1962 IRE International Convention Record

PART 9

Sessions Sponsored by

IRE Professional Groups on

Bio-Medical Electronics

Human Factors in Electronics

Instrumentation

Nuclear Science

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 9 - BIO-MEDICAL ELECTRONICS; HUMAN FACTORS IN ELECTRONICS; INSTRUMENTATION; NUCLEAR SCIENCE

TABLE OF CONTENTS

Page

Space Age Requirements for Electronic Measurements
(Session 2: sponsored by PGI)

Nuclear Instrumentation
(Session 19: sponsored by PGNS)

(Session 27: sponsored by PGNS)

Bio-Medical Electronics
(Session 32: sponsored by PGBME)

Page

Panel: Engineering in the Life Sciences, a Growing Challenge
(Session 39: sponsored by PGBME)

Instrumentation Systems and Devices
(Session 40: sponsored by PGI)

Human Factors in Electronics
(Session 44: sponsored by PGHFE)

SPACE PROGRAM REQUIREMENTS FOR IMPROVED ELECTRONIC MEASUREMENTS

Frank E. Goddard
Jet Propulsion Lab.
Pasadena, Calif.

Abstract

After a period of extensive study and planning the National Aeronautics and Space Administration is launching an expanded space program which should lead to the landing of the first American on the moon about 1967. Other NASA programs include global systems of communication and weather satellites. In all of these programs the total cost, the safety of human life, and the prestige of the United States are strongly dependent upon the reliability of electronic systems. Achieving the necessary reliability will require the development of many types of improved instrumentation, measurement standards, and measurement techniques. This paper points out some areas in which increased emphasis is required.

ABSOLUTE AND RELATIVE MEASUREMENTS

A. G. McNish
Natl. Bur. of Standards
Washington, D. C.

Abstract

Our scientific measuring system is based on the adoption of units and standards for a few elemental quantities and the measuring of other quantities in terms of these in accordance with established physical laws. Such are called absolute measurements. As experimental procedures become highly complicated absolute measurements become inaccurate, as in measurement of microwave power in terms of length, mass, and time. Relative measurement of one microwave power source in terms of another may be performed with much greater precision. If one of these sources is arbitrarily adopted as a standard, then, by definition, the other source may be calibrated in terms of it with high accuracy. Care must be observed in adoption of such arbitrary standards in order to assure stability of each standard and that all measurements of the quantity for which it is a standard are traceable to it.

The Dilemma of Measurement

Martin A. Mason
Dean, School of Engineering
The George Washington University

Summary

The conditions contributing to the "measurement gap" are reviewed briefly. The experiences of the Center for Measurement Science in its efforts to contribute to improvement of our measurement position are described. It is suggested that the dilemma of measurement can be resolved only if technical people convince others of the implications and consequences of our measurement deficiencies.

It is characterististic of our times that the pace of life is accelerating. Almost no field of human endeavor is immune; in some fields the pace seems to be explosive. This is the case in technology, where political, economic, and other pressures have given it importance far beyond that attaching in earlier times. Today a nation's science and technology are the determining factors in its very existence; and therefore any situation jeopardizing the technological stature of a people can be considered as most dangerous.

In this environment we find common agreement among technical people in the United States that there is a serious discrepancy between what we are doing in measurement and what is required to be done. There is feeling that this "measurement gap" must be closed as rapidly as possible. The Instrument Society of America's special journal reprint Measurement Standards Report states that "This must be the concern of every researcher, educator, manufacturer and user of industrial and scientific products".

Examples of the gap abound. Secretary of Commerce Hodges, writing in Industrial Research in December 1961, states that the standards and calibration services of the National Bureau of Standards satisfy only half of the needs of its users. He states further "The behavior of extremely hot gases is so singular that many scientists consider plasma the fourth state of matter. Yet we simply do not have the basic scientific tools required to measure plasmas". It is well known in technical circles that inability to precisely measure rocket thrust is costing the nation hundreds of millions of dollars. Secretary Hodges makes the point that military and industrial requirements are growing so fast, that even if current efforts in the field of radio standards were doubled immediately, new demands would in two years place us back in our current deficient position.

Several studies and surveys of the measurement situation have been made by government, the Aerospace Industries Association, and others. The results can be summed up as showing that technological development is being retarded because of needs for more:

1) accurate data on which to base the engineering design of complex equipment;

2) accurate monitoring and control of manufacturing processes to assure realization of potential and necessary reliability;

3) extension of precise measurement to new areas and to values formerly considered extreme

4) reliable traceability of standards to a common national standard.

Among all the information of various sort collected for these studies there appears a contінously recurring three part set, viz. capable personnel, improved and expanded facilities, and research. Studies by the Center for Measurement Science of The George Washington University have led to the conclusion that there is critical need now for greatly increased numbers of competent measurement personnel; for the development of improved and new standards of measurement and precision measuring equipment; for the development of procedures and techniques of precise measurement in frontier areas of technology and science; and for expansion of improved precise calibration facilities.

