 75-84, October 10-12, (1960).

12. Yudkin, H.L., "Some Results in Measurement of Impulse Noise on Several Telephone Circuits", NEC, Data Transmission Session, T 123, pp. 65-75, October 10-12, (1960).

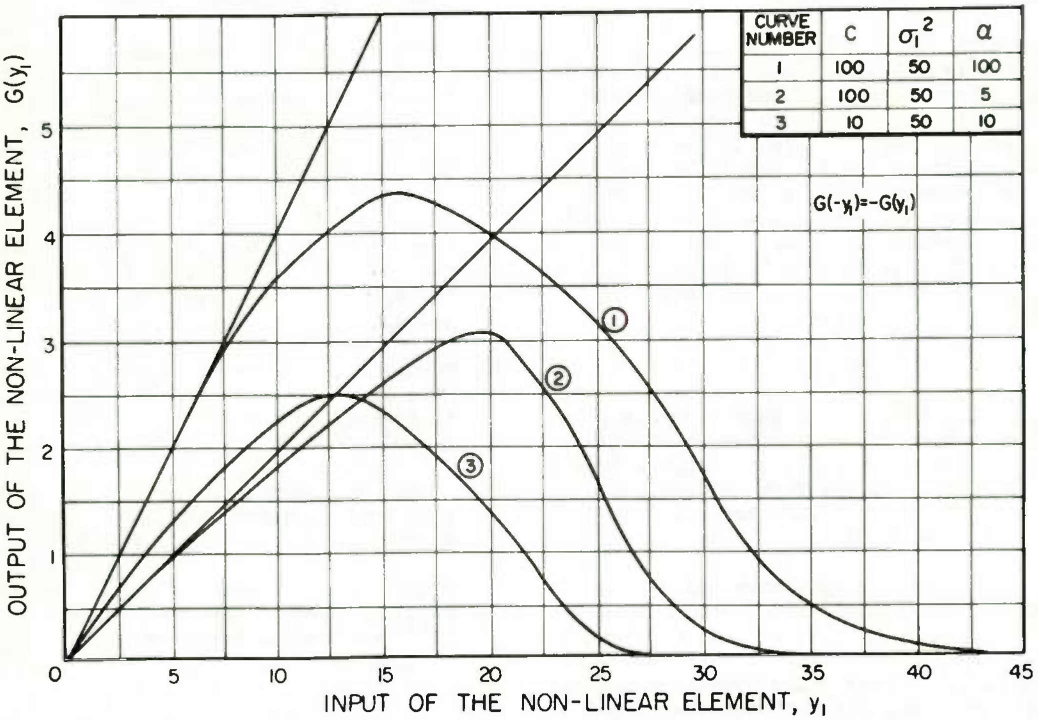


Fig. 1

Transfer Characteristics Of The Non-Linear Element Of The Optimum Detector For The One Bit Case

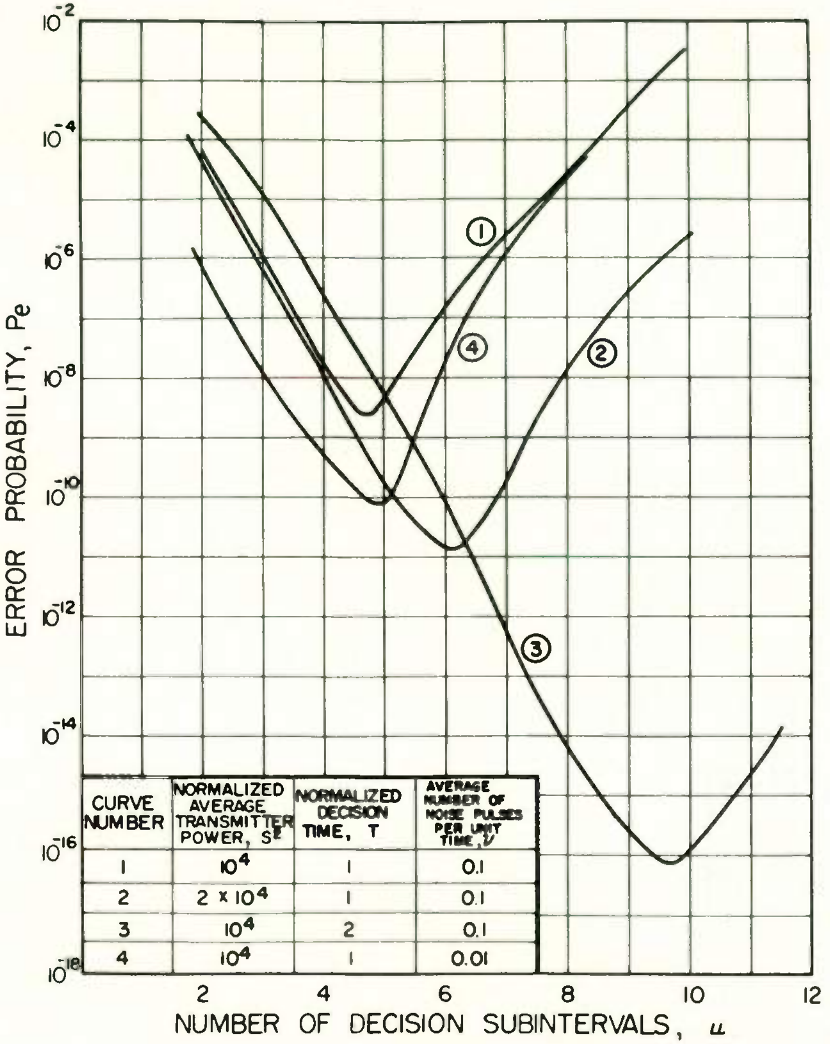


Fig. 2

The Error Probability As A Function Of The Number Of Decision Subintervals For A Communication Link Transmitting One Bit Of Information And Using An Ideal Discarding Detector. The Gaussian Noise Power Spectrum Is Rapidly Rising With Frequency, $\Phi_{nl}(j\omega) = \omega^2 + 1$.

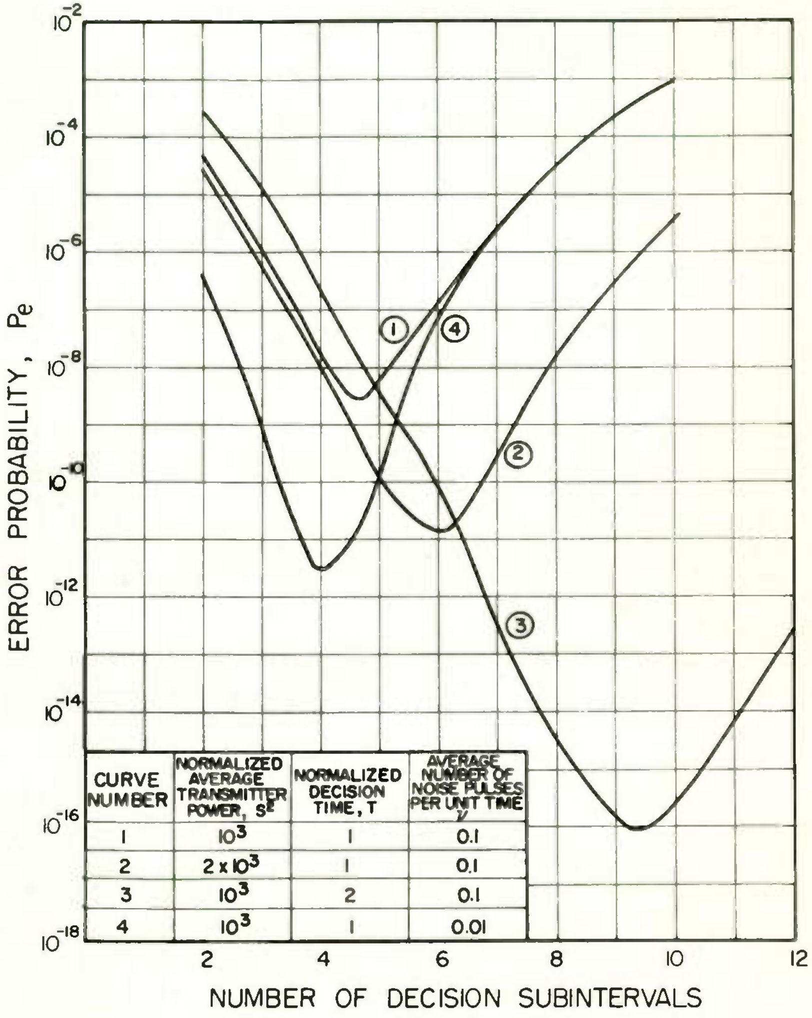


Fig. 3

The Error Probability As A Function Of The Number Of Decision Subintervals For A Communication Link Transmitting One Bit Of Information And Using An Ideal Discarding Detector. The Gaussian Noise Power Spectrum Is Slowly Rising With Frequency, $\Phi_{nl}(j\omega) = 0.1\omega^2 + 1$.

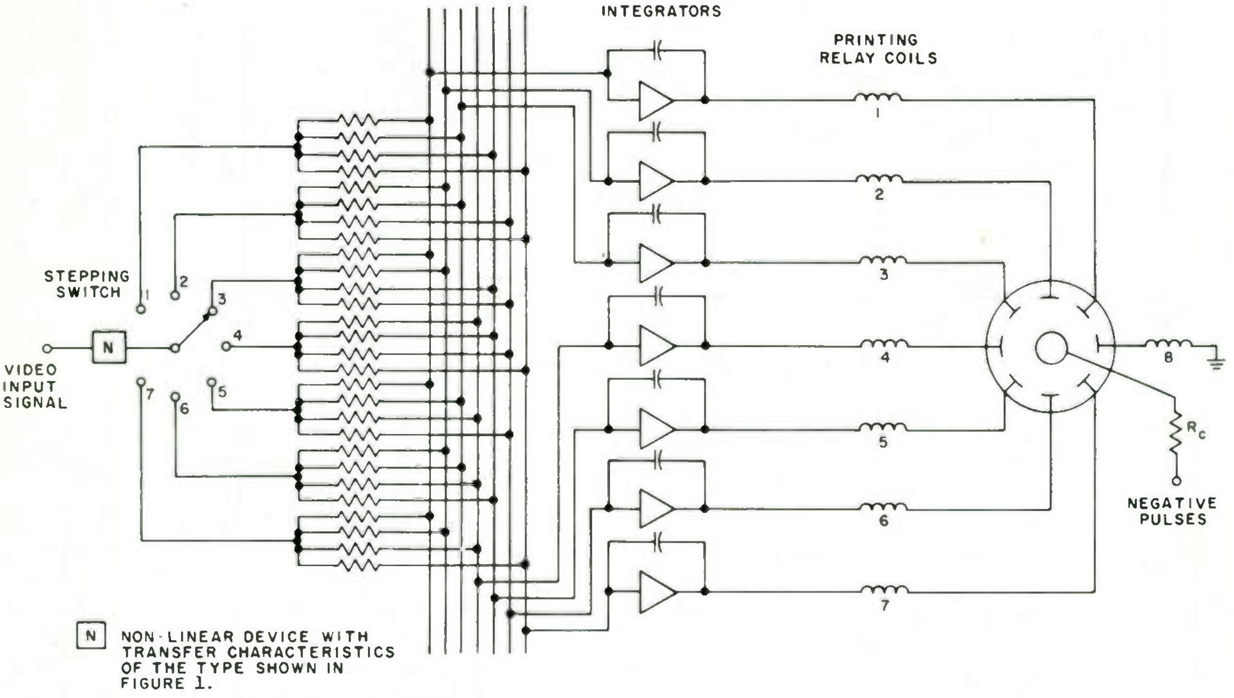
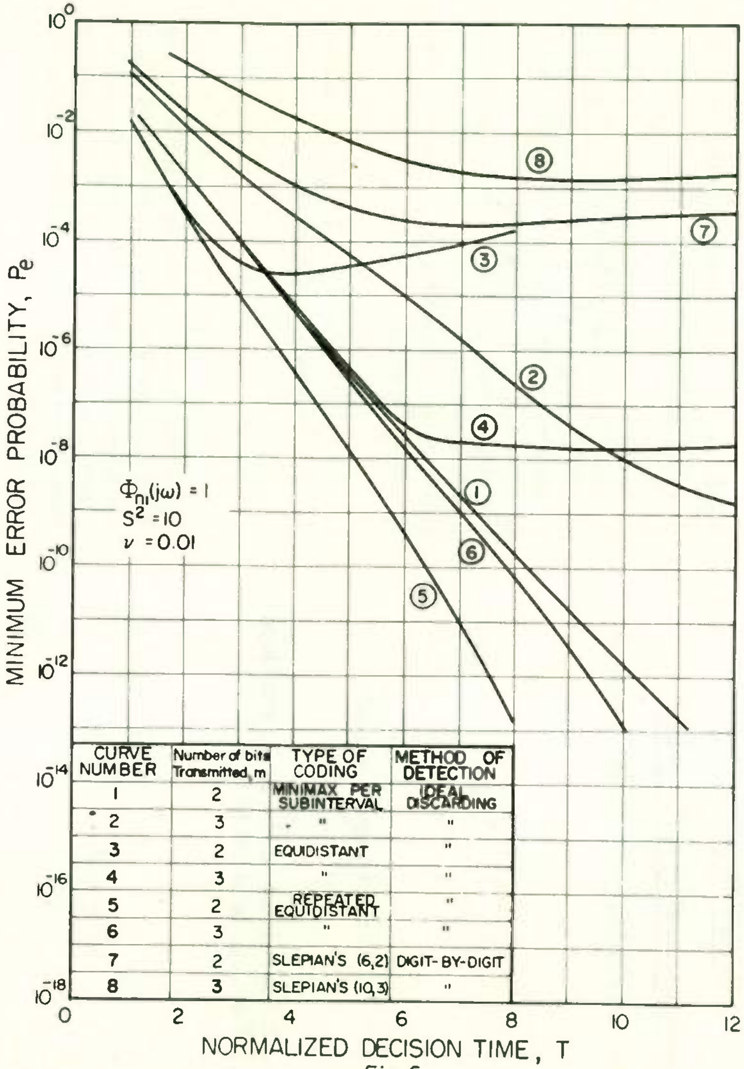


FIGURE 4.

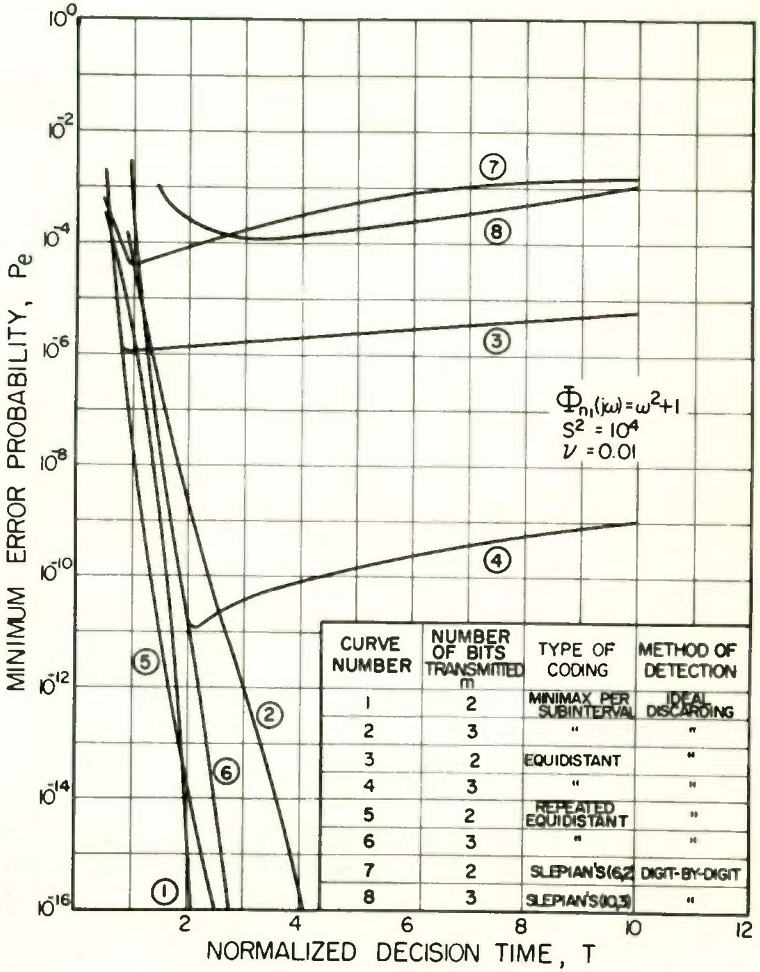
THE ESSENTIAL WIRING DIAGRAM FOR THE DETECTOR



CURVE NUMBER	Number of bits Transmitted, m	TYPE OF CODING	METHOD OF DETECTION
1	2	MINIMAX PER SUBINTERVAL	IDEAL DISCARDING
2	3	"	"
3	2	EQUIDISTANT	"
4	3	"	"
5	2	REPEATED EQUIDISTANT	"
6	3	"	"
7	2	SLEPIAN'S (6,2)	DIGIT-BY-DIGIT
8	3	SLEPIAN'S (10,3)	"

Fig. 5

Minimum Error Probability As A Function Of The Normalized Decision Time



CURVE NUMBER	NUMBER OF BITS TRANSMITTED m	TYPE OF CODING	METHOD OF DETECTION
1	2	MINIMAX PER SUBINTERVAL	IDEAL DISCARDING
2	3	"	"
3	2	EQUIDISTANT	"
4	3	"	"
5	2	REPEATED EQUIDISTANT	"
6	3	"	"
7	2	SLEPIAN'S(6,2	DIGIT-BY-DIGIT
8	3	SLEPIAN'S(10,3)	"

Fig. 6

Minimum Error Probability As A Function Of The Normalized Decision Time

EFFECTS OF PHASE-MODULATION ERRORS ON PULSE COMPRESSION SIGNALS

Charles E. Cook
Air Armament Division
Sperry Gyroscope Company
Division of Sperry Rand Corporation
Great Neck, L.I., New York

Introduction

Paired-echo theory has been used to describe the effects of network and time-function distortions both amplitude, and phase) of pulsed signals. [1-3] The advent of coded-waveform techniques (e.g., pulse compression) that generate waveforms having sidelobes extending in time has led to the application of paired echo theory in the establishment of design criteria for modern radar techniques. [4-6] This paper introduces somewhat more general conditions into the paired echo analysis as a basis for conducting experimental work on the effects of phase-error modulation of the time signal in a pulse compression system. The first phase of the experimental program provides a quantitative confirmation of paired echo theory, and the second phase yields waveform and spectrum data for the added conditions assumed in the analysis.