The measurement gap shows no signs of closing, rather it is widening. Thoughtful people concerned with measurement problems are particularly distressed with our inability to make the measurements required if our technology is to progress. These areas of inability (perhaps incompetence is a better designation) are growing. Not long ago our chief area of deficiency was precise large force measurement, say a million pounds force, today the deficiency continues and our need has extended to several million pounds force, with ten million pounds or more in sight as a requirement. Yesterday we were little concerned with extremely hot gases, today we need precise measurements of characteristics and properties we do not know how to measure. In microwave power measurements we are not only not sure of what we can measure, we are not reasonably certain of how to go about measuring what we need to know.

It must be recognized that more than measuring tools are required. We at the Center for Measurement Science have talked with many persons responsible for measurement matters in government and in industry, with those whose job it is to make measurements, and with management persons whose concern with measurement is simply that it serves needs efficiently and economically. In all groups there appeared to be most concern for the tools of measurement, the instruments, techniques, and procedures. The brains to guide and control the tools seemed to be of less concern, yet it is admitted that competent people are the key to the problem of closing the measurement gap. With very few exceptions managements stated their most difficult problem was competent people, but almost none accept the idea management had a responsibility to do anything about increasing the supply of people.

Adequate numbers of competent measurement specialists and of technicians are absolutely essential. There is little agreement on what is meant by "adequate""competent""technician and "specialist", as we at the Center for Measurement Science have discovered. Usually people such as you and I see little merit in quibbling about word meanings. In this instance what is meant by the words is all important, for the interpretations point up one of the serious aspects of the measurement dilemma.

On the one hand is the belief held by management (and many technical people) that measurement is a low-level activity which can be performed cheaply yet adequately by technicians trained for a few weeks in specific procedures; on the other is the reality in many instances of exceedingly difficult and complex measurement needs requiring high level technical capability for their satisfaction. There seems to be little understanding between these groups as to when the technician's capability is not enough, or the measurement specialist's capability may be too much. In this atmosphere of lack of common understanding there is considerable evidence that the measurement problem is handled on the basis of getting it done with what is at hand, rather than on the basis of what is technically and economically best or required.

Measurement in American technology has been an art traditionally. When a measurement need developed the problem was given to someone who knew and worked with measuring instruments, and he rigged some sort of scheme, many times most ingenious, to obtain some data. No one except himself worried much about the accuracy of the data so long as some reasonably acceptable answer was obtained. This has been an excellent scheme because the men given the responsibility were conscientious, careful, and competent within the demands made on them.

Today it is no longer permissible to rely on the art of measurement. Requirements have exploded into realms in which an art has not had time to develop, into regions of precision beyond the capabilities of existing instrumentation, and into fields on the frontiers of scientific knowledge. Our formerly competent measurement people have not lost their competence, but we now have requirements for competence they have not had opportunity to develop.

The problems that used to be found only in the laboratories are now often the commonplace problems of precision measurement in the production complex. The problems of measurement in the research laboratories are now often exceedingly difficult and demanding research problems in themselves. It is as though almost overnight all yardsticks became deficient and were replaced with interferometers. None would have lost his yardstick competence but very few satisfactory measurements would be made until the new knowledge and skills needed to use interferometers had been developed. This seems an absurd illustration, yet it is not far from describing what has happened and is happening in measurement.

Not only have our demands for precision exceeded our capabilities,in more instances than not these demands are coupled with extension of our measuring operations into levels several orders more complex than formerly.

Undoubtedly, universities and colleges have been slow to undertake education in measurement. There are many valid reasons for their being reluctant: one of the strongest has been the low intellectual demands of measurement. Measurement has not been and is nct now generally recognized as a first-class professional occupation. The test engineer is the low man on the professional totem pole, a less than creative detail man who engages in dull routine.

A student demand for measurement education has not and does not exist. True, there have been job opportunities for competent measurement technicians, but there is likewise a demand for stock and bond salesmen; in the student view it's easier and cheaper to become a salesman and the rewards may be much higher. The average prospective student finds little in the rewards of measurement capability to justify the rigors of the professional education required to attain capability.

The elements of the dilemma of measurement now become clear to us. First, there is incontrovertible need, now and in the future, in both research and industrial activity, for persons educated in measurement science. Second, the knowledge required by these persons relates to many areas of science and technology, and is of relatively high intellectual level. Third, there is very little attraction in the current status of measurement activity to draw new minds to the rigorous preparation required to fill the measurement need. Fourth, unless and until new minds are prepared to fill the need American technology will be hamstrung in its efforts to meet world competition. Fifth, the time for talk is long past; now only fast, extensive, and vigorous action by those having needs can salvage the situation.

Obviously this dilemma is similar to a classic dilemma, in that consequences can be fatal to American industry and possibly to the American nation. It need not be so. It will not be so if those having needs, both government and industry, take the feasible and necessary steps to remedy the situation.

The program of the Center for Measurement Science at The George Washington University is an example of a constructive effort to resolve the dilemma. The Center program has three principal activities; they are, the education and training of measurement specialists, at all the levels of competence required, from technician to research doctorates; research in measurement science as a concomitant part of the education of measurement specialists and to increase knowledge in the field; and development of a staff of measurement consultants available to those having temporary needs for highly specialized measurement knowledge and capability.