A General Paired Echo Analysis for Phase Errors Only

Frequency-Domain (i.e. Filter) Distortion

A complex video signal may be expressed as:

$$f_1(t) = \frac{1}{2\pi} \int_{-\omega_1}^{\omega_1} S(\omega) e^{j[\omega t + \theta(\omega)]} d\omega \tag{1}$$

where $S(\omega)e^{j\theta(\omega)}$ is the complex signal spectrum.

An i-f (or bandpass) signal of the same envelope shape, centered at ω_o, is given by:

$$f_2(t) = \frac{1}{2\pi} \int_{\omega_o - \omega_1}^{\omega o + \omega_1} S(\omega - \omega_o) e^{j[\omega t + \theta(\omega - \omega_o)]} d\omega \tag{2}$$

If the Fourier translation theorem is applied:

$$f_2(t) = \frac{1}{2\pi} e^{+j\omega_o t} \int_{-\omega_1}^{\omega_1} S(\omega) e^{j[\omega t + \theta(\omega)]} d\omega \tag{3}$$

or

$$f_2(t) = f_1(t)_e{}^{+j\omega_o t} \tag{4}$$

When the real-time function is considered, then:

$$f_2(t) = f_1(t) \cos \omega_o t, \tag{5}$$

which is the usual representation of a signal having a carrier frequency ω_o.

In the standard form of paired echo analysis, it is assumed that either $f_1(t)$ or $f_2(t)$ are transmitted through a filter having a certain form of distortion error. The filter amplitude distortion is assumed to be a periodic, even function, and the phase distortion is assumed to be a periodic, odd function of the same frequency. Thus, the filter transmission function in complete form is assumed to be:

$$H(\omega) = [a_o + a_1 \cos c\omega] e^{j[b_o\omega + b_1 \sin c\omega]} \tag{6}$$

In the type of all-pass time delay networks used for pusle compression filters, the error functions need not be exclusively odd or even. Letting a_o=1 and a_1=0 for ease of analysis, a more general phase-error term for the bandpass case can be written:

$$H(\omega) = e^{j[\, b_o(\omega - \omega_o) + b_1 \sin\{ c(\omega - \omega_o) + \phi \}]} \tag{7}$$

where ϕ is an arbitrary phase constant (Fig. 1).

When the phase expression given above is inserted into the Fourier integral for the i-f signal, the following expression is obtained:

$$f_2(t) = \frac{1}{2\pi}\, e^{j\omega_o t} \int_{-\omega_1}^{\omega_1} S_1(\omega) e^{j\,[\omega(t + b_o) + \theta(\omega) + b_1 \sin(c\omega + \phi)]}\, d\omega \tag{8}$$

Making use of the relationship:

$$e^{jb_1 \sin x} = J_o(b_1) + \sum_{n=1}^{\infty} J_n(b_1) e^{jnx} + (-1)^n \sum_{n=1}^{\infty} J_n(b_1) e^{-jnx} \tag{9}$$

(8) becomes:

$$\begin{aligned} f_2(t) = \frac{1}{2\pi} \Bigg[& e^{j\omega_o t} J_o(b_1) \int_{-\omega_1}^{\omega_1} S_1(\omega) e^{j[\,(t + b_o)\omega + \theta(\omega)]}\, d\omega \\ & + e^{j\omega_o t} \sum_{n=1}^{\infty} J_n(b_1) e^{jn\phi} \int_{-\omega_1}^{\omega_1} S_1(\omega) e^{j\,[t + b_o + nc)\omega + \theta(\omega)]}\, d\omega \\ & + (-1)^n e^{j\omega_o t} \sum_{n=1}^{\infty} J_n(b_1) e^{-jn\phi} \int_{-\omega_1}^{\omega_1} S_1(\omega) e^{j\,[(t + b_o - nc)\omega + \theta(\omega)]}\, d\omega \Bigg] \end{aligned} \tag{10}$$

and the real time function is:

$$f_2(t) = J_o(b_1)f_1(t+b_o)\cos\omega_o t$$

$$+\sum_{n=1}^{\infty} J_n(b_1)f_1(t+b_o+nc)\cos(\omega_o t+n\phi) \qquad (11)$$

$$+(-1)^n\sum_{n=1}^{\infty} J_n(b_1)f_1(t+b_o-nc)\cos(\omega_o t-n\phi)$$

The two summation terms in eq (11) represent the paired echo signals. It is seen that the result of adding the arbitrary phase ϕ, which essentially positions the error term relative to the band center, is to change the relative phase of each of the nth set of paired echoes by $n\phi$ and $-n\phi$ respectively, taking the phase of the carrier of the $J_o(b_1)$ term as a reference.

Time-Function Distortion

A general distortion expression for a time signal is:

$$E(t)=(a_o+a_1\cos c_m t)e^{j[\omega_o t+b_1\sin c_m t]},\ -\frac{T}{2}\le t\le\frac{T}{2} \qquad (12)$$

If E(t) is the output of a hard-limiting transmitter it is realistic to assume that $a_o=1$ and $a_1=0$. Introducing the arbitrary phase constant as before results in the time function:

$$E(t)=e^{j[\omega_o t+b_1\sin(c_m t+\phi)]},\ -\frac{T}{2}\le t\le\frac{T}{2} \qquad (13)$$

The use of eq (9) again leads to:

$$E(t)=J_o(b_1)e^{j\omega_o t}+\sum J_n(b_1)e^{j[(\omega_o+nc_m)t+n\phi]} \qquad (14)$$

$$+(-1)^n\sum J_n(b_1)e^{j[(\omega_o-nc_m)t-n\phi]},\ -\frac{T}{2}\le t\le\frac{T}{2}$$

The effect of the phase-modulation error is to produce the well-known paired sidebands centered at $(\omega_o \pm nc_m)$. The effect of the positioning term ϕ is to establish the relative phases of these carrier frequencies.

In a normal pulsed signal the paired sideband signals occur at the same interval in time, and are generally observed as ringing phenomena and other pulse-shape distortions.

In the case of linear FM pulse compression it has been shown [4,7-9] that, to a first approximation, a shift in frequency at the compression-filter input by an amount ω_d causes a relative delay at the filter output of $(\omega_d/\Delta\omega)T$, where $\Delta\omega$ is the linearly swept bandwidth and T is the input signal duration.[4,7-9] Thus, if the signal expression of eq (14) is a modulation-distorted, pulse compression signal, the action of the compression filter on the modulation sidebands will be to separate their respective waveforms by time increments $(nc_m/\Delta\omega)/T$ referred to the nominal pulse output. Under this condition the filter-output time function can be expressed:

$$E_o(t) = J_o(b_1)\bar{E}(t)\cos\omega_o t$$

$$+\sum k_n J_n(b_1)\Big\{\bar{E}(t)+\frac{nc_m}{\Delta\omega}T)\cos[(\omega_o+nc_m)t+n\phi] \qquad (15)$$

$$+(-1)^n\bar{E}(t-\frac{nc_m}{\Delta\omega}T)\cos[(\omega_o-nc_m)t-n\phi]\Big\}$$

where

$\Delta\omega$ = swept bandwidth

T = uncompressed pulse width

$\bar{E}(t)$ = compressed pulse envelope

k_n = Weighting factor of passband frequency response on nth pair of sidebands

This equation reprsents a good approximation when the phase-error induced sidebands of interest are mainly within the band limits of the weighting network. A more exact expression would have to consider the waveform-distortion effects of both the weighting network and the realizeable approximation of the matched filter.

The significance of the phase angles ± $n\phi$ in eq (11) and eq (14) is demonstrated in the results of the second phase of the experimental program, and is discussed in greater detail in that section of this paper.

Experimental Program

Phase I - Quantitative Verification of Paired Echo Theory

Although the paired echo analysis approach has not been seriously questioned, quantitative supporting data have been difficult to obtain in normal applications for the following reasons:

1. Network transmission functions cannot be controlled precisely enough to obtain a meaningful experiment.
2. Data reduction for phase modulation errors by means of a spectrum analyzer is not feasible because the sideband spectra usually overlap, rendering accurate measurement difficult.

The application of pulse compression as a laboratory analysis technique provides a method for overcoming the difficulties cited above. Equation (15) shows that a small displacement of paired sidebands from center frequency ω_o can be translated into a relatively large time displacement of the waveforms if the parameter T is large enough. Figure 2 blocks out in basic detail an experimental setup designed to obtain quantitative data on the paired echo, or paired sideband, phenomenon. A critical test of the theory would be to determine, for sinusoidal-phase-error modulation of the pulse compression input signal, if the various output time functions of (15) can be made to go to zero in conformance with the values of b_1 that cause a null in the Bessel functions $J_n(b_1)$. An additional test would be to determine how closely the amplitudes of the paired signals follow the Bessel-function curves, although, as n increases, so does the effect of the weighting factor k_n.

In the program conducted the values of b_1 ranged over limits far exceeding that acceptable for a system performance requirement. The pulse compression parameters used were:

Input Pulse Width	=	30 μsec
Swept Bandwidth	=	1.6 mc
Error-Modulating Frequencies $\frac{c_m}{2\pi}$	=	100 kc 200 kc

The value of b_1 was computed from the relationship $\Delta f = (c_m/2\pi)\ b_1 \cos c_m t$, where Δf was extablished from the voltage-vs.-frequency characteristic of the voltage-controlled oscillator and represents the peak frequency deviation caused by the error-modulation voltage.

In Fig. 3a, the sweep voltages applied to the voltage-controlled oscillator are compared, and in Fig. 3b the frequency-modulated waveforms with and without the error modulation added are compared.

Figure 4 shows the pulse-compression-filter outputs for various values of b_1. Note the emergence of the J_2 term in the last waveform. Figure 5 illustrates, on an expanded scale, the suppression of the compressed pulse at its nominal location, this being the case of $J_o(b_1) = 0$. Figure 6 shows the signal spectrum for this condition.

The variations of the measured $J_o(b_1)$ and $J_1(b_1)$ terms are compared to the theoretical values in Figure 7. The measured and theoretical values of b_1 that cause nulls in the Bessel functions up to n = 3 are given in Table I. The results obtained in this phase of the experimental program agree accurately with those predicted by theory, thus establishing the validity of this experimental technique. It will be noted that the percentage error of the measured value of b_1 increases for both higher orders of n and the higher order null signals. This can be attributed to driving the voltage-controlled oscillator into the region where it no longer maintains a linear frequency-vs.-voltage characteristic.

Phase II - Spectrum and Waveform Data

In phase II of the experimental program, the technique outlined in Section II was used to obtain data on the effects of time-function phase-error modulation on pulse compression waveforms and spectra. In addition, data concerning the effect of the error-modulation positioning factor ϕ on the compressed-pulse, paired-echo signals also was obtained.

The signal introduced into the pulse compression filter is:

$$E(t) = \cos\left[\omega_o t + \frac{1}{2}\mu t^2 + b_1 \sin(c_m t + \phi)\right], \quad -\frac{T}{2} \leq t \leq \frac{T}{2} \qquad (16)$$

As decribed previously, the action of the phase-modulation-error term $b_1 \sin(c_m t + \phi)$ is to produce paired sidebands. The signals associated with these sidebands are separated out in time by the compression filter to produce a paired-echo type signal, and this is how the error signals will be designated.

The variables in the experiments are:

b_1 - peak phase modulation, in radians

c_m - error-modulation rate, radian frequency

ϕ - starting phase, or positioning factor, of the modulation frequency, in radians

The error-modulation rate can be referred to a critical modulation rate by the following postulate:

> The critical modulation rate is that modulation frequency at which the first paired echo signals are distinctly separated from the main signal for small values of b_1 (say $b_1 \leq 0.5$ radian).

Using the above as a standard, and letting $\bar{c}_m$ denote the critical modulation frequency, the time shift caused by $\bar{c}_m$ is:

$$\bar{t} = \frac{c_m}{\Delta\omega} T \tag{17}$$

Since this is a generalized time shift dependent on the compressed-pulse width, eq (17) can be written in an alternate form, noting that $2\pi \bar{f}_m = \bar{c}_m$:

$$\frac{\bar{f}_m}{\Delta f} T = \frac{n}{\Delta f} = \bar{t} \tag{18}$$

where n will be some number, probably between 1 and 3, that is dependent partly on the compressed-pulse-waveform shape, particularly when sidelobe reduction results in less-well-defined skirts on the waveform. From eq (18) a value of $\bar{f}_m$ is obtained:

$$\bar{f}_m = \frac{n}{T} \tag{19}$$

Thus, the critical modulation rate will depend only on the uncompressed pulse width T and not on any of the other pulse compression factors such as the compression ratio or $\Delta\omega$. This is not surprising when it is realized that changes in $\Delta\omega$ result in compensating changes in the compression-filter-delay characteristic that, for a fixed f_m and T, will maintain the paired echoes in the same <u>relative</u> position to the main signal. In essence, then, the critical factor is not the actual value of f_m, but is the number of cycles of phase-error modulation that occur during the uncompressed pulse interval T. This number is independent of any of the explicit parameters of the pulse compression signal, and thus the values of $\bar{f}_m$ used in any experimental setup should yield results that, in general, hold for all ranges of linear FM sweep parameters.

The experimental data are compiled as a series of waveform photographs for the various conditions imposed. Figure 8 illustrates the effect of sinusoidal phase modulation in creating the paired echoes associated with pulse compression signals. The spectra and waveforms for values of b_1 between 0 and 1.0 radians are shown. The basic conditions of this experiment were:

T	=	30 μsec
f_m	=	6 cycles/30 μsec (200 kc)
ϕ	=	0
compression ratio	=	50, before sidelobe reduction
sidelobe level	=	-25 db

Figures 9 and 10 show the effect of the positioning factor ϕ on the paired echo waveforms. The added variables were:

ϕ = -90, -45, 0, 45 and 90 degrees (eq 13)

f_m = 2 and 3 cycles in the 30-microsecond pulse interval

The waveforms clearly show that the factor ϕ has a noticeable effect when f_m is of such a value that the first pair of echoes overlap the main signal (and even each other). Since the net result ϕ is to change the phase of the carriers of the various paired echo components, as shown in eq (15), then, when these components overlap, a pattern of waveform interference and reinforcement is set up that depends on the value ϕ assumed. Thus the mirror-waveform symmetry of a function of ϕ (as shown in Fig. 9) generally can be expected when the component time waveforms overlap. When f_m increases and the paired echo components are clearly separated, then the effect of ϕ is negligible. The critical-modulation cases for the waveform parameters used in these experiments is shown in Fig. 10. For these experiments, f_m was set equal to 3 cycles/wide-pulse width, or 100 kc in this instance.

The results shown in Fig. 9 illustrate the effect of the time phase, ϕ, in creating waveforms having mirror symmetry when ϕ is changed from a leading value to a lagging value of the same magnitude. Since these waveforms represent an addition of i-f signals, a symmetrical set of waveforms such as shown are obtainable only at a specific center frequency. Figure 11 illustrates that successive shifts in the center frequency of 1, 2, and 3 per cent of the bandwidth (i.e., 1.6 kc, 3.2 kc, and 4.8 kc) distort the symmetry relationship.

Conclusion and Acknowledgment

A method of performing paired echo experiments with pulse compression signals has been demonstrated to yield results conforming to theory. When the timing phase ϕ of the modulation-error signal is varied, the composite paired echo time signal is effected if the modulation is such that it causes the paired echoes to overlap the main signal.

The efforts of Mr. J. N. Cerar in devising the experimental techniques and in obtaining the data are appreciated.

References

1. H.A. Wheeler, "The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes," Proc. IRE, Vol. 27, June 1939.

2. C.R. Burrows, "Discussion on Paired Echo Distortion Analysis," Proc. IRE, Vol. 27, June 1939.

3. S. Goldman, Frequency Analysis, Modulation and Noise, McGraw-Hill, New York, 1948.

4. J.R. Klauder, A.C. Price, S. Darlington, and W.J. Albersheim, "The Theory and Design of Chirp Radars," BSTJ, Vol. 39, July 1960.

5. J.V. DiFranco and W.L. Rubin, "Distortion Analysis of Radar Systems," Proc. 7th Annual East Coast Conf. on Aeronautical and Navigational Electronics, October 1960.

6. P.M. Liebman, Unpublished Notes on Pulse Compression Distortion.

7. C.E.Cook, "Pulse Compression - Key to More Efficient Radar Transmission," Proc. IRE, Vol. 48, March 1960.

8. C.E. Cook, "General Matched Filter Analysis of Linear FM Pulse Compression," Proc. IRE, Vol. 49, April 1961.

9. C.E. Cook, "Linear FM Pulse Compression," Chapter 14, Modern Radar Techniques, R.S. Berkowitz, editor, John Wily and Sons, New York, to be published.

TABLE I

VALUES OF b_1 FOR $J_n(b_1) = 0$

Bessel Function	First Null $-b_1$		Second Null $-b_1$		Third Null $-b_1$	
Order, n	Calculated	Measured	Calculated	Measured	Calculated	Measured
0	2.405	2.39	5.520	5.40	8.654	8.83
1	3.832	3.87	7.016	7.44	10.173	10.41
2	5.136	5.36	8.417	8.93	11.62	12.5
3	6.38	6.80	9.761	10.41	13.015	14.10

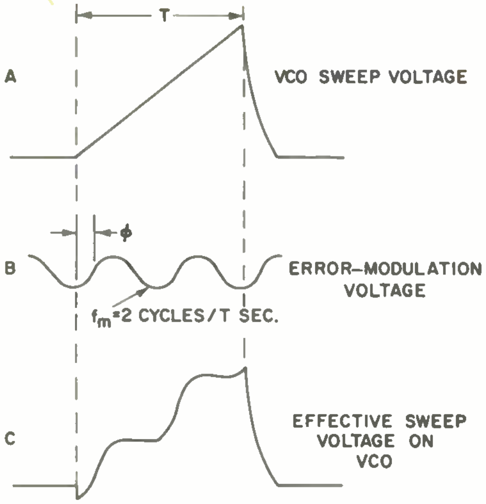
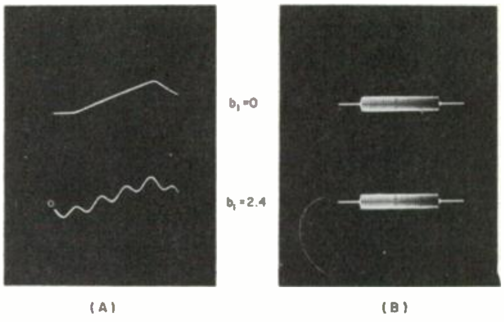


FIG. I SWEEP WAVEFORM COMBINATION



(A) SWEEP VOLTAGES APPLIED TO VCO

(B) OUTPUTS OF VCO

FIG. 3 NORMAL AND ERROR-MODULATED SIGNALS OF VOLTAGE-CONTROLLED OSCILLATOR FOR f_m = 3 CYCLES/INPUT PULSE WIDTH

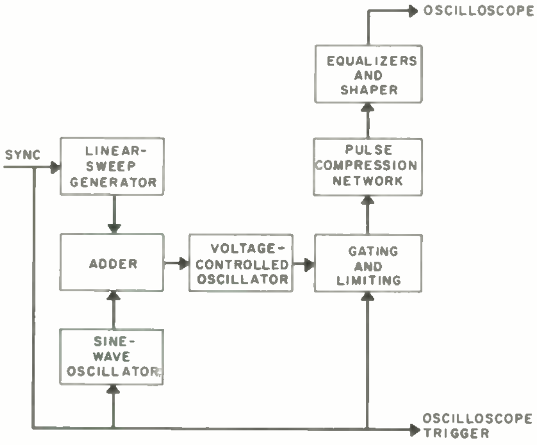


FIG. 2 BLOCK DIAGRAM OF PAIRED ECHO EXPERIMENT

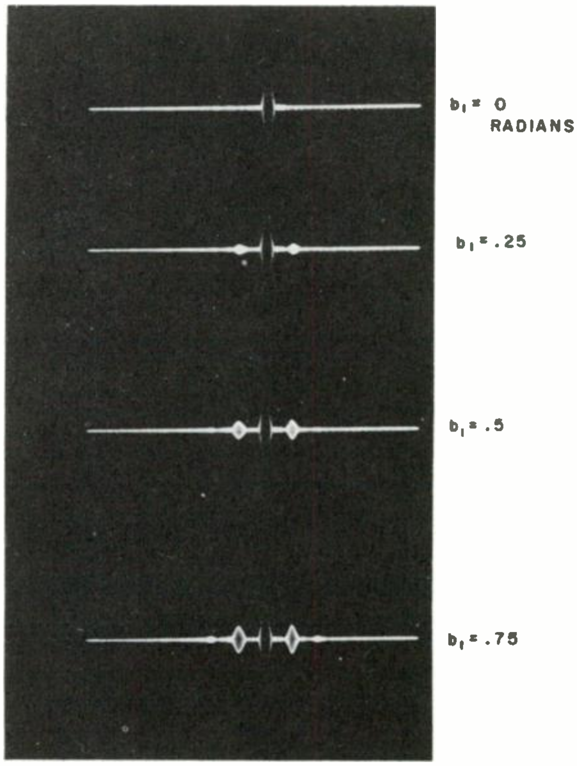


FIG. 4 COMPRESSION-OUTPUT WAVEFORMS AS A FUNCTION OF b_1 f_m = 6 CYCLES/INPUT PULSE WIDTH

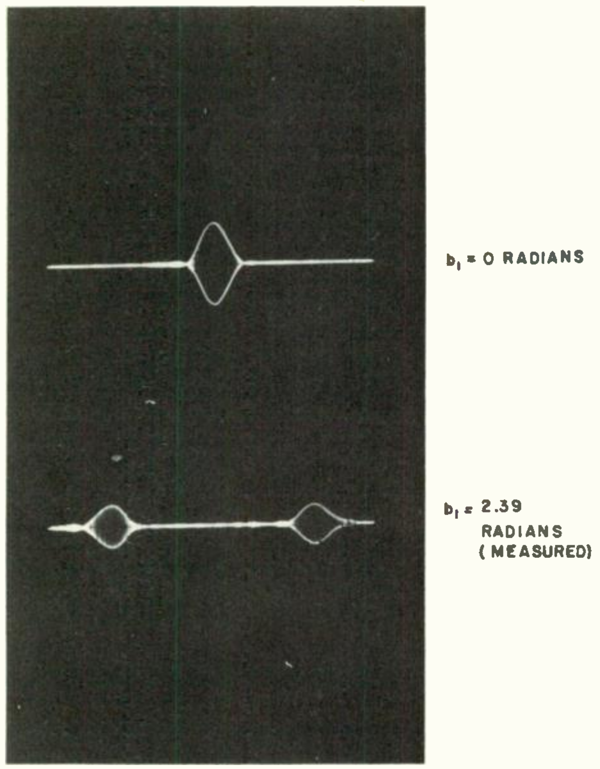


FIG. 5 EXPERIMENTAL CONFIRMATION OF SUPPRESSION OF NORMAL PULSE COMPRESSION SIGNAL ($J_0(b_1) = 0$)
f_m = 6 CYCLES/INPUT PULSE WIDTH

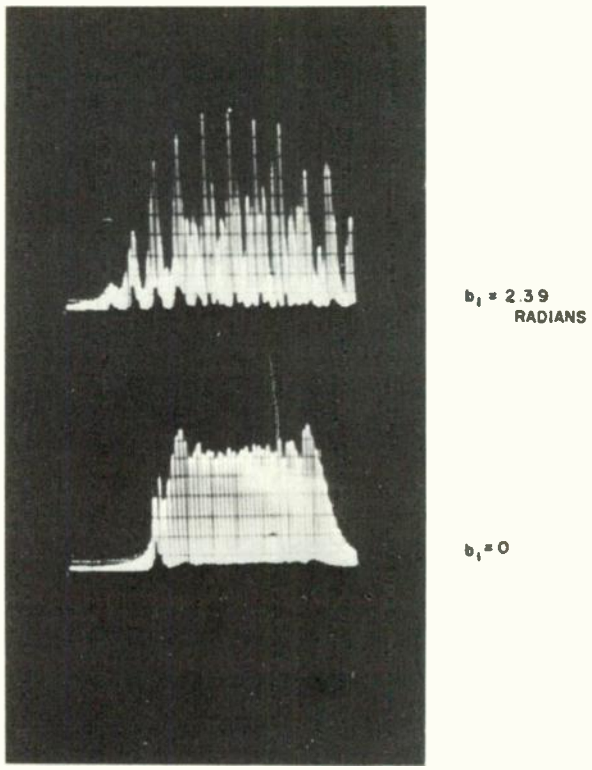


FIG. 6 EFFECT OF PHASE MODULATION ON PULSE COMPRESSION SPECTRUM
f_m = 6 CYCLES/INPUT PULSE WIDTH

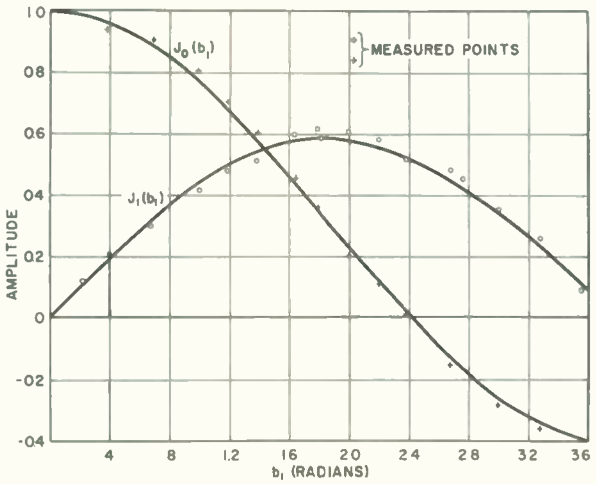
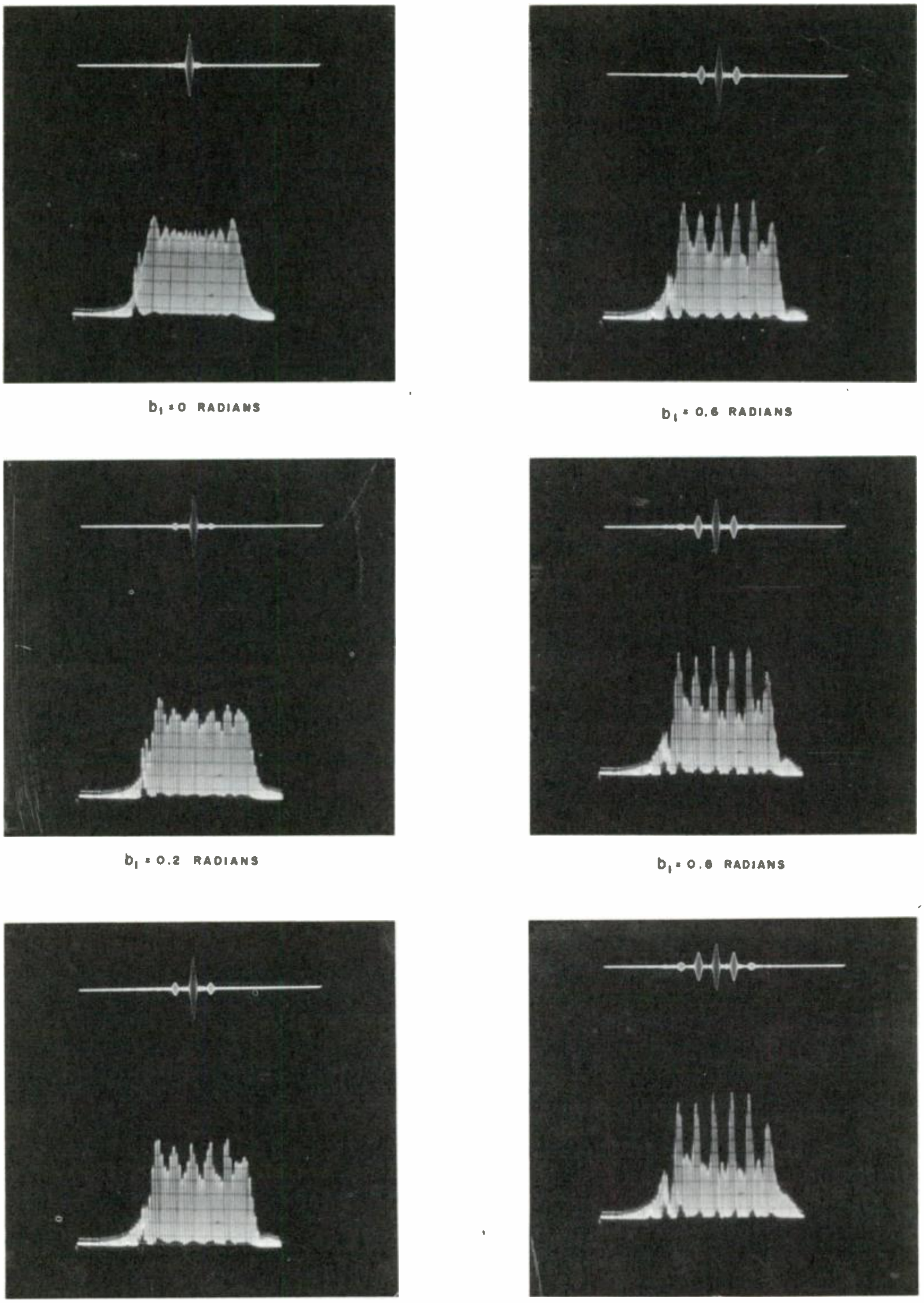


FIG. 7 AMPLITUDE OF MODULATION-DISTORTED PULSE COMPRESSION SIGNALS



$b_1 = 0$ RADIANS

$b_1 = 0.2$ RADIANS

$b_1 = 0.4$ RADIANS

$b_1 = 0.6$ RADIANS

$b_1 = 0.8$ RADIANS

$b_1 = 1.0$ RADIANS

FIG. 8 EFFECT OF SINUSOIDAL PHASE MODULATION ON PULSE COMPRESSION WAVEFORM AND SPECTRUM
$f_m = 6$ CYCLES/INPUT PULSE WIDTH