The education activity was started in February 1961, with the assistance of the National Bureau of Standards and the Martin Company - Baltimore. Student response was small. One of the unexpected developments was that many students having interest were deficient in fundamental preparation in mathematics, probability and statistics, and the engineering sciences. Of the forty odd initial students a few more than half either have not continued their measurement studies or are now undertaking basic studies to improve their preparation. Courses in measurement theory and in electrical measurements have been most popular.

It is too early to arrive at any conclusions about education for measurement from experience with this program. However evidence to date is that no large number of students is likely to undertake this program without substantial encouragement and support. There is no evidence that the top level student is likely to be attracted to education for measurement under existing conditions.

Obviously many factors can and do influence student response to this program. It is noted that the curriculum in the Center's program, which is one important factor, differs from the established engineering curricula of the School only in the approximately 36 semester hours study in measurement science. The faculty for the program are about divided between regular University faculty and adjunct faculty who are National Bureau of Standards staff members. Courses are offered during both day and evening hours, but exclusively on-campus or at the National Bureau of Standards. The influence of these factors is believed to be favorable to developing student interest.

The Center for Measurement Science education program has included also non-credit short courses of one or two weeks intensive study. Two were given in August 1961, one in Foundations of Precise Measurements, and one in Precise Electrical Measurements. Forty-two students attended the first; nineteen attended the second. The attendees were largely from east of the Mississippi and divided about equally from government agencies and industry. Although the courses were quite elementary in terms of collegiate level, deficiencies in preparation and variance of interests presented some difficulties.

Certainly no valid generalizations should be drawn from these meagre experiences. It does appear feasible however to surmise that:

1) Substantial encouragement and support of students in education for measurement science will most likely be required, if there are to be sufficient numbers of specialists produced by this means to help alleviate the existing and growing shortage of competent personnel.

2) The possibility of rapid training of personnel in intensive short courses is good at very elementary levels. Training at higher levels (above technician, say one year beyond high school) of current measurement personnel is likely to be quite difficult because of problems of preparation and interest patterns or attitudes.

The Center's programs in research and consultation are just starting. Their future depends upon the financial support which can be obtained from government and industry. We are firmly of the opinion that financial support and sharing of responsibility for action to improve the nation's measurement position must come from industry and government rather than primarily from university sources. The needs and the benefits derived from satisfaction of the needs are most intimately related to industrial activity and the associated profit to be derived. If the needs are as real and critical as they appear to be their satisfaction should justify considerable expenditures. The question then cannot be whether or not financial support will be given, it has to be what specific activity will receive financial support.

The experience of the Center in its quest for financial support is revealing. There has been widespread interest in industry in the Center's program, an avid interest in the names and addresses of students (for obvious reasons), a universal commendation of the Center's plans; and in one year of campaigning for financial support, a single contribution of funds from an industrial concern. Much better results have been obtained from manufacturers of measuring equipment, many leaders in their fields having contributed to the Center's support.

It is recognized that many factors influence financial contributions by industrial organizations to such activities as the Center for Measurement Science. Perhaps we at the Center have not been successful in establishing understanding of what the Center is and what it proposes to do. On the other hand we are much impressed by the high concern of the technical personnel in industry with the measurement gap, as compared to the low concern or total lack of concern, of industrial management with the problem. It is probably this latter attitude that accounts for much of the lack of interest in financial support.

Evidence is again too meager to generalize, yet it seems adequate to indicate that genuine concern about measurement is limited at this time to technical and professional personnel in government and industry. Concern sufficient to be translated into funds apparently does not yet exist in any appreciable degree among industry's management personnel or government appropriations committees. The real and vital problems seen by the scientists and engineers have not been appreciated by management sufficiently to result in necessary funds being made available.

I think this is the heart of the dilemma of measurement. Professional concern for the state of measurement has not been made known in whatever ways are necessary to result in the support and the funds necessary for remedial actions to be undertaken becoming available, from either industry or government. We professionals have the responsibility to translate into terms understandable by others, the consequences and implications of the measurement gap. I am confident that until we do so there is little real hope of improving the situation; if we are successful in doing so the dilemma of measurement as we see it now will disappear.

SOME PROBLEMS OF IMPROVING ACCURACY OF MEASUREMENT

A. V. Astin, Director
National Bureau of Standards
Washington 25, D. C.

The best information we have concerning the properties of matter and natural phenomena is quantitative information derived from measurement. The usefulness and reliability of such information increases with our ability to state it more accurately. This characteristic applies both to the progress of science and to progress in the application of scientific knowledge in a technological economy. In space age technology, and specifically in space age electronics, we find that the problems of accuracy of measurement are of critical importance. Their solution requires knowledge and techniques of unusual sophistication.

It is of interest in a symposium devoted to the subject of Space Age Requirements for Electronic Measurements to review briefly the evolution of some of our present concepts involving the role of measurement in the progress of science and technology.

Approximately 80 years ago Lord Kelvin made a classic and frequently quoted statement concerning the role of measurement. In essence Kelvin stated that you have to be able to measure what you are talking about and express it in numbers if the subject is to be classified as science. Although there is much truth in Kelvin's assertion it can lead to an oversimplification of the role of measurement. It may lead to a failure to appreciate the many pitfalls that can accompany measurement processes that are inadequately understood and applied.