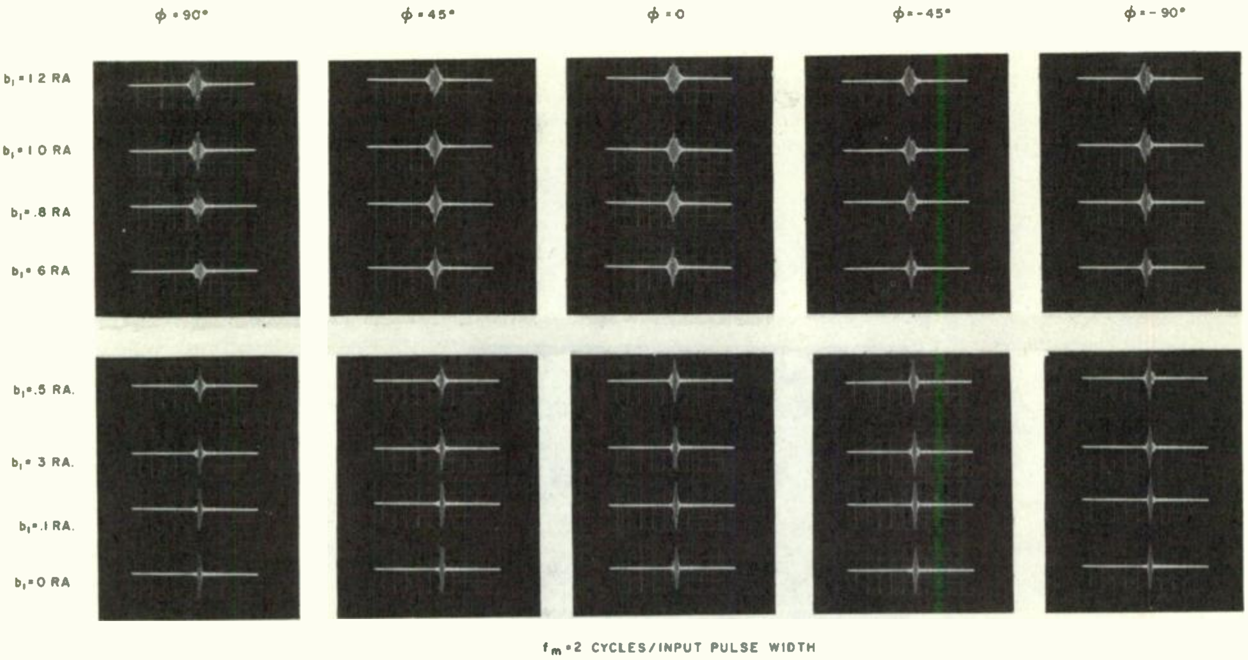


FIG. 9 EFFECT OF ϕ ON PULSE COMPRESSION PAIRED ECHOES CAUSED BY SINUSOIDAL PHASE-ERROR MODULATION

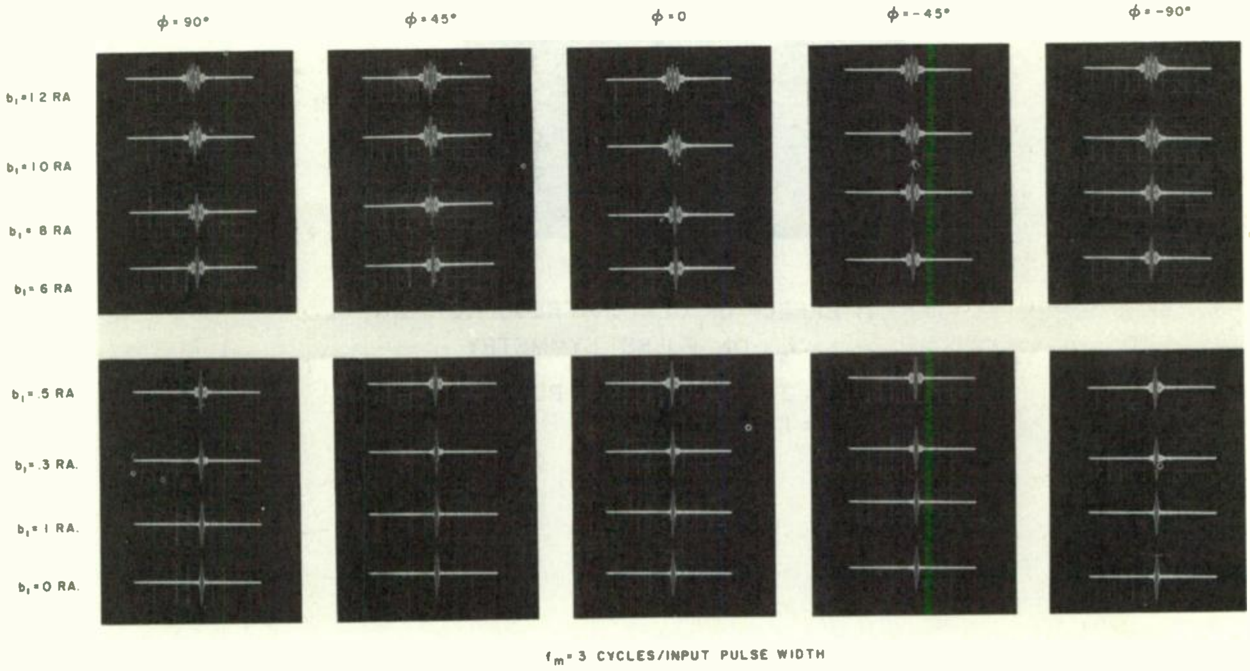


FIG. 10 EFFECT OF ϕ ON PULSE COMPRESSION PAIRED ECHOES CAUSED BY SINUSOIDAL PHASE-ERROR MODULATION

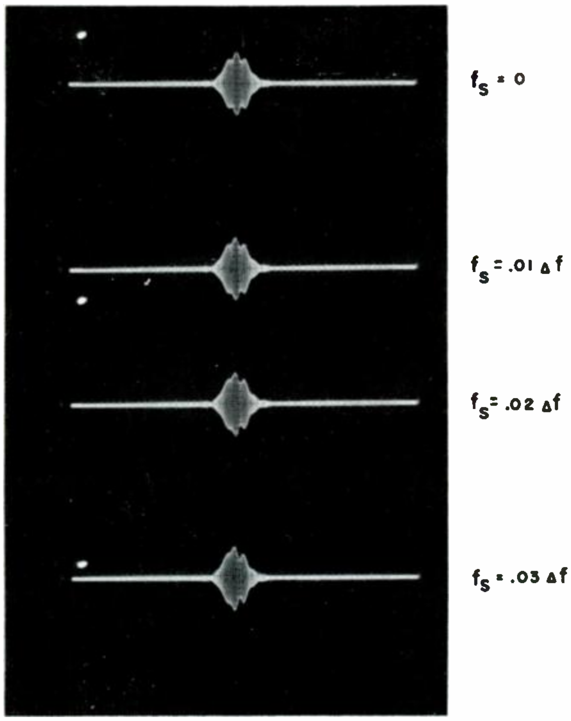


FIG. 11 EFFECT OF CENTER-FREQUENCY SHIFT, f_s, ON PULSE SYMMETRY

f_m = 2 CYCLES/INPUT PULSE WIDTH
b_1 = 1.4 RADIANS

THE RELATIVE EFFICIENCIES OF VARIOUS BINARY DETECTION SYSTEMS

Richard O. Rowlands
Ordnance Research Laboratory
Pennsylvania State University

Summary

This paper attempts to fill the need of the engineer who is designing a signal detection system and who wants a rough idea of the relative merits of the various basic detection methods that are available. Five common binary detection systems are analyzed and formulas are derived for the signal-to-noise ratios required to give the same probabilities of error in each system. The effects of varying the duration of the signal and the bandwidth of the noise are discussed for each system, and the region is defined within which the detectability of the signal is improved by clipping.

Introduction

A signal y is received and a decision has to be made whether y consists of noise only or whether it contains a signal x of duration T in addition to the noise. The detectors considered in this analysis are assumed to operate in a binary manner, i.e., a threshold is set and, if the output of the detector exceeds this threshold, the decision D_1 is made that the signal x is present; but, if the threshold is not exceeded, the decision D_o is made that x is absent. It is assumed that the noise is Gaussian limited to a frequency band W and that samples taken at Nyquist intervals are independent. There are, therefore, a total of $2WT = n$ samples. Although the physical operations to be considered are performed on the continuous signal, the same mathematical results may be obtained more simply by using the sampling technique. The output of each detector will be designated Z_o in the absence of the signal x and Z_1 in the presence of the signal. The noise voltage will be denoted by v and, for convenience, its power will be assumed to be unity; therefore, the signal power S is also the signal-to-noise ratio.

Systems

1. Ideal Correlator

The ideal correlator performs the operation $\sum x_i v_i$. In the absence of x, the signal $y = v$ and the output Z_o of the correlator is given by $Z_o = \sum x_i y_i$. Since each of the terms v_i has a Gaussian distribution with a mean of zero and a variance 1, the terms $x_i v_i$ will also have the same distribution with variance x_i^2. Therefore, Z_o will have a mean equal to zero and a variance equal to the sum of the variances of the individual terms, i.e., a variance equal to $\sum x_i^2 = nS$.

When x is present in y, the output Z_1 is given by $Z_1 = \sum x_i(x_i + v_i) = \sum x_i^2 + \sum x_i v_i$. The first term is a constant of value nS while the second is identical to Z_o.

Output	Mean	Variance
Z_o	0	nS
Z_1	nS	nS

2. Square-Law Detector

The square-law detector performs the operation $\sum y_i^2$. When x is absent, $Z_o = \sum v_i^2$. This is the well known Chi Squared distribution. Its mean is n and its variance is 2n. When x is present,

$$Z_1 = \sum (x_i + v_i)^2 = \sum x_i^2 + \sum 2\, x_i v_i + \sum v_i^2 .$$

The first term is a constant of value nS, the second is Gaussian with a variance of 4 nS, and the third is Chi Squared. The distribution of Z_1 will, therefore, be the convolution of a Gaussian with a Chi Squared distribution. Its variance will be the sum of their variances.

Output	Mean	Variance
Z_o	n	2n
Z_1	nS + n	4nS + 2n

3. Two-Channel Correlator

In certain receiver systems, such as the sum and difference patterns of arrays, the signal x is either present or absent simultaneously in the two channels, but the noise signals in the two channels are uncorrelated. The two-channel correlator multiplies the signal in one channel by the signal in the other.

When x is absent $Z_o = \sum v_{1i} v_{2i}$. Because each v_{2i} has a Gaussian distribution with a variance of 1, the variance of $v_{1i} v_{2i}$ will be v_{1i}^2; therefore, it would appear at first sight that Z_o has a Gaussian distribution with a variance of $\sum v_{1i}^2$, but this itself is a variable with a mean value of n. The probability density of Z_o will, therefore, be the sum of Gaussian probability densities of variance $\sum v_{1i}^2$ each weighted by the probability density $p(\sum v_{1i}^2)$ of the variance, which is a Chi Squared distribution, i.e.,

$$p(Z_o) = \int_0^{\infty} \frac{1}{\sqrt{2\pi\psi}}\, e^{-\frac{Z_o^2}{2\psi}}\; \frac{\psi^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\left(\frac{n}{2}-1\right)!}\, e^{-\frac{\psi}{2}}\, d\psi .$$

This distribution is rather similar to the Gaussian distribution: its variance is n, but its kurtosis of 3n(n + 2) is slightly greater than the Gaussian value of $3n^2$. In general the level of the wanted signal will not be the same in both channels. Let it be (1 + a) x in one channel and (1 - a) x in the other; then

$$
\begin{aligned}
Z_1 &= \Sigma \; \{(1 + a)x_i + v_{1i}\} \; \{(1 - a)x_i + v_{2i}\} \\
&= \Sigma \; (1 - a^2)x_i^2 + \Sigma \; \{(1 + a)x_i v_{2i} \\
&\quad + (1 - a)x_i v_{1i}\} \\
&\quad + \Sigma \; v_{1i} v_{2i}.
\end{aligned}
$$

In the two previous systems, the signal-to-noise ratio was unambiguously defined since each system had a single-channel input; but, in this system there are two channels, each with a different signal-to-noise ratio. The signal-to-noise ratio of the system will, therefore, be defined as the highest ratio obtainable by adding the two channels. This optimum value is obtained by giving to each channel a weight that is proportional to its signal-to-noise ratio. This process yields: $(1 + a)^2x + (1 + a)v_1 + (1 - a)^2x + (1 - a)v_2$. The signal power is $4(1 + a^2)^2 x^2$ and the noise power is $2(1 + a^2)$; hence, $S = 2(1 + a^2)x^2$. The mean of Z_1 is the first term in the expression and is equal to $n(1 - a^2)\, S/2\, (1 + a^2)$. The second term is Gaussian with a mean of zero and a variance of nS; the third term has already been discussed.

Output	Mean	Variance
Z_o	0	n
Z_1	$n(1 - a^2)\, S/2\, (1 + a^2)$	nS + n

4. Sign Correlator

A technique that is sometimes employed when the signal envelope is rectangular is that of clipping the received wave form before feeding it into the correlator, which compares its sign with that of x.

When x is absent, each sample of the clipped noise will have probabilities of 0.5 that it is either positive or negative. The probability distribution of the sum of a number of samples is binomial. Let the output of the correlator be + 1 when the samples correlate and - 1 when they do not: the mean, of Z_o, will then be zero and its variance n.

When a signal of amplitude x is also present, x + v will have the same sign as x for $-x < v < \infty$; the probability that a single sample will correlate is the cumulative Gaussian probability

$$q\ (x) = \int_{-|x|}^{\infty} \frac{1}{\sqrt{2\ \pi}}\ e^{-u^2/2}\ du.$$

This is related to the error function defined as follows:

$$
\begin{aligned}
\operatorname{erf} |x/\sqrt{2}| &= \frac{2}{\sqrt{\pi}} \int_0^{|x/\sqrt{2}|} e^{-t^2}\, dt \\
&= \frac{2}{\sqrt{2\ \pi}} \int_0^{|x|} e^{-u^2/2}\, du \\
&= 2\ q\ (x)\ -1.
\end{aligned}
$$

For signal plus noise, each sample will, therefore, produce an output of 1 with probability q(x) and an output of - 1 with a probability of 1 - q(x). It will be necessary here to simplify the problem by considering x to be a pulsed square wave so as to make q(x) a constant q. Having made this simplification, the results that will be obtained will be in the nature of an upper bound on the detectability of the signal. The mean of a single sample is

$$2q\ -1 = \operatorname{erf}\ |\ x/\sqrt{2}\ | = \operatorname{erf} |\sqrt{S/2}\ |.$$

The variance is the mean of the square minus the square of the mean, which is $1 - \operatorname{erf}^2 \sqrt{S/2}$. The output Z_1 is the sum of n samples.

Output	Mean	Variance
Z_o	0	n
Z_1	$n \operatorname{erf} \sqrt{S/2}$	$n(1 - \operatorname{erf}^2\sqrt{S/2})$

5. Two-Channel Sign Correlator

The two-channel sign correlator is similar to the two-channel correlator except that a clipper is used in both channels. As before, the levels of the wanted signals in the two channels will be designated (1 + a)x and (1 - a)x. The probabilities q_1 and q_2 that clipped samples of each channel have the same sign as x are given by

$$
\begin{aligned}
2q_1 - 1 &= \operatorname{erf}\ (1 + a)x/\sqrt{2} \\
&= \operatorname{erf}\ (1 + a)\ \sqrt{S/4(1 + a^2)}
\end{aligned}
$$

and

$$
\begin{aligned}
2q_2 - 1 &= \operatorname{erf}\ (1 - a)x/\sqrt{2} \\
&= \operatorname{erf}\ (1 - a)\ \sqrt{S/4(1 + a^2)},
\end{aligned}
$$

where the signal-to-noise ratio of the system is as previously defined. Two samples will correlate with each other if they are both of the same sign as x or both different from x. The probability r of this occurring is given by

$$
\begin{aligned}
r &= q_1 q_2 + (1 - q_1)(1 - q_2) \\
&= 1 - (q_1 + q_2) + 2q_1 q_2.
\end{aligned}
$$

The distribution of Z_1 will, therefore, be similar to that of the single-channel sign correlator except that the probability r has to be substituted for q. The mean of Z_1, then, will be $(2r - 1)$ and its variance will be $1 - (2r - 1)^2$. For equal signal-to-noise ratios in the two channels, the means and variances will be:

Output	Mean	Variance
Z_o	0	n
Z_1	$n \operatorname{erf}^2 \sqrt{S/4}$	$n(1-\operatorname{erf}^4 \sqrt{S/4})$

Comparison of the Systems

The means and variances of the probability distribution of Z_o and Z_1 have been calculated. The probabilities of error are not affected by changes in the origin or scale of the distribution; therefore, it will be advantageous to shift origins and to change scales to make the mean of Z_o zero and its variance unity for each system. The resulting means and variances of Z_1 will be as given in Table 1. It will be noted that for the two-channel system a value of "a" equal to zero has been chosen for the calculations to be performed from here on. This value is obtained when the signal-to-noise ratios in the two channels are equal and it represents the condition of maximum efficiency of the systems.

TABLE 1

	Z_1	
System	Mean	Variance
1. Ideal Correlator	$\sqrt{nS}$	1
2. Square-Law Detector	$\sqrt{n}\ S/\sqrt{2}$	$1 + 2S$
3. Two-Channel Correlator	$\sqrt{n}\ \ S/2$	$1 + S$
4. Sign Correlator	$\sqrt{n} \operatorname{erf} \sqrt{S/2}$	$1-\operatorname{erf}^2 \sqrt{S/2}$
5. Two-Channel Sign Correlator	$\sqrt{n} \operatorname{erf}^2 \sqrt{S/4}$	$1-\operatorname{erf}^4 \sqrt{S/4}$

Before going on to compare the systems, it will be profitable at this point to discuss the significance of some of the results summarized in the table. It will be noticed that the number of samples affects each system in exactly the same way, namely, that the mean of Z_1 is directly proportional to $\sqrt{n}$. When n is increased by increasing the duration of the signal, its detectability is increased to the same extent in each system because S is held constant. If, however, the bandwidth of the noise is altered, a trade is made between n and S such that the product nS remains constant, and each system must be considered separately.