Some of Lord Kelvin's contemporaries went much further than he in attributing an almost infallible role to physical measurement. For example, the famous American physicist Michelson, in a speech delivered at a convocation of the University of Chicago in 1894, said "It is never safe to affirm that the future of physical science has no marvels in store which may be even more astonishing than those of the past but it seems probable that most of the underlying principles have now been firmly established and that further advances are to be sought chiefly in the rigorous applications of these principles to all the phenomena which comes under our notice."

A little later in his talk Michelson went on to say: "An eminent physicist has remarked that the future truths of physical science are to be looked for in the sixth place of decimals." I have been unable to identify the eminent physicist referred to by Michelson but the viewpoint expressed appears to have been widely representative of a large group of physical scientists in the latter part of the 19th century.

Such smugness was of course rudely shocked at the turn of the century. Momentous experimental discoveries and the advent of the special theory of relativity and of the quantum theory completely disrupted these complacent points of view. But here it is important to emphasize that these revolutionary theoretical concepts were stimulated to an appreciable extent in an effort to explain phenomena whose baffling properties had been reliably described through painstakingly accurate physical measurement. One of the phenomena which had to be explained was the result of Michelson's own beautifully precise ether drift experiment.

Although it is now firmly established that there is much more to science than measurement alone, the broadened understanding of the progressive nature of science brought also a broadened understanding of the basic and progressive role of measurement.

A major consequence of both the special theory of relativity and the quantum theory was to add an important new dimension to physical measurement. The new theories related, as a general and fundamental principle, the process of measurement itself to the phenomena being measured. Consequently, it has become unrealistic to think of a measured property without consideration also of the act of measurement.

A basic element in the special theory of relativity is that one cannot adequately describe an event unless consideration is given to the transit time of the signal by means of which the event is observed or its properties measured. Of course in countless situations the transit time can be completely ignored in comparison with other uncertainties in an observation. But if one becomes realistic about the measurement process then the need for signals of some sort between the measured object and the observer has to be admitted together with the conclusion that in certain situations the transit time of those signals may be significant in comparison with other properties of the phenomena under observation. The consequences of this simple admission are far-reaching but from the point of view of meaningful precision measurement it is merely essential to recognize that one of the factors that may unavoidably influence the value of a particular measured property is the transit time of the signals essential to the measurement process.

The quantum theory also relates measured values to the measurement process but in quite a different way. Here we must recognize that measurement signals or measurement probes possess finite rather than infinitesimal levels of energy and that at least one quantum of energy must be involved in any measurement. Thus in observing processes where the energy levels involved approach the limits necessary for observation serious roadblocks develop. In the situation of attempting to observe electrons in atoms the quantum of energy necessary to react with an electron for observational purposes must have sufficient energy to modify significantly the properties of the electron in the atom. This situation seriously limits the nature of the observations that can be made upon atoms. It led to the formulation of the Heisenberg uncertainty principle which relates necessary uncertainties in the simultaneous measurement of position and momentum (or velocity) to the limiting quantum of action essential to an observation. This conclusion also has far-reaching implications but from the point of view of precision measurement it mainly means that in some cases the accuracy and interpretation of an observation may be unavoidably affected by the level of a single quantum of energy.

From the overall viewpoint of precision metrology the quantum uncertainty principle was probably responsible for a better appreciation of the need to consider in any situation the possible effects of the measuring process upon the values being measured. It is now a basic principle of metrology that any measured value may be influenced by the act of measurement. This overall need to be realistic about the possible interaction of the measurement system and process upon the value being measured has sometimes erroneously been called the uncertainty principle. It differs primarily from the quantum uncertainty principle in that in the general case one can usually, through proper design of the experiment or evaluation of the data, bring about significant reductions in the interaction of the measurement process upon the property being measured. The Heisenberg uncertainty principle, on the other hand, specifies an irreducible minimum, regardless of the degree of ingenuity in the design of the experiment, for the combined uncertainties in the measurement of position and momentum.

Another important impact of the quantum theory upon the science of measurement stems from the statistical or probability aspects of the theory. Here we are forced to recognize that many natural properties can be specified only in statistical terms. Such properties cannot be described by absolutely precise values but there will always be an associated degree of uncertainty. The degree of uncertainty can in many situations be made vanishingly small if the measurement can be made on a sufficiently large sample of similar objects but the possible existence of a statistical uncertainty must enter into an evaluation of the results of measurement.

Just as the Heisenberg uncertainty aspect of the quantum theory had a broadening impact on the concept of the measurement process so does the statistical aspect. Statistical considerations and probability estimates are involved not only in the values of ultimate properties but also the behavior of the observer and in the links between the observer and the object observed.

The process of measurement involves men or observers, instruments and equipment, objects and events to be measured, a host of environmental conditions, energy, time, and a system of interactions among the foregoing factors. Any or all of these may influence a measured value.