The product nS is equal to the signal energy divided by the noise power per cycle and is the basis of the well-known result that the detectability of a signal with an ideal correlator is independent of the form of the signal or the bandwidth of the noise.

For the square-law detector and the two-channel correlator, it is an advantage to trade n for S since this will increase the product $\sqrt{nS}$ and, hence, the mean of Z_1. The variance also increases but not enough to offset the advantage gained by the increasing mean. In practical terms this means that all noise outside the bandwidth of the signal should be eliminated before it reaches the detection system.

The reverse is true for the sign correlator since the maximum value of the error function is one; therefore, a decrease in S by an increase in the bandwidth of the noise in the system is a good trade.

The two-channel sign correlator has some of the characteristics of both the sign correlator and the two-channel correlator. When the signal-to-noise ratio is high, it is an advantage to trade some of this for an increase in the noise bandwidth; but, when the signal-to-noise ratio is low, erf^2 is proportional to S and it becomes advantageous to filter out some of the noise. The cross-over point, at which the mean of Z_1 is a maximum, occurs when $S = 4$.

The normalized output signal-to-noise ratio, which is equal to the mean of Z_1 divided by the square root of its variance, is often quoted as a criterion for judging the relative merits of the various systems; however, this is a very loose criterion that fails badly when the signal is clipped. The proper criterion to use is the input signal-to-noise ratio, S, required to give the same probability of error in each system. One system may then be said to be "x" db better than another system if it gives the same probability of error with a signal-to-noise ratio that is "x" db lower. There are, of course, two kinds of error; false alarm, the probability of which will be designated $P_o(D_1)$; and false rest, the probability of which will be designated $P_1(D_o)$. To reduce the problem to manageable dimensions, it will now be assumed that the number of samples involved is large so that the probability distributions of both Z_o and Z_1 become approximately Gaussian.

The threshold, which determines what decision is made, will be designated $K\sqrt{n}$. The probability of false alarm will, therefore, be equal to the area of the tail of the Z_o distribution lying beyond this threshold and will be given by

$$P_o(D_1) = \frac{1}{\sqrt{2\pi}} \int_{K\sqrt{n}}^{\infty} e^{-u^2/2}\, du.$$

The probability of false rest will be equal to the area of the tail of the Z_1 distribution lying below the threshold and will be given by

$$P_1(D_o) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{K\sqrt{n}} e^{-(u-\mu)^2/2\sigma^2}\, du,$$

where μ is the mean and σ^2 is the variance of Z_1. If the variable in the latter expression is changed by letting $u_1 = (u-\mu)/\sigma$, then

$$P_1(D_o) = \frac{1}{\sqrt{2\ \pi}} \int_{-\infty}^{\frac{K\sqrt{n} - \mu}{\sigma}} e^{-u_1^2/2}\, du_1$$

$$= \frac{1}{2\ \pi} \int_{-\infty}^{K_1\sqrt{n}} e^{-u_1^2/2}\, du_1 .$$

The normalized threshold $K_1\sqrt{n}$ of Z_1 is related to $K\sqrt{n}$ by the expression

$$K_1\sqrt{n} = (K\sqrt{n} - \mu)/\sigma. \qquad (1)$$

In most cases of practical interest, K_1 will be approximately equal to -K because the area under the tail of the Gaussian curve changes rapidly for small changes in threshold; e.g., when the probability of error is 0.0001, a change of 3 per cent in the threshold increases the probability of error by 60 per cent, and a change of 20 per cent increases the probability of error tenfold. The expression $K_1\sqrt{n} = -K\sqrt{n}$, therefore, becomes $-K\sqrt{n} = (K\sqrt{n} - \mu)/\sigma$ and by using the values of μ and σ given in Table 1, the following relationships between S and K are obtained for the various systems:

1. Ideal Correlator	$S = 4K^2$
2. Square-Law Detector	$S = 4K^2 + 2\sqrt{2K}$
3. Two-Channel Correlator	$S = 4K^2 + 4K$
4. Sign Correlator	$\operatorname{erf}\sqrt{S/2} = \frac{2K}{1 + K^2} \quad K \leq 1$
5. Two-Channel Sign Correlator	$\operatorname{erf}\sqrt{S/4} = \sqrt{\frac{2K}{1 + K^2}} \quad K \leq 1$

The curves representing these equations are plotted in Fig. 1 and, although only strictly applicable to equal probabilities of error, they are approximately correct for quite wide variations in these probabilities as explained above.

Discussion

It has been assumed that the value of n has been sufficiently large to enable each distribution to be approximated by a Gaussian distribution. A rough estimate of the resulting error will now be obtained.

The distribution that differs to the greatest extent from the Gaussian is the Chi Squared. Its lower tail stops short at zero, whereas its upper tail is drawn out. An interesting thing about this distribution, however, is that the distance between two ordinates representing the boundaries between two equal areas under the tails is nearly the same as if the distribution were Gaussian (except when one ordinate is very close to zero). For example: when n = 8, the distance between the Gaussian and Chi Squared ordinates representing error probabilities of 0.001 are 6.18 and 6.33 respectively; and for error probabilities of 0.00001, they are 8.5 and 9.15. The restriction that n must be large may, therefore, be removed when the Z_o and Z_1 distributions are both Chi Squared, which occurs in the square-law detector when the signal-to-noise ratio is low. When the signal-to-noise ratio is high, the performance of the square-law detector is not as good as is indicated on the chart; it deteriorates by 1 db for an error probability of 0.01 and n = 8. For the same degree of deterioration, if the error probability is reduced to 0.001 the value of n must be increased to 27. The graph for the two-channel correlator has been computed on the assumption that the signal-to-noise ratios in the two channels are equal. When they differ by 10 log $(1 + a)/(1 - a)$ db, the loss in efficiency is 10 log $(1 + a^2)/(1 - a^2)$. A graph of this loss versus the ratio difference is plotted in Fig. 2.

A Further Comparison of the Ideal and Sign Correlators

In Fig. 1 where the magnitudes of the thresholds are equal, the ideal correlator is always superior to the sign correlator, but this is not the case under all conditions. The following equations are obtained for these two systems by inserting the appropriate values of μ and σ in equation (1):

Ideal Correlator	$K_1 = K - S$
Sign Correlator	$K_1 = \frac{K - \operatorname{erf}\sqrt{S/2}}{\sqrt{1 - \operatorname{erf}^2\sqrt{S/2}}}$.

It is possible to find values of K_1, K, and S that simultaneously satisfy both equations. For these values the two systems are equally efficient. A graph of K_1 versus K is shown in Fig. 3 for this condition and represents the boundary between the regions where the relative efficiencies of the two systems are reversed. The reason why the sign correlator can become superior is that a high noise spike that is negatively correlated with the signal can nullify a number of positively correlated samples in the ideal detector, but the effect of clipping is to cause all samples to have the same weight so that only one correlated sample can be nullified by the noise spike in the sign correlator.

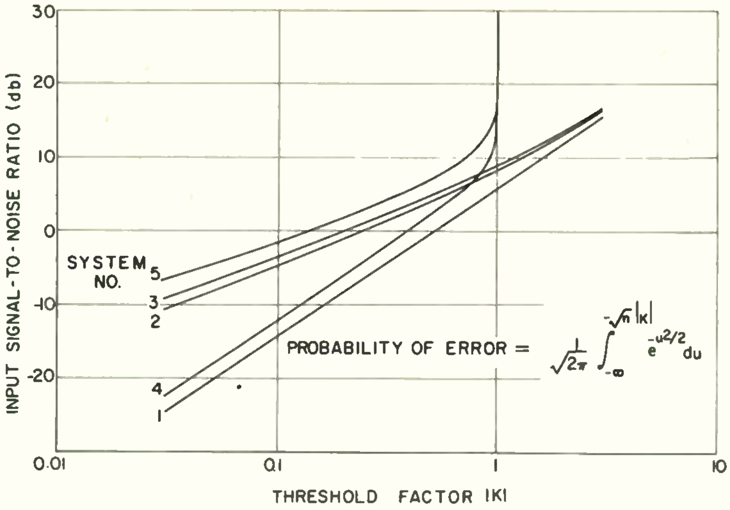


Fig. 1. Input signal-to-noise ratio versus threshold factor for various systems when $k_o = -k_1$.

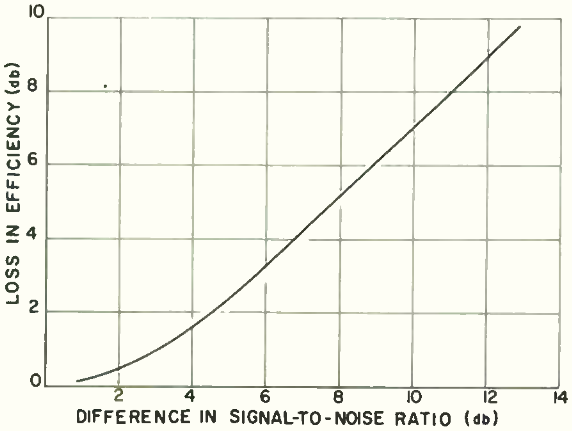


Fig. 2. Loss in efficiency of the two-channel system when the signal-to-noise ratios in the two channels differ.

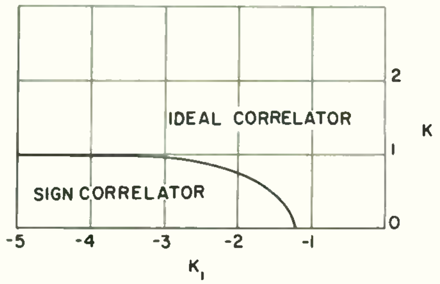


Fig. 3. Showing the regions where the ideal and sign correlators are the optimum detectors.

PARTIAL ORDERING OF DISCRETE CHANNELS[*]

T. T. Chang and J. G. Lawton
Cornell Aeronautical Laboratory, Inc.
Buffalo 21, New York

Summary

A discrete, finite, memoryless channel K_1 is said to include another such channel K_2 if it is possible to duplicate all of the transition probabilities of K_2 by interposing K_1 between pairs of suitably chosen pre- and post-channels. In general, it is not known what the necessary and sufficient conditions for inclusion are. The problem has been solved for the following cases:

1. K_2 and K_1 are symmetric channels.
2. K_2 is an n-ary symmetric channel, and K_1 is an m-ary channel with $m \leq n$.
3. K_2 is an n-ary symmetric channel, and K_1 is an m-ary channel with all column maxima contained in t rows where $t \leq n$.
4. K_2 and K_1 are binary channels.

A channel K_1 is said to be "not more lossy" than another channel K_2 if, for every assignment of losses to the possible transitions and all probability distributions of the inputs, the minimum expected loss is not greater for K_1 than for K_2. A channel K_1 is "more informative" than K_2 if K_1 has an information rate which is not less than that of channel K_2 regardless of the probability distribution of the inputs. Partial orderings of binary channels by the criteria of "includes," "not more lossy" and "more informative" are derived and compared.

Introduction

Digital communications systems may be quantitatively compared if a performance index, such as error probability, expected loss or information rate can be computed. The value attained may depend not only on the characteristics of the channel used, but also on such parameters as the probability distribution of inputs, assignment of losses, etc. On this basis, one can conclude only that one channel is superior to another in conjunction with the particular parameters of the system which were used in the computations. However, if the performance obtained when using one channel is superior to that when using another channel regardless of the auxiliary parameters, then these channels can be ordered absolutely. We will be concerned with such orderings for n-ary symmetric and general (asymmetric) binary channels.

Channel Inclusion

Shannon formally defined channel inclusion as follows:[1]

"Let $q_i(j)$ $(i=1,\ldots,a;\ j=1,\ldots,b)$ be the transition probabilities for a discrete memoryless channel K_1 and $p_k(\ell)$ $(k=1,\ldots,c;\ \ell=1,\ldots,d)$ be those for K_2. We shall say that K_1 includes K_2, $K_1 \supseteq K_2$, if and only if there exist two sets of transition probabilities, $r_{\alpha k}(i)$ and $t_{\alpha j}(\ell)$, with

$$r_{\alpha k}(i) \geqq 0, \textstyle\sum_i r_{\alpha k}(i) = 1,$$

and

$$t_{\alpha j}(\ell) \geqq 0, \textstyle\sum_\ell t_{\alpha j}(\ell) = 1,$$

and there exists

$$g_\alpha \geqq 0, \textstyle\sum_\alpha g_\alpha = 1$$

with

$$\sum_{\alpha, i, j} g_\alpha r_{\alpha k}(i) q_i(j) t_{\alpha j}(\ell) = p_k(\ell) \qquad (1)$$

"Roughly speaking, this requires a set of pre- and post-channels R_α and T_α, say, which are used in pairs, g_α being the probability for the pair with subscript α. When this sort of operation is applied, the channel K_1 looks like K_2."

The conditions under which the required set of transition matrices $[R_\alpha]$, $[T_\alpha]$ and their probabilities g_α, can be found are, in general, unknown. For symmetric channels necessary and sufficient conditions for inclusions may easily be derived. Furthermore, the required pre- and post-channels may be easily constructed.

One is frequently only interested in the probability of correct transition $p_i(i)$. The distribution of the probability of error $p_i(e) = 1 - p_i(i) = \sum_{i \neq j} p_i(j)$ may be of little or no interest. One may then inquire under which conditions a channel can duplicate the probability of correct transition of another channel. Following Shannon, we shall use the term "includes" and the symbol $\supseteq$ to indicate the ability to duplicate the entire set of transition probabilities $p_i(j)$ and the term "includes on an error probability basis" to indicate the ability to duplicate the probabilities of correct transition $p_i(i)$.

Symmetric Channels

A memoryless n-ary symmetric channel is specified by an $n \times n$ transition matrix having:

$$p_i(i) = p \text{ for all } i \qquad p_i(j) = \frac{1-p}{n-1} \text{ for } i \neq j \qquad (2)$$

In practice, one is ordinarily only interested in these channels when $p > \frac{1}{n}$; this inequality will be assumed for all symmetric channels. In this case, it is evident that channels having the same size alphabet, n, may be ordered in accordance with their probability of correct transition p.[*] We will examine the question of how one may compare symmetric channels having alphabets of differing sizes and some related problems.

[*] This work was sponsored by the Rome Air Development Center under Contract AF 30(602)-2210.

[*] An n-ary symmetric channel with probability of correct transition q can be obtained by cascading two such channels with probabilities of correct transition p_a and p_b, satisfying:

$$p_a - q = \triangle \geq 0 \qquad p_b = \frac{nq - 1 + \triangle}{nq - 1 + n\triangle} \geqslant q$$

Inclusion of an n-ary Symmetric Channel by an m-ary Symmetric Channel for $n \geq m$

Consider the situation depicted in Figure 1, which shows a cascade consisting of a pure[*] pre-channel, an m-ary symmetric channel with transition probabilities given by $[Q]_m$ and a noisy post-channel. Note that in Figure 1 the post-channel is just a reflection of the pure pre-channel.

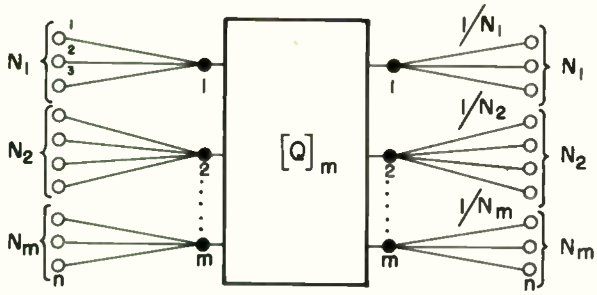


Figure I

We call a channel in which the input is determined with certainty by specification of the output a reflected pure channel.[**] The n inputs are connected to the input terminals of the m-ary channel in such a manner that $N_i \geq 1$ inputs are connected to the ith terminal of the m-ary channel. Let $p_j(c)$ be the probability of correct transition through the cascade of the jth input symbol. Then $p_j(c) = q_i(i)\,p_i(j) = q\,\frac{1}{N_{i,j}}$ where $p_i(j)$ denotes the transition probability of the post-channel and $N_{i,j}$ the number of inputs connected to the same ith input of $[Q]_m$ as the jth input. The average value of the numbers $p_j(c)$ is

$$\bar{p}(c) = \frac{1}{n}\sum_j p_j(c) = \frac{1}{n}\sum_j q\,\frac{1}{N_{i,j}} \qquad (3)$$
$$= \frac{q}{n}\left\{\sum_{k=1}^{N_1}\frac{1}{N_1} + \sum_{k=1}^{N_2}\frac{1}{N_2} + \cdots + \sum_{k=1}^{N_m}\frac{1}{N_m}\right\} = \frac{mq}{n}$$

Note that $\bar{p}(c)$ is also the statistical average of the probability of correct transition if all input symbols are equally probable. Unless all N_i are equal, the probability of correct transition $p_j(c)$ of the different input symbols are not all equal. Furthermore, the various possible errors do not occur with equal probabilities (even if all N_i are equal). Therefore, the cascade as described above does not have all the properties of an n-ary symmetric channel. These properties can, however, be attained simply by relabeling input and output terminals. $n!$ such relabelings are possible[*] and, if used with equal probability, will result in cascade having all the characteristics of an n-ary symmetric channel with probability of correct transition equal to $\bar{p}(c)$. It is, obviously, possible to form a cascade using the same m-ary channel which has the characteristics of an n-ary symmetric channel with $p < \frac{mq}{n}$. Therefore, for $n \geq m$, an m-ary symmetric channel with probability of correct transition q includes all n-ary symmetric channels with probability of correct transition p for which $np \leq mq$.