By systematic alteration of some of the test conditions, estimates can sometimes be made of the influence of a particular factor upon the measured value. Or by using prior knowledge of the characteristics of properties of some of the test parameters reasonable estimates may be made of their influence upon a measured value. The straightforward repetition of one observation under as near identical conditions as feasible may be used to reduce the so-called random errors. Such techniques can be employed to reduce the magnitude of the uncertainties in the observed values of particular properties.

Thus we find that the business of extending precision or accuracy in physical measurement is one of removing or reducing several layers of fuzziness or uncertainty leaving perhaps a residue of uncertainty which may be an essential characteristic of the object whose properties are under observation.

In recent decades there have been significant developments in statistical techniques both in reducing the effects of random errors associated with the measurement process and of systematic errors associated with the reaction of the measuring process on the property being measured. For the purposes of this discussion it is not necessary to summarize the important statistical advances which can be used to improve the accuracy of observed values. The important thing here is to emphasize that error or uncertainty has been recognized as an unavoidable part of scientific observation and that the approach to true quantitative description of the properties of nature is to identify, reduce, or isolate and evaluate the factors extraneous to the properties under observation.

In other words there is no perfection or absolute truth in measured values. Rather we establish a progressive approach to true values by continuously seeking to describe, reduce, or estimate the elements which create inevitable uncertainties in quantitative descriptions. The direction of progress in the physical measurement part of experimental science is to identify the areas of uncertainty associated with a particular measured value and then systematically seek to reduce them or quantitatively to estimate their magnitude or effects.

The first step in an effort to increase the accuracy in the measurement of the value of any property is to define and examine the uncertainties associated with the existing state of the art. This process, if rigorously applied, may reveal that the presumed accuracy of a known quantity is less than had been assumed. It is impossible to conclude with 100% certainty that all of the factors that may influence a measured value have been adequately accounted for. Moreover, it is extremely easy to confuse precision with accuracy. One of the most interesting examples of this situation is our forced acceptance over the past decade of a new value for the velocity of light (based on modern electronic measurement techniques) that differed significantly from prior values. The difference between the old and new values was much greater than would have been expected from the confidence limits assigned to earlier work. Apparently too much confidence had been given to the precision associated with earlier measurements of the velocity of light and insufficient attention given to probable systematic errors.

A basic element in reducing the uncertainties associated with measured values is the accuracy associated with the standards for physical measurement. The major purpose of standards for measurement is to provide a basis to assure the stability and compatibility of measurements from time to time and from place to place. If the accuracies with which measurement standards are known and utilized are less than the precisions with which measurements are carried out in laboratories and on production lines then problems ensue. Data cannot be exchanged with optimum confidence and tests performed on the same materials or devices in different places may be incompatible or uncertain in terms of predicted performance. It is therefore imperative that those responsible for developing, maintaining, and disseminating the measurement standards keep ahead of the major needs of science and industry. This requirement presents a major and never ending problem to the National Bureau of Standards.

Reference was made earlier to the great variety of factors that can influence a measured value. These complex interactions provide the greatest difficulty in extending measurement accuracy. This is natural and inevitable. Extraneous factors that may have a completely negligible influence on a measured value if the desired accuracy is low may have a major or predominate influence as the attempted accuracy increases. We may even reach a situation where the property we are attempting to measure has inherent fluctuations. Consider the cases when thermal noise in a resistor is the limiting factor or when molecular structure and molecular vibration limit the accuracy by which a surface may be defined. In such situations it is imperative that the property we are attempting to measure be accurately and realistically defined.

Space age requirements present to metrologists their greatest problems. The first is the problem of reliability, the second is associated with the variety and quantity of components that must fit and operate together, and the third involves extrapolation of performance based on measurement to extreme and frequently unknown environmental conditions.

Reliability cannot of course be tested into a product but it can only be assured and predicted confidently through measurement. Tolerance limits have to be established and controlled through measurements and as reliability requirements increase tolerances have to be more precisely controlled. One of the critical problems arises because tolerance limits are directly affected by the limits of measurement accuracy. If the realistic tolerance limits approach the limits of measurement accuracy then the tolerances practicably available are sharply reduced.

If, for example, it is desired to control a pulsed voltage so that it is more than 20 and less than 25 volts but we can measure the pulsed voltage only to within plus or minus 2 volts we will have a very difficult problem. Moreover, any uncertainty in measurement will effect some reduction in performance tolerance limits. Hence it is urgent, if we hope to increase reliability, to keep the reductions of measurement uncertainties upon tolerances to an absolute minimum.

An essential characteristic of space age technology is the magnitude and the diversity of the components and sub-systems that must operate in synchronism. In such technology we are confronted with an array of devices that are orders of magnitude in excess of prior situations. Instead of an assembly of thousands of components we may have an amalgamation of upwards of a thousand systems many of which may have thousands of individual components. Furthermore, the individual sub-systems must be fabricated and quality controlled by hundreds of specialty manufacturers throughout the nation. All of these individual equipment fabricators must control the quality of their products through measurement. It is not enough that each sub-system operate reliably in the plant where it is made -- it must be compatible with all the other sub-systems. Here again it is urgent that measurement uncertainties between different laboratories or production lines not significantly reduce the tolerances essential to overall reliable performance.