That this condition is also necessary for inclusion is shown below. Note that $\bar{p}(c) = \frac{mq}{n}$ for any pure pre-channel provided $N_i \geq 1$ for all i. Since the central m-ary channel is given, we need only to show that permitting $N_i = 0$ results in a lowering of $\bar{p}(c)$ and that the best post-channel is a reflection of the input channel.[**]

Consider the arrangement yielding the greatest $\bar{p}(c)$ with some $N_k = 0$. Then there exists at least one $N_\ell \geq 2$. Let $p(c,i)$ designate the joint probability of correct transition through the cascade and excitation of the ith output terminal of the m-ary channel. Then

$$\bar{p}(c) = \sum_{i=1}^{m} p(c,i) = \sum_{i \neq \ell, k} p(c,i) + p(c,\ell) + p(c,k) \qquad (4)$$

or, since $p(c,k) \leq \frac{1}{n}(1-q)/(m-1)$,[***]

$$\bar{p}(c) \leq \sum_{i \neq \ell, k} p(c,i) + p(c,\ell) + \frac{1}{n}\,\frac{1-q}{m-1} \qquad (5)$$

[*] A pure channel is defined[1] as a channel in which all transition probabilities are either 0 or 1, so that each input letter is transferred with certainty into some output letter.

[**] If we designate the output symbols by y and the input symbols by x, then we have

$p_x(y) = 0,\ 1$ for a pure channel, and

$p_y(x) = 0,\ 1$ for a reflected pure channel.

(Note that in a reflected pure channel, the transition probabilities need not be equal to $\frac{1}{n_i}$, only $\sum_{k=1}^{n_i} p_i(k) = 1$ is required.)

[*] Although $n!$ relabelings are possible, they are not all needed in order to ensure that the overall cascade has symmetric characteristics. Since relabeling within a group of N_i inputs does not alter any of the transition probabilities, we may divide by the number of such relabelings; there are $\prod_{i=1}^{m} N_i!$ of these. The number of required relabelings may be further reduced by noting that interchange of groups having the same number of inputs does not affect the overall transition probabilities.

[**] We can dismiss the possibility of the input channel not being a pure channel, since any stochastic channel can be built up from pure channels selected with proper probabilities.

[***] The equality is achieved if and only if in the post-channel $\sum_{j=1}^{n} p_k(j) = 1$.

Note that we assumed all input symbols to be equally probable. (This is permissible in a proof of necessity, since the property of inclusion must hold for all input probabilities.)

Let s , t designate two of the N_ℓ inputs of the cascade connected by the pre-channel to the ℓth input of the m-ary channel and consider the effects of the following changes.

In the pre-channel, reduce N_ℓ by one by setting $p_s'(\ell)=0$ and increase N_k to one by setting $p_s'(k)=1$. In the post-channel, set $p_k'(s)=1$ and $p_\ell'(t)=p_\ell(t)+p_\ell(s)$. (Where primed quantities refer to the modified channels.) Then we find

$$\sum_{i\neq \ell,k} p'(c,i) = \sum_{i\neq \ell,k} p(c,i) \tag{6}$$

$$p'(c,\ell) = p(c,\ell) \tag{7}$$

$$p'(c,k) = \frac{1}{n} q > \frac{1}{n} \frac{1-q}{m-1} \geq p(c,k) \tag{8}$$

so that $\bar{p}'(c) > \bar{p}(c)$. Therefore, none of the N_i should be zero.

To prove that the post-channel should be a reflection of the pure pre-channel with all $N_l \geq 1$, consider a post-channel with transition probability $p_k(t)>0$ when the t^th input terminal to the cascade is not connected to the k^th input terminal of the m-ary channel ($p_t(k)=0$ in the pre-channel). For this cascade

$$n p(c,k) = \sum_{j\in A} q p_k(j) + \sum_{j\notin A} \frac{1-q}{m-1} p_k(j) \tag{9}$$

where the set A contains those inputs connected by the pre-channel to the k^th input of the m-ary channel. If, now, the post-channel is modified such that $p_k'(t)=0$ and

$$\sum_{j\in A} p_k'(j) = \sum_{j\in A} p_k(j) + p_k(t) \tag{10}$$

then

$$p'(c,k) = p(c,k) + \left(q - \frac{1-q}{m-1}\right)\frac{1}{n} p_k(t) \tag{11}$$

and $$p'(c,i) = p(c,i) \qquad i\neq k \tag{12}$$

so that $$\bar{p}'(c) > \bar{p}(c) . \tag{13}$$

Inclusion of an n-ary Symmetric Channel by an m-ary Symmetric Channel for $n< m$

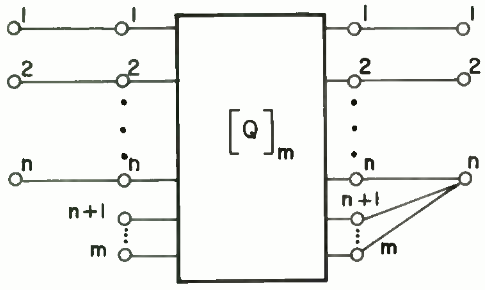


Figure 2

By the use of pure pre- and post-channels, as in Figure 2, it is seen that

$$n\bar{p}(c) = nq + (m-n)\frac{1-q}{m-1} . \tag{14}$$

From our previous arguments, it follows that the necessary and sufficient conditions for the inclusion of an n-ary symmetric channel with probability of correct transition p by an m-ary symmetric channel for $n<m$ is that $p \leq \bar{p}(c)$ as given by Equation (14).

Inclusion of an n-ary Symmetric Channel by an m-ary Asymmetric Channel for $n\geq m$.

Let a given m-ary channel be labeled such that its trace $T[Q]$ attains its maximum value.* Then, by proceeding as in the section before the previous one, we find $\bar{p}(c)=\frac{T[Q]}{n}$. Therefore, a sufficient condition for inclusion of an n-ary symmetric channel with probability of correct transition p , by an m-ary channel, for the case $n>m$ is $T[Q]\geq np$. If

$$q_i(j) \leq q_j(j), \qquad i\neq j \tag{15}$$

then the column maxima fall on the main diagonal. We show, below, that the necessary and sufficient condition for inclusion of an n-ary symmetric channel $[P]_n$ by an m-ary asymmetric channel $[Q]_m$ for $m<n$ is that

$$(\text{Sum of the column maxima of } [Q]_m) \geq np \tag{16}$$

To show the sufficiency of condition (16), let the asymmetric channel be labled such that the first $t\leq m$ rows of $[Q]_m$ contain all the column maxima of $[Q]_m$. Figure 3 illustrates the manner in which the channel Q represented by the $m\times m$ matrix $[Q]_m$ which does not satisfy (15) can be converted into a channel Γ represented by a $t\times t$ matrix $[\Gamma]_t$ which satisfies (15).

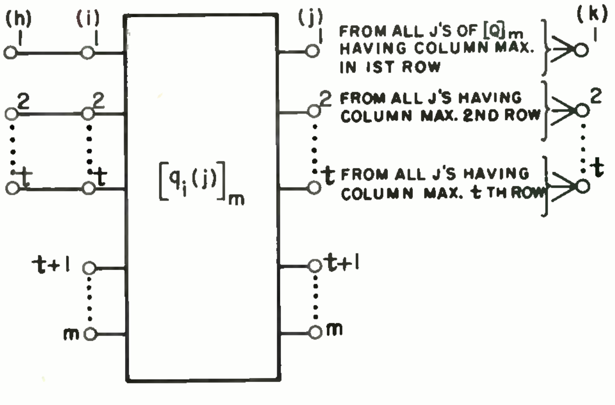

(h)
(i)
(j)
(k)
1
1
1
1
2
2
2
2
t
t
t
t
t+1
t+1
m
m
$[q_i(j)]_m$
FROM ALL J'S OF $[Q]_m$ HAVING COLUMN MAX. IN 1ST ROW
FROM ALL J'S HAVING COLUMN MAX. 2ND ROW
FROM ALL J'S HAVING COLUMN MAX. t TH ROW

Figure 3

$[\Gamma]_t$ is obtained from $[Q]_m$ as follows:

$$[\Gamma]_t = [R]_{t\times m} \; [Q]_m \; [S]^T_{m\times t} \tag{17}$$

*The trace of an $m\times m$ matrix $[Q]$ is defined as $T[Q] = \sum_{i=1}^{m} q_i(i)$.

where

$$r_i(i) = \begin{cases} 1, & i \leq t \\ 0, & elsewhere \end{cases}$$
$$r_i(j) = 0, \quad i \neq j$$
$$s_i(j) = \begin{cases} 1, & i = i_j \,,\, j \leq t \\ 0, & elsewhere \end{cases}$$

i_j is the row of $[Q]_m$ in which the jth column maximum occurs

Note that $s_i(j)$ is the ijth element of $[S]_{t \times m}$ whereas $[S]^T_{m \times t}$ occurs in the product (17).

The elements of $[\Gamma]_t$ are given by

$$r_h(k) = \sum_j a_{jk} \; q_h(j) \tag{18}$$

where

$$a_{jk} = \begin{cases} 1, & \text{if the jth column maximum of } [Q]_m \text{ occurs in the kth row.} \\ 0, & \text{elsewhere.} \end{cases}$$

Thus

$$r_k(k) = \sum_{j'} a_{jk} \; q_k(j)$$

= (sum of column maxima of $[Q]_m$ that fell in the kth row)

and

$$r_h(k) \leq r_k(k) \leq 1, \quad h \neq k \tag{19}$$

Equation (19) proves that $[\Gamma]_t$ satisfies condition (15). Further, $\tau\,[\Gamma]$ = (Sum of column maxima of $[Q]_m$) so that (Sum of column maxima of $[Q]_m) \geq np$ is a sufficient condition for $[Q]_m \supseteq [P]_n$. To prove the necessity of (16), we show that for n, equally probable inputs $n\,\bar{p}(c) \leq$ (Sum of the column maxima of $[Q]_m$).

Since any stochastic channel can be built up from a set of pure channels, $\bar{p}(c)$ of the cascade can be no greater than that attainable with pure pre- and post-channels. With pure pre- and post-channels, if every output of $[Q]_m$ is connected to one output of the cascade, there is only one transition probability $q_i(j)$ for every j to result in a correct transition through the cascade. With all inputs equally likely, the probability of a correct transition involving $q_i(j)$ is $\frac{1}{n} q_i(j)$. The probability of any correct transition is $\frac{1}{n} \sum_{j=1}^{m} q_i(j) \leq \frac{1}{n} \sum_{j=1}^{m} q_{i_0}(j)$ where $q_{i_0}(j)$ is the jth column maximum.

Inclusion of an n-ary Symmetric Channel by an m-ary Asymmetric Channel for $n<m$

In this case, one can first convert the $[Q]_m$ channel into $[\Gamma]_t$ channel as previously described. If $t \leq n$, then the problem is equivalent to the $n \geq m$ case which has already been discussed and the necessary and sufficient condition for $[Q]_m \supseteq [P]_n$ when $t \leq n$ is (16).

As an example, consider $n = 3$ and $m = 4$ with

$$[Q]_4 = \begin{bmatrix} \enclose{circle}{.6} & 0 & 0 & \enclose{circle}{.4} \\ .2 & \enclose{circle}{.5} & 0 & .3 \\ .1 & .2 & \enclose{circle}{.4} & .3 \\ .3 & .4 & .3 & 0 \end{bmatrix} \tag{20}$$

where the column maxima have been encircled. Here $t = 3 = n$. By inspection, or by application of (18),

$$[\Gamma]_3 = \begin{bmatrix} 1 & 0 & 0 \\ .5 & .5 & 0 \\ .4 & .2 & .4 \end{bmatrix} \tag{21}$$

so that (16) requires $p \leq \frac{1.9}{3}$.

The matrices $[R]_{3\times4,}$ $[S]^T_{4\times3}$ take the forms

$$[R] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad [S]^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}^T$$

in accordance with (17).

The cascade $[R]\,[Q]\,[S]^T$ which is equivalent to the channel $[\Gamma]_3$ is sketched in Figure 4.

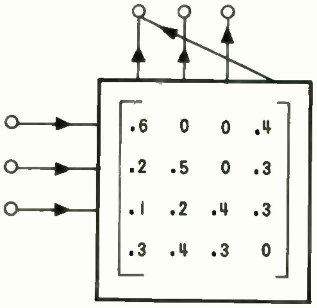


Figure 4

When $t > n$, a reduction of $[Q]_m$ to $[\Gamma]_t$ does not solve the problem. It is difficult to make meaningful statements of general validity for this case. However, the examination of specific cases can proceed readily in the following manner. For a given (pure) pre-channel, the best (pure) post-channel is the one which connects output j of $[Q]_m$ to that output of the cascade which corresponds to the largest $q_{\hat{i}}(j)$ where $\hat{i}$ are those of the inputs of $[Q]_m$ which are connected by the pre-channel. No simple manner of determining which of the $\binom{m}{n}$ possible pure pre-channels should be used has been found. However, the number of pre-channels which need be examined may be greatly reduced if the $[Q]_m$ channel has partial symmetries. Table 1 gives two numerical examples of this technique for the case $m = 4$, $t = 3$, $n = 2$. Example (a) has a unique best solution; while, in example (b), all possible pre-channels result in the same $\bar{p}(c)$. Note that the post-channel is not always uniquely specified by the pre-channel and $[Q]_m$.

TABLE I

	PRE CHANNEL CONNECTS TO INPUT	POST CHANNEL COMBINES OUTPUTS	$2\bar{p}(c)$
(a.) $[Q]_4 = \begin{bmatrix} .5 & .1 & .1 & .3 \\ .2 & .5 & 0 & .3 \\ .1 & .2 & .4 & .3 \\ .3 & .4 & .3 & 0 \end{bmatrix}$	1,2	1-3,2-4	.6 + .8 = 1.4
	1,3	1-4,2-3	.8 + .6 = 1.4
	1,4	1-4,2-3	.8 + .7 = 1.5 → OPTIMUM
	2,3	1-2,3-4	.7 + .7 = 1.4
	2,4	2-4,1-3	.8 + .6 = 1.4
	3,4	3-4,1-2	.7 + .7 = 1.4
(b.) $[Q]_4 = \begin{bmatrix} 5 & .2 & .1 & .2 \\ .2 & .5 & 0 & .3 \\ .1 & .2 & .4 & .3 \\ 3 & .4 & .3 & 0 \end{bmatrix}$	1,2	1-3,2-4	.6 + .8 = 1.4
	1,3	1,2-3-4	.5 + .9 = 1.4
	1,4	1-4,2-3	.7 + .7 = 1.4
	2,3	1-2-4,3	1.0 + .4 = 1.4
	2,4	2-4,1-3	.8 + .6 = 1.4
	3,4	3-4,1-2	.7 + .7 = 1.4

Inclusion on an Error Probability Basis

One may be interested in reproducing the probabilities of correct transition $p_i(i)$ only. A channel can reproduce all $p_i(j)$ of a symmetric channel which it includes; if only the $p_i(i)$ need be reproduced, the number of relabelings can be reduced.

Another application is to channels in which all $p_i(i)$ are equal, but all $p_i(j)$ are not equal (a practically important example is a channel which has as its inputs a set of $\frac{n}{2}$ orthogonal signals and their negatives).

The techniques described apply only to those n-ary channels in which $p_i(i) = p$ independent of i. If an m-ary channel includes an n-ary symmetric channel with probability of correct transition p, then it also includes on an error probability basis all n-ary (asymmetric) channels for which $p_i(i) = p' \leq p$.