The environmental problem for space age measurements is no less critical. We are accustomed to measuring performance over limited ranges of temperatures and pressures. But how realistically can we duplicate the temperatures and pressures of outer space? Or the effect of ionizing radiations? Or of vibrations? Or of zero g? Or of other factors still unknown? One thing is reasonably certain. We must learn how to make measurements in new types of environmental chambers, or we must extend our ability to make measurements through telemetry or most likely through a combination of both.

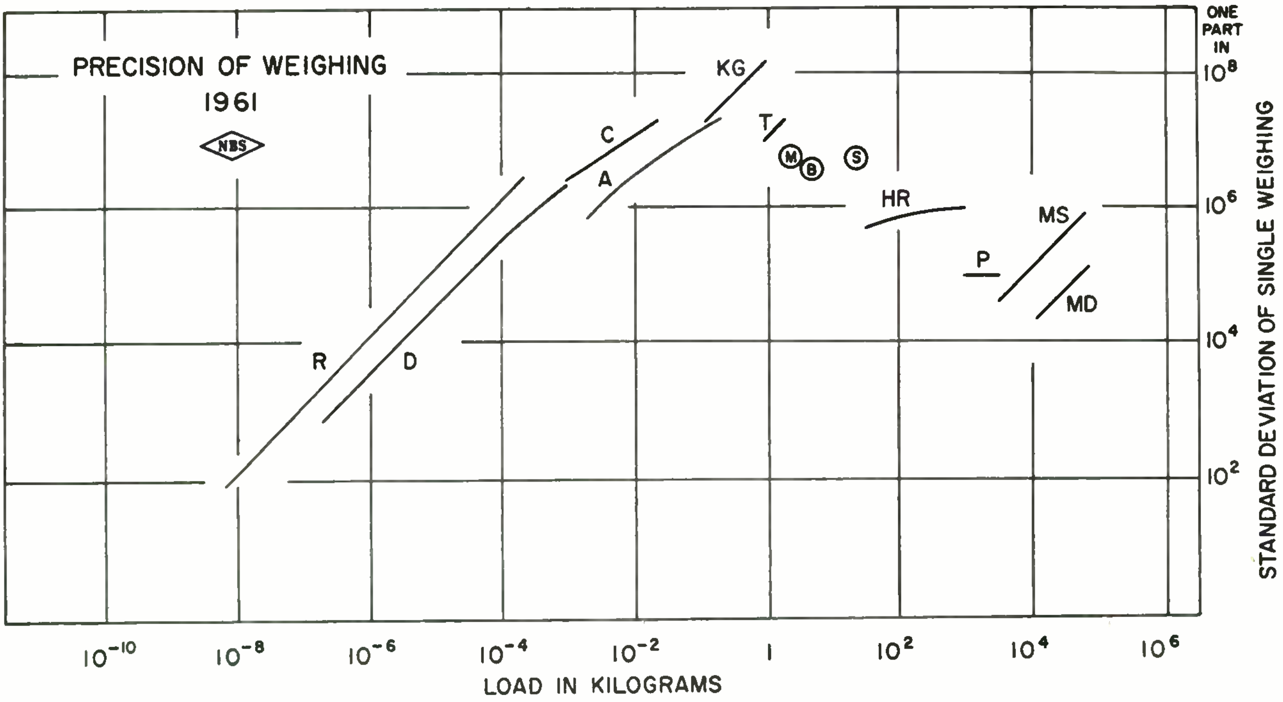
In approaching the problem of space age measurement we must continually evaluate the state of the art. This, as noted earlier, is an important first step in reducing uncertainties in measurement. It is also of basic importance in predicting reliability. Appraising the state of the art also helps to identify the measurement problems most urgently needing attention.

At the National Bureau of Standards a program to evaluate the state of measurement competence has been initiated in specific areas. Figures 1, 2, 3, 4 and 5 show the results of this appraisal for five typical and different properties. The first four figures, dealing with weighing, temperature measurement, pressure measurement, and resistance measurement respectively show one important common characteristic: namely that the highest precision is associated with the reference standard usually at a familiar and convenient value. As we go to higher and lower values the precision as well as the accuracy tend to drop. In Figure 5, dealing with RF voltage measurements, we need to display another parameter namely the frequency of measurement. The state of the art for measuring radio frequency properties varies substantially with the frequency of measurement, as well as with the magnitude of the quantity measured. The simplest way to show precision as a function of both magnitude and frequency in a two-dimensional display is to use the contour map technique as in Figure 5. The standards for RF voltage measurements are derived from d.c. voltage standards which in turn are maintained by standard cells. Hence we find maximum precision in the range of the familiar voltage of the standard cell and at the lowest frequencies.

The great attenuation of precision at the very high and very low values of the property measured represents a continuing and important problem of measurement. The attenuation of precision is amply demonstrated by the figures. If we recall, as we examine the figures, that much of the urgent new development activity is in the region of extreme values then it becomes evident that the measurement precision available for exploration in these regions is extremely limited. Since space age technology is very much involved with work at the extreme values of temperature, force and pressure, as well as with utilization of ultra high frequency and microwave radio waves, the progress of this technology is very much limited by the available measurement techniques.

The figures indirectly point to another basic problem in the extension of the accuracy of measurement. Since all accuracy from the point of view of stable and compatible measurement is derived from the standards these provide a limiting factor for accuracy in all areas of measurement. Thus it is imperative that the accuracy with which the standards can be effectively utilized be adequate to all measurement requirements for a particular property.

The national standards laboratories throughout the world are continuously seeking to improve the basic measurement standards. These efforts led during the past 10 years to important improvements in the status of three of the four basic standards, length, time, and temperature. Further improvement is expected in the time standard by 1966. It is hoped by then that international accord can be reached on an atomic definition for the second to replace the astronomical second. In these efforts also we find that the requirements of space technology provide one of the most compelling reasons for extending the accuracy with which time and frequency measurements can be made.



T Selected Equal-Arm 2kg Balance (NBS T-1)
M Quick-Weighing Single-Arm 6-Pound Balance
B Selected Equal-Arm 10kg Balance (NBS B-1)
S {Special 25kg blance (NBS S-1)
Quick-Weighing Single-Arm 50-Pound Balance
R Quartz - Fiber Ultra-Microbalance
D Special Assay Balance
C Corwin Balance

A Selected Equal-Arm 200g balance (NBS A-1)
KG Rueprecht 1 kg balance
HR Russell Balance, 2500 Pounds
P Platform Scale, 10,000 Pounds
MS Master Scale, 150,000 Pounds, Substitution Weighing
MD Master Scale, Direct Reading

Fig. 1. The state of the art for precision in weighing. Along the abscissa are displayed increasing loads expressed in kilograms, and along the ordinate are shown increasing accuracies expressed as the ratio of the standard deviation of a single weighing to the total load weighed. The letters designate different techniques of weighing.