Comparison of Binary Channels on the Basis of "Inclusion"

A binary channel is completely defined by the 2 x 2 matrix of its transition probabilities $P = \begin{bmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{bmatrix}$ Such a channel may be represented as a point in the unit square with coordinates (p_1, p_2), and the relationship of channel K to those channels which are "included" by it is shown in the unit square (Figure 5). The use of the pure pre-channel having matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and line diagram $\times$ results in a cascade having matrix $\begin{bmatrix} 1-p_2 & p_2 \\ p_1 & 1-p_1 \end{bmatrix}$, that is, a matrix in which the rows have been interchanged. The point representing this channel is plotted as K_1. The use of a pure post-channel having matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ results in interchange of columns $\begin{bmatrix} 1-p_1 & p_1 \\ p_2 & 1-p_2 \end{bmatrix}$ and representation by point K_2. The use of both the above pure pre- and post-channels results in interchange of both rows and columns $\begin{bmatrix} p_2 & 1-p_2 \\ 1-p_1 & p_1 \end{bmatrix}$ and representation by point K_3. The use of the pure post-channel having matrix $\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ results in over-all transition matrix $\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ represented by point K_4; similarly, the use of the pure post-channel having matrix $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ is represented by point K_5. The set of channels included by channel K (or K_1, K_2, K_3) is bounded by the hexagon $K\ K_5 K_1 K_2 K_4 K_3$. Any point on the boundary, such as point A, may be reached by using two cascades with appropriate probabilities, e.g., use K_3 with probability $g_3 = \frac{\overline{K_4 A}}{\overline{K_4 K_3}}$ and K_4 with probability $g_4 = \frac{\overline{A K_3}}{\overline{K_4 K_3}}$. Any point inside the hexagon, such as point B, may be reached by using three cascades, e.g., use K_5 with probability $g_5 = \frac{\overline{AB}}{\overline{AK_5}}$, use K_3 with probability $g_{3B} = g_A g_3 = \frac{\overline{BK_5}}{\overline{AK_5}} \cdot \frac{\overline{K_4 A}}{\overline{K_4 K_3}}$ and use K_4 with probability $g_{4B} = \frac{\overline{BK_5}}{\overline{AK_5}} \cdot \frac{\overline{AK_3}}{\overline{K_4 K_3}}$.

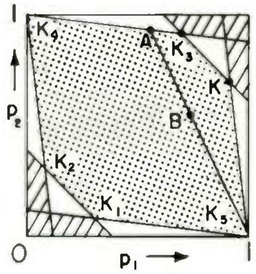


Figure 5

In Figure 5, the shaded regions represent channels which include channel K, the dotted hexagon represents channels included by K, and the clear areas represent channels which neither include K nor are included by K. These results were already obtained by Shannon.[1]

Comparison of Binary Channels on the Basis of "Not More Lossy".

A channel K_A is said to be "not more lossy" than another channel K_B if for all assignment of losses $L(\omega_i, a_j)$, corresponding to acceptance of the hypothesis that symbol ω_j was sent when symbol ω_i actually was sent, and all possible probability distributions of the input symbols the minimum expected loss when channel K_A is used is equal to or less than the minimum expected loss when channel K_B is used.

Before commencing with the actual analysis, it is worth noting that use of the 0 - 1 loss matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ results in an expected loss which is equal to the probability of error. It follows that if channel K_A is "not more lossy" than channel K_B, then the min. probability of error when channel K_A is used is not greater than the min. probability of error when channel K_B is used, regardless of the distribution of input symbols. We shall prove that the inverse of this statement also holds; i.e., if for all distributionsof inputs the min. probability of error when using channel K_A is not greater than when using channel K_B, then K_A is "not more lossy" than K_B.

It does not appear intuitively obvious that there exist channels which have the "not more lossy" relationship or that this relationship should be implied by the criterion based on error probabilities. Abramson[3] has proven the existence of such channels and derived conditions which must exist among them by the use of the theory of statistical decisions[4]. The system under investigation is shown in Figure 6.

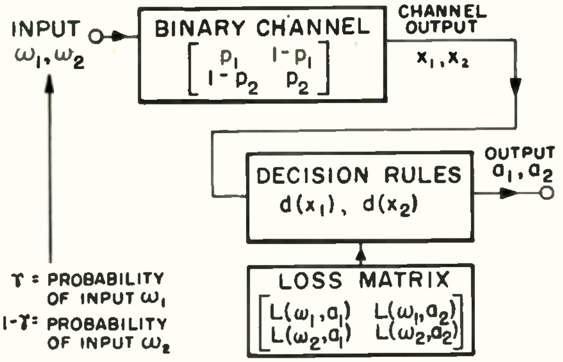


Figure 6

The decisions a_1 or a_2, to accept the hypothesis that the channel input was ω_1 or ω_2, are made upon observing the channel output x_1 or x_2 in accordance with decision rules $d(x_1)$ or $d(x_2)$. The Bayes decision rules are to be used, that is those which result in the smallest expected loss.

The decision rules take the form

$$d(x_1) = a_1 \text{ or } a_2$$
$$d(x_2) = a_1 \text{ or } a_2$$

The expected loss is $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, where $\mathcal{L}_i$ is the expected loss given that the channel output is x_i, $i = 1, 2$. Since the probability of the channel output x_i occurring is independent of the decision rules, one can optimize these rules independently.

$$\mathcal{L}_1 = \gamma p_1 L\{\omega_1, d(x_1)\} + (1-\gamma)(1-p_2) L\{\omega_2, d(x_1)\}$$

If $d(x_1) = a_1$, this yields

$$\mathcal{L}_{11} = \gamma p_1 L(\omega_1, a_1) + (1-\gamma)(1-p_2) L(\omega_2, a_1)$$

and if $d(x_1) = a_2$

$$\mathcal{L}_{12} = \gamma p_1 L(\omega_1, a_2) + (1-\gamma)(1-p_2) L(\omega_2, a_2)$$

One, therefore, chooses

$$d(x_1) = \begin{cases} a_1, \text{ if } \mathcal{L}_{11} < \mathcal{L}_{12} \\ a_2, \text{ if } \mathcal{L}_{11} \geq \mathcal{L}_{12} \end{cases}$$

The threshold of discrimination may be determined by setting $\mathcal{L}_{11} = \mathcal{L}_{12}$.

Thus,

$$\gamma p_1 L(\omega_1, a_1) + (1-\gamma)(1-p_2) L(\omega_2, a_1) = \gamma p_1 L(\omega_1, a_2) + (1-\gamma)(1-p_2) L(\omega_2, a_2),$$

which yields

$$\frac{p_1}{1-p_2} = \frac{1-\gamma}{\gamma} \; \frac{L(\omega_2, a_1) - L(\omega_2, a_2)}{L(\omega_1, a_2) - L(\omega_1, a_1)}$$

Note that the right-hand side is independent of the channel characteristics and the left-hand side is dependent only on the channel characteristics.

Similarly,

$$\mathcal{L}_2 = \gamma(1-p_1) L\{\omega_1, d(x_2)\} + (1-\gamma) p_2 L\{\omega_2, d(x_2)\}$$

If $d(x_2) = a_1$, this yields

$$\mathcal{L}_{21} = \gamma(1-p_1) L(\omega_1, a_1) + (1-\gamma) p_2 L(\omega_2, a_1)$$

and if $d(x_2) = a_2$,

$$\mathcal{L}_{22} = \gamma(1-p_1) L(\omega_1, a_2) + (1-\gamma) p_2 L(\omega_2, a_2)$$

Therefore,

$$d(x_2) = \begin{cases} a_1, \text{ if } \mathcal{L}_{21} < \mathcal{L}_{22} \\ a_2, \text{ if } \mathcal{L}_{21} \geq \mathcal{L}_{22} \end{cases}$$

Again, the threshold may be determined by setting $\mathcal{L}_{21} = \mathcal{L}_{22}$, yielding

$$\frac{1-p_1}{p_2} = \frac{1-\gamma}{\gamma} \; \frac{L(\omega_2, a_1) - L(\omega_2, a_2)}{L(\omega_1, a_2) - L(\omega_1, a_1)}$$

Define

$$\beta = \frac{1-\gamma}{\gamma} \cdot \frac{L(\omega_2, a_1) - L(\omega_2, a_2)}{L(\omega_1, a_2) - L(\omega_1, a_1)} = \frac{1-\gamma}{\gamma L}$$

Although it is not necessary, it is most reasonable to assign greater losses to wrong decisions than to right ones. In this case, $\beta \geq 0$, $L > 0$ and the Bayes decision rules are*

$$d(x_1) = \begin{matrix} a_1, \text{ if } \dfrac{p_1}{1-p_2} > \beta \\ a_2, \text{ if } \dfrac{p_1}{1-p_2} \leqslant \beta \end{matrix} \qquad d(x_2) = \begin{matrix} a_1, \text{ if } \dfrac{1-p_1}{p_2} > \beta \\ a_2, \text{ if } \dfrac{1-p_1}{p_2} \leqslant \beta \end{matrix}$$

and it is possible to write the loss matrix in the form $\begin{bmatrix} 0 & L \\ 1 & 0 \end{bmatrix}$. If $L > 0$ then a lossless channel** has a minimum expected loss of zero, and this is also the smallest expected loss attainable by any channel. The $0 \cdot 1$ loss matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ yields an expected loss numerically equal to the probability of error.

If one permits $L < 0$, then one obtains $d(x_1) = d(x_2) = a_i$, $i = 1$ or 2 depending on $L(\omega_1, a_2) >$ or $< L(\omega_1, a_1)$, as the Bayes decision rules. Thus, for all input distributions and all channels, the channel output is ignored in making the decisions and the expected loss becomes independent of channel properties. This situation appears to be of little interest, and we will assume that $0 < L < \infty$ henceforth.

The Bayes decision rules may be conveniently depicted on the unit square. The two lines AD and BC which represent the decision threshold have been drawn on Figure 7 for the case $\beta < 1$.

These lines divide the unit square into three regions with decision rules:

$\left.\begin{array}{l} d(x_1) = a_1 \\ d(x_2) = a_2 \end{array}\right\}$ for channels represented by points to the right of line BC, such as point X,

$\left.\begin{array}{l} d(x_1) = a_1 \\ d(x_2) = a_1 \end{array}\right\}$ for channels represented by points between lines BC and AD, such as point Y, and

$\left.\begin{array}{l} d(x_1) = a_2 \\ d(x_2) = a_1 \end{array}\right\}$ for channels represented by points to the left of line AD, such as point Z.

*If $L(\omega_1, a_2) < L(\omega_1, a_1)$ these decision rules are to be reversed.

**A lossless channel is a channel having zero equivocation for all input probability distributions (Reference 5, p. 26); the two lossless binary channels have $p_1 = p_2 = 1$ and $p_1 = p_2 = 0$, respectively.

The minimum expected Bayes losses corresponding to any channel can be found on the graph above the unit square by projecting the point representing the channel parallel to lines AD and BC. The losses $\mathcal{L}_X, \mathcal{L}_Y, \mathcal{L}_Z$ corresponding to points X, Y, Z are shown on Figure 7.

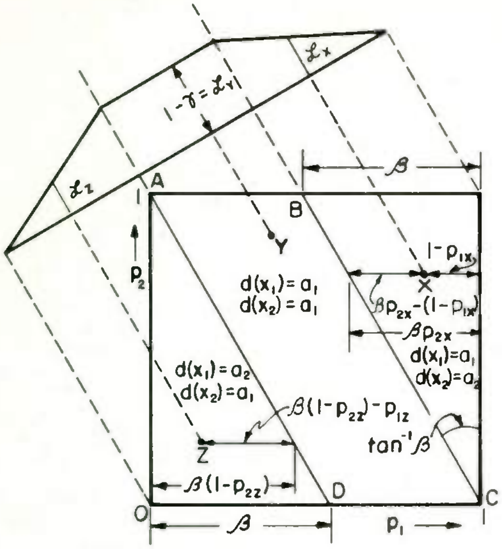


Figure 7

Any channel represented by a point in the parallelogram ABCD, such as point Y satisfies $\frac{p_{1Y}}{1-p_{2Y}} > \beta$ and $\frac{1-p_{1Y}}{p_{2Y}} > \beta$ so that the decision rules are $d(x_1) = d(x_2) = a_1$. That is, the hypothesis that the input ω_1 was sent is accepted regardless of the output of the channel. The expected loss is then $\mathcal{L}_Y = (1-\gamma)$ for such a channel.

For any channel represented by a point to the right of the line BC, such as point X,

$$\frac{p_{1X}}{1-p_{2X}} > \beta \ , \qquad \frac{1-p_{1X}}{p_{2X}} < \beta$$

Therefore, the Bayes decision rules are

$$d(x_1) = a_1, \ d(x_2) = a_2,$$

the expected loss is

$$\mathcal{L}_X = \mathcal{L}_{22} + \mathcal{L}_{11} = \gamma L\,(1-p_{1X}) + (1-\gamma)(1-p_{2X})$$

Substituting $\beta = \frac{1-\gamma}{\gamma L}$ yields

$$\mathcal{L}_X = (1-\gamma)\left[1 - \frac{1}{\beta}\left\{\beta p_{2X} - (1-p_{1X})\right\}\right]$$

$$= (1-\gamma)\left[1 - \frac{\sqrt{1+\beta^2}}{\beta}(\text{distance of point } X \text{ from line } BC)\right].$$

For any channel represented by a point to the left of the line AD, such as point Z,

$$\frac{p_{1Z}}{1-p_{2Z}} > \beta, \quad \frac{1-p_{1Z}}{p_{2Z}} < \beta$$

so that the Bayes decision rules are

$d(x_1) = a_2$ and $d(x_2) = a_1$, the expected loss is

$$\mathcal{L}_Z = \mathcal{L}_{12} + \mathcal{L}_{21} = \gamma L\, p_{1Z} + (1-\gamma)\, p_{2Z}$$

$$= (1-\gamma)\left[1 - \frac{1}{\beta}\left\{\beta(1-p_{2Z}) - p_{1Z}\right\}\right]$$

$$= (1-\gamma)\left[1 - \frac{\sqrt{1+\beta^2}}{\beta}\ (\text{distance of point } Z \text{ from line } AD)\right]$$

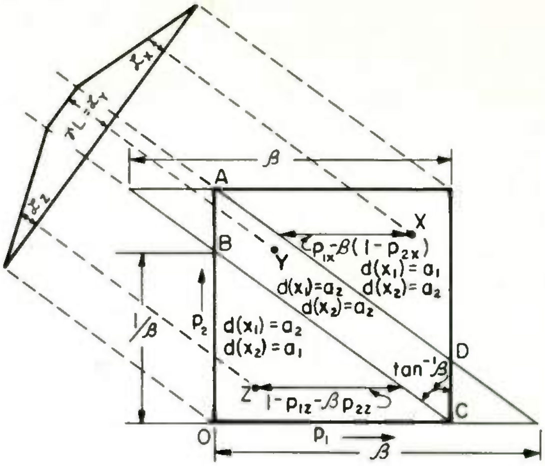


Figure 8

Figure 8 illustrates the three decision regions for the case $\beta > 1$. The decision rules for channels represented by points above the line AD, such as point X, are $d(x_1) = a_1, d(x_2) = a_2$; for channels represented by points between lines AD and BC, such as point Y, $d(x_1) = d(x_2) = a_2$; and for channels represented by points below the line BC, such as point Z, the decision rules are $d(x_1) = a_2, d(x_2) = a_1$. The Bayes losses corresponding to any channel can be found on the graph to the left of the unit square by projecting the point representing the channel parallel to lines AD and BC. The losses corresponding to points X, Y, Z are shown on Figure 8.

For any channel represented by a point in the parallelogram ABCD, such as point Y, $\frac{p_{1Y}}{1-p_{2Y}} < \beta$, $\frac{p_{1Y}}{1-p_{2Y}} < \beta$ so that the decision rules are $d(x_1) = d(x_2) = a_2$ The hypothesis that the input ω_2 was sent is accepted regardless of the output of the channel and the expected loss is $\mathcal{L}_Y = \gamma L$.

For any channel represented by a point above line AD, such as point X, $\frac{p_{1X}}{1-p_{2X}} > \beta$, $\frac{1-p_{1X}}{p_{2X}} < \beta$ so that the Bayes decision rules are $d(x_1) = a_1$, $d(x_2) = a_2$ and the expected loss is

$$\mathcal{L}_X = \mathcal{L}_{22} + \mathcal{L}_{11} = \gamma L(1-p_{1X}) + (1-\gamma)(1-p_{2X}),$$

which upon substituting $\beta\ \frac{1-\gamma}{\gamma L}$ becomes

$$\mathcal{L}_X = \gamma L\left[1 - \left\{p_{1X} - \beta(1-p_{2X})\right\}\right]$$

$$= \gamma L\left[1 - \sqrt{1+\beta^2}(\text{distance of point } X \text{ from line } AD)\right]$$

For any channel represented by a point below the line BC. such as point Z, $\frac{p_{1Z}}{1-p_{2Z}} < \beta$, $\frac{1-p_{1Z}}{p_{2Z}} > \beta$ so that the Bayes decision rules are $d(x_1) = a_2, d(x_2) = a_1$ and the expected loss is

$$\mathcal{L}_Z = \mathcal{L}_{12} + \mathcal{L}_{21} = \gamma L\, p_{1Z} + (1-\gamma)\, p_{2Z}$$

$$= \gamma L\left[1 - (1 - p_{1Z} - \beta p_{2Z})\right]$$ (con'td next page)

$= \gamma L \left[1 - \sqrt{1+\beta^2} \text{ (distance of point z from line } BC) \right]$

The "not more lossy" relationship of any channel to a given channel K can be determined most easily by plotting the point representing channel K in the unit square and dividing the unit square, as shown in Figure 9, by the solid lines. By the reasoning used to derive the loss relationships depicted in Figures 7 and 8, one concludes:

(a) that channel K is "not more lossy" than channels represented by points in the dotted parallelogram $AKCK_2$;

(b) that channels represented by points in the shaded regions are "not more lossy" than channel K;

(c) that points in the clear regions (marked I or II) represent channels which are not comparable by the "not more lossy" criterion with channel K (because for some values of β, the loss will be less than that of channel K, while for some other values of β, it will be greater than that of channel K).

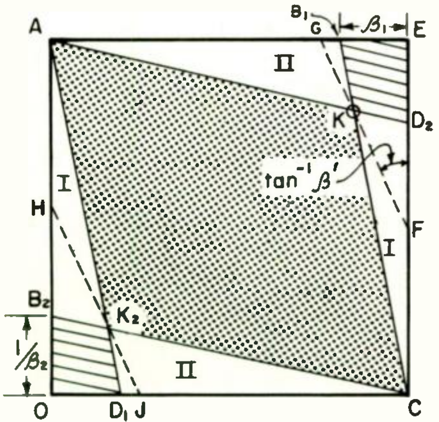


Figure 9

For any positive value of β the Bayes loss for channel K_2, which is the channel obtainable by interchanging the output connections of channel K, is equal to that for channel K. For $0 \leq \beta \leq \beta_1$, channels represented by points in the triangular areas CEB_1 and AOD_1 have losses which are not greater than that of channel K, while all other channels have loss $1-\gamma$. For $\beta_2 \leq \beta \leq \infty$, channels represented by points in the triangular areas COB_2 and AED_2 have losses which are not greater than that of channel K, while all other channels have loss γL. For $\beta_1 < \beta < \beta_2$, such as $\beta = \beta'$, channels represented by points in the two triangular areas EFG, OHJ, have losses not greater than the loss of channel K; while all other channels have losses which are not smaller than that of channel K.

It will be noted that the boundaries of the three regions in Figures 7 and 8 are determined solely by the value of β which can range over $0 \leq \beta \leq \infty$ for any $\infty > L > 0$. Therefore, the boundaries of the sets of points representing channels comparable with a given channel, depicted by solid lines in Figure 9, do not depend on the value of L. In particular, these boundaries are applicable to the case $L = 1$ in which case the expected loss equals the probability of error. It follows that K_A is "not more lossy" than channel K_B if and only if the minimum probability of error when using channel K_A is equal to or less than that when using channel K_B for all probability distributions of the inputs.

If K_A, K_B are two channels such that neither is "not more lossy" than the other, then there will exist a channel K_I which is the greatest lower bound of the set of channels which are "not more lossy" than channels K_A and K_B and a channel K_{II} which is the least upper bound of the set of channels, such that K_A is "not more lossy" than K_{II}, and K_B is "not more lossy" than K_{II}. This situation is illustrated in Figure 10. Thus, the "not more lossy" criterion has the properties of a lattice.[6]

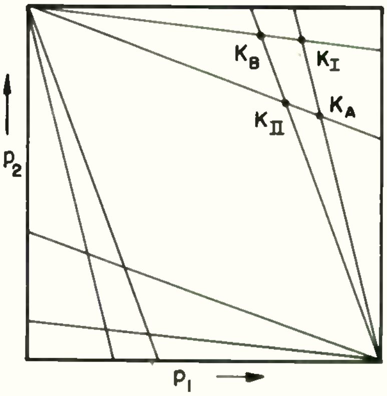


Figure 10

Comparison of the "Inclusion" and "Not More Lossy" Criteria

Comparison of Figures 5 and 9 shows that K_A "not more lossy" than K_B is a sufficient condition for K_A includes K_B. With reference to Figure 5, all the channels represented by points inside the hexagon $K_5 K_1 K_2 K_4 K_3 K$ are included by channel K (or K_1, K_2, K_3). According to Figure 7, for $0 \leq \beta \leq 1$, the loss of channel K (or K_2) is not greater than that of any channel represented by a point inside the hexagon; and, in accordance with Figure 8, for $1 \leq \beta \leq \infty$, the loss of channel K_1 (of K_3) is not greater than that of any channel inside the hexagon. Therefore, designating $L(B)$ as the Bayes loss of any channel B, we can state:

If K includes A, then $L(A) \geq \min. \left\{ L(K), L(K_1) \right\}$

where channel K_1 is obtainable by interchanging the input connections of channel K.

By considering the position of points representing two channels, such that neither includes the other, it is easily seen that the "inclusion" criterion also has the lattice property.[1]

Comparison of Binary Channels on the Basis of "More Informative"

It is important that the criterion described by the term "more informative" be clearly kept in mind, as there exists considerable literature in which the term "more informative" has various meanings.[3,4,7] In particular, the terminology "more informative" is used in Reference 3 to describe the "not more lossy" criterion as used here. The reason that the terminology "not more lossy" rather than "more informative" seems more appropriate to describe that criterion is as follows. For any pair of channels K_A and K_B represented by points K_A, K_B as in Figure 10, there exist two ranges of β such that with either K_A or K_B the smallest loss is attained by ignoring the channel output in making decisions, i. e., $d(x_1) = d(x_2) = a_1$ or a_2. Under these conditions, the information about the channel input which becomes available at the channel output (and which is always greater than zero on the average)[7,8] is ignored in making these designations.

We shall use the following definition of "more informative". Channel K_A is "more informative" than channel K_B if for all probability distributions of the inputs the rate of information transfer when using channel K_A equals or exceeds the rate of information transfer when using channel K_B.

The rate of information transfer for any channel is given by:

$$\begin{aligned} R &= H(\Omega)-H(\Omega/X) = \text{(Entropy of Input) - (Equivocation)} \\ &= H(X)-H(X/\Omega) = \text{(Entropy of Output) - (Conditional Entropy of Output given the Input)} \\ &= \sum_{\omega}\sum_{x} p(\omega,x) \log \frac{p(\omega,x)}{p(\omega)p(x)} \quad \text{for discrete channels} \\ &= \int\int p(\omega,x) \log \frac{p(\omega,x)}{p(\omega)p(x)}\, d\omega\, dx \quad \text{for continuous channels} \end{aligned}$$

where ω represents the channel input and X the channel output.

It has not been possible to derive the necessary and sufficient conditions under which an asymmetric binary channel is "more informative" than another such channel. However, the boundaries in the unit square of a subset of all channels comparable with a given channel by the "more informative" criterion are known. The actual boundaries in the unit square have been obtained by machine computation for an illustrative case.

Considerable insight into this problem may be obtained by the following geometric approach conceived by Shannon.[9] (The approach is also valid for n-ary discrete channels in which case the use of n dimensional geometry is required.)

For a binary channel, the entropy of the output, $H(X)$, and the conditional entropy of the output given the input, $H(X/\Omega)$, appearing in the expression for the information rate $R = H(X)-H(X/\Omega)$ are:

$$H(x) = \sum_{j=1}^{2} -p(x_j) \log p(x_j)$$

$$H(x|\Omega) = -\sum_{i=1}^{2} p(\omega_i) \sum_{j=1}^{2} p(x_j|\omega_i) \log p(x_j|\omega_i)$$

Using $p(x_1|\omega_1) = p_1$, $p(x_2|\omega_2) = p_2$, $p(\omega_1) = \gamma$, etc., and

$$p(x_j) = \sum_{i=1}^{2} p(\omega_i)\, p(x_j|\omega_i), \text{ and defining } d = p_1-(1-p_2),$$

$$H(x) = -(1-p_2+\gamma d) \log (1-p_2+\gamma d) - (p_2-\gamma d) \log (p_2-\gamma d)$$

$$H(x|\Omega) = \gamma\left\{-p_1 \log p_1 - (1-p_1) \log (1-p_1)\right\} + (1-\gamma)\left\{-(1-p_2) \log (1-p_2) - p_2 \log p_2\right\}$$

Define $H_2(\xi)$, for $0 \le \xi \le 1$, by

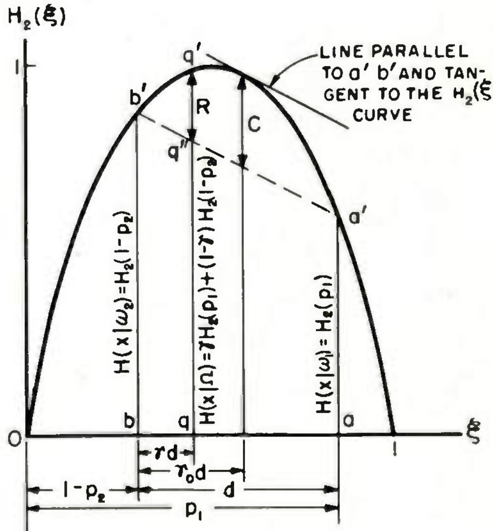
$$H_2(\xi) = -\xi \log \xi - (1-\xi) \log (1-\xi) = H_2(1-\xi)$$

which is the dome-shaped curve plotted in Figure 11. Then

$$H(X) = H_2(1-p_2+\gamma_d) = \overline{qq'}$$

$$H(X|\Omega) = \gamma H_2(p_1) + (1-\gamma) H_2(1-p_2) = \overline{qq''}$$

and, hence, the information rate, R, is equal to $q'q''$ of Figure 11. (The capacity C which is defined as the greatest R attainable by variation of γ is also shown in Figure 11.)



H₂(ξ)
1
q′
LINE PARALLEL TO a′ b′ AND TANGENT TO THE H₂(ξ) CURVE
b′
R
C
q″
a′
H(x|ω₂)=H₂(1-p₂)
H(x|Ω)=γH₂(p₁)+(1-γ)H₂(1-p₂)
H(x|ω₁)=H₂(p₁)
O
b
q
a
1
ξ
γd
γ₀d
1-p₂
d
p₁

Figure 11

There exist three regions such that channels represented by points located in region I are less informative than a comparison channel K_A, channels represented by points in region II are more informative than channel K_A, and points located in region III represent channels not comparable to channel K_A by the "more informative" criterion. Proof: Let channel K_A be represented by a point not on the diagonals, then channel K_B with $p_{1B} = 1$, $p_{2B} = 0$ is in region I since $R_B = 0$. Channel K_C with $p_{1C} = p_{2C} = 1$ is in region II since its equivocation is zero (Rate = Entropy of Input - Equivocation), and channel K_D with $p_{1D} = p_{2A}$, $p_{2D} = p_{1A}$ in region III since $R_A \neq R_D$ if $\gamma \neq 1/2$ and it $R_A > R_D$ for $\gamma > 1/2$ then $R_A < R_D$ if $\gamma < 1/2$ and vice versa.

If K_A is "not more lossy" than K_B, then K_A is "more informative" than K_B. Proof: It can be shown (e. g., Reference 7, Theorem 6) that, for any γ, the rate of information is a convex function of the transition probabilities, i. e., if K_A, K_B, K_c are three channels, such that

$$p_{1B} = \lambda p_{1A} + (1-\lambda) p_{1c},$$
$$p_{2B} = \lambda p_{2A} + (1-\lambda) p_{2c},$$
$$0 \leq \lambda \leq 1,$$

then $R_B \leq \lambda R_A + (1-\lambda) R_c$. For any point B on the boundary of the set of points representing channels "not more lossy" than channel K_A, channel K_c can have $p_{1c} = 1, p_{2c} = 0$ (or $p_{1c} = 0, p_{2c} = 1$) so that $R_c = 0$. (For points on the other side of the $(1,0)-(0,1)$diagonal, K_A can be replaced by K_A' which has $R_A = R_A'$ for all γ, where $p_{1A}' = 1-p_{1A}, p_{2A}' = 1-p_{2A}$.) Therefore $R_B \leq R_A$ for all γ; the proof may easily be extended to the interior of the boundary.

This theorem establishes that the channels which are "not more lossy" than K_A form a subset of the channels which are "more informative" than K_A. And the set of channels such that K_A is "not more lossy" than any channel contained in the set forms a subset of the channels such that K_A is "more informative" than any channel in the subset.

The capacity of a binary channel is given by

$$C(p_1, p_2) = \log \left[\exp \frac{(p_2-1) H_2(p_1) + p_1 H_2(p_2)}{1-p_1-p_2} + \exp \frac{(p_1-1) H_2(p_2) + p_2 H_2(p_1)}{1-p_1-p_2} \right]$$

This expression has the symmetries

$$C(p_1, p_2) = C(p_2, p_1) = C(1-p_1, 1-p_2) = C(1-p_2, 1-p_1)$$

corresponding to the original channel and channels obtainable from it by interchange of both input and output, output, and input connections.

As indicated in Figure 11, a particular distribution of inputs γ_0, $1-\gamma_0$ must be used in order to achieve an information rate equal to the channel capacity. It is found that[2] if a line of constant capacity is drawn through a point representing channel K_A, then there is at most a single point on that line which corresponds to a given value of γ_0. It follows that the line of constant capacity lies entirely in region III representing channels which are not comparable with K_A by the "more informative" criterion.

Figure 12 illustrates the relations discussed above, the point K_A has coordinates $p_{1A} = .85, p_{2A} = .6$. Region I representing channels which are "less informative" than K_A is dotted, region II representing channels which are "more informative" than channel K_A is shaded, and region III representing channels which are not comparable by the "more informative" criterion with channel K_A has been left blank. The linesof constant capacity (C = .165 bits = capacity of channel K_A)are also shown. (Note that these lines lie in region III.) The straight line boundaries of the regions representing channels which may be compared with K_A by the "not more lossy" criterion are also shown for comparison.

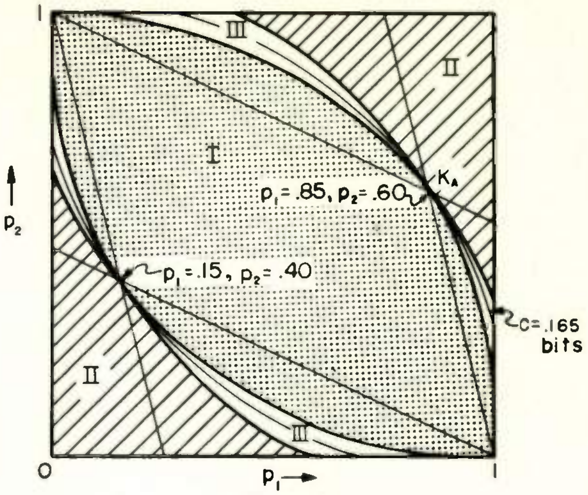


Figure 12

REFERENCES

1. Shannon, C.E. A Note on a Partial Ordering for Communication Channels Information and Control, Vol. 1, 1958.

2. Silverman, R.A. and Chang, Sze-Hou Topics in the Theory of Discrete Information Channels New York University Institute of Mathematical Sciences Research Report No. EM-152 April 1960.

3. Abramson, N.M. A Partial Ordering for Binary Channels Stanford Electronics Laboratories Technical Report No. 2001-1 15 April 1960.

4. Blackwell, D. and Girshick, M.A. Theory of Games and Statistical Decisions John Wiley & Sons New York 1954.

5. Feinstein, A. Foundations of Information Theory McGraw-Hill New York 1958.

6. Birkhoff, G. and MacLane, S. A Survey of Modern Algebra MacMillan Company New York 1958.

7. Lindley, D.V. On a Measure of the Information Provided by an Experiment The Annals of Mathematical Statistics, Vol. 27 December 1956.

8. Woodward, P.M. Probability and Information Theory, With Applications to Radar Pergamon Press London 1957.

9. Shannon, C.E. Geometrische Dentung einiger Ergebnisse bei der Berechnung der Kanalkapazitat Nachrichtentechnische Zeitschrift Jahrgang 10. Heft 1. January 1957.