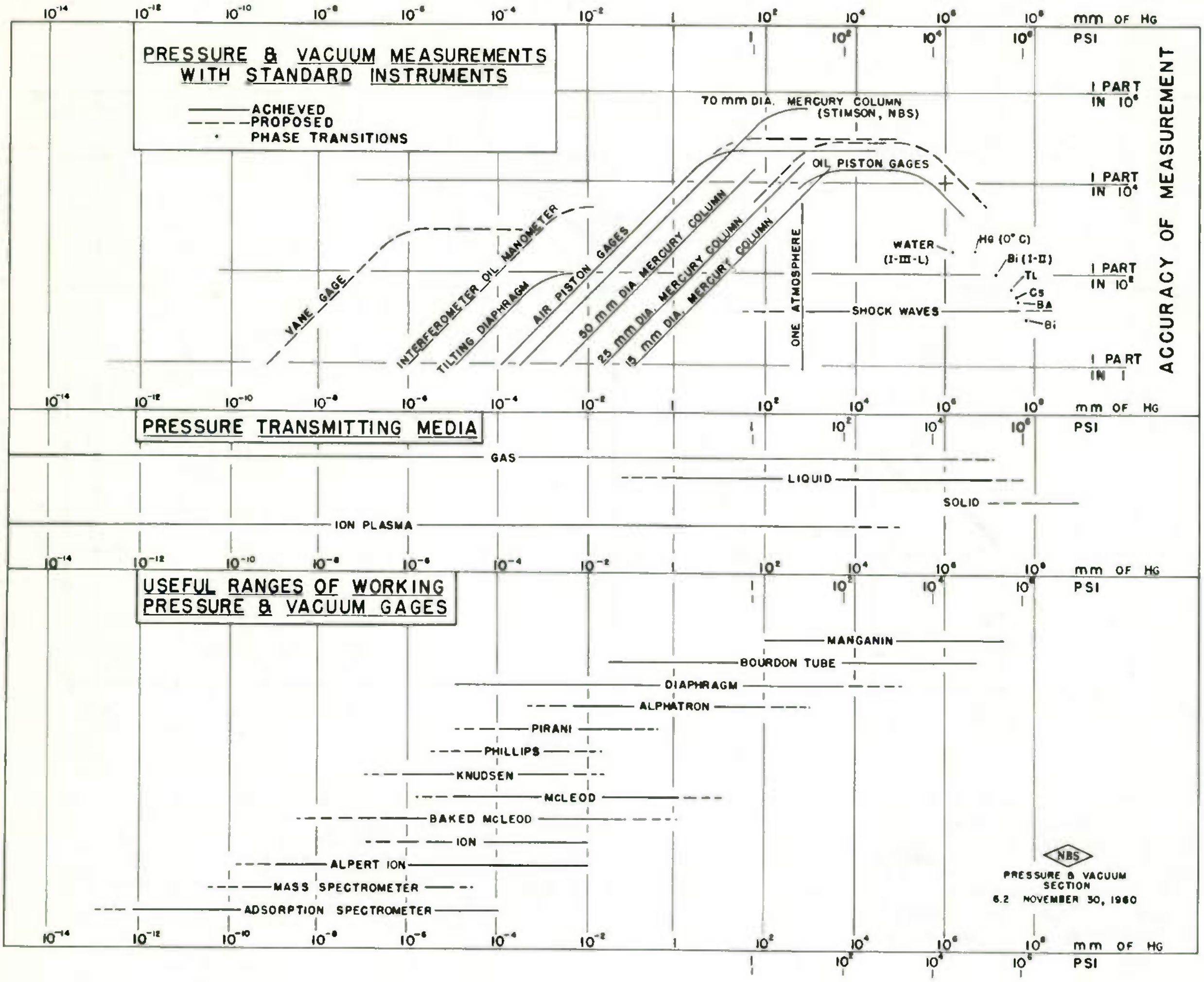


Fig. 2. The state of the art in pressure measurements. Along the abscissa in the upper curve are displayed increasing pressures in terms of mm of mercury or pounds per square inch. Along the ordinate are shown increasing values of accuracy expressed in terms of ratio of the uncertainty to the total pressure measured. The lower half of the figure displays techniques of measurement over different pressure ranges.

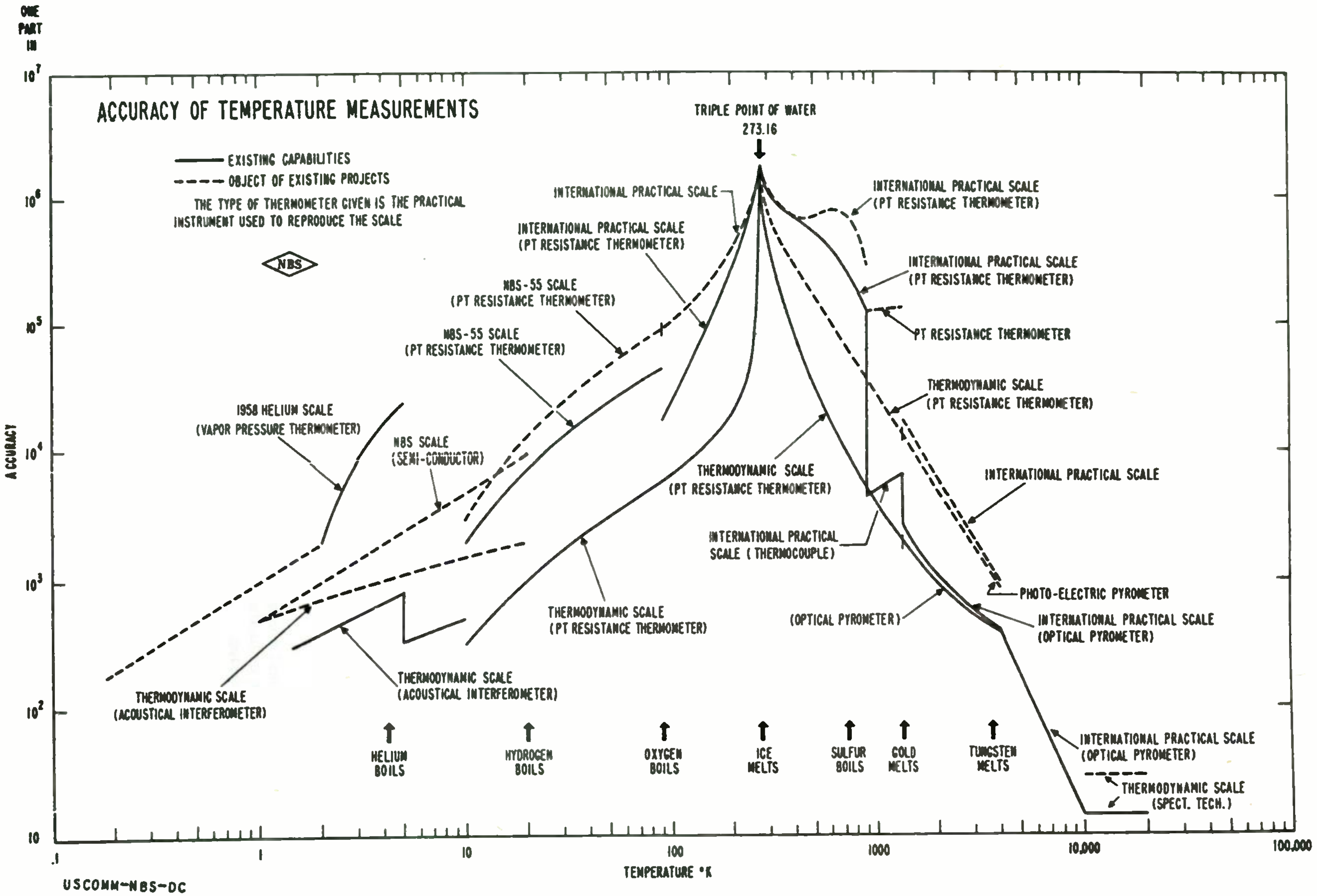


Fig. 3. State of the art in temperature measurements. Along the abscissa are displayed increasing values of temperature, and along the ordinate increasing levels of accuracy expressed in the ratio of the uncertainty to the absolute temperature value. The labels on the figure refer to different techniques or instruments of measurement that are useful over specific ranges of temperature.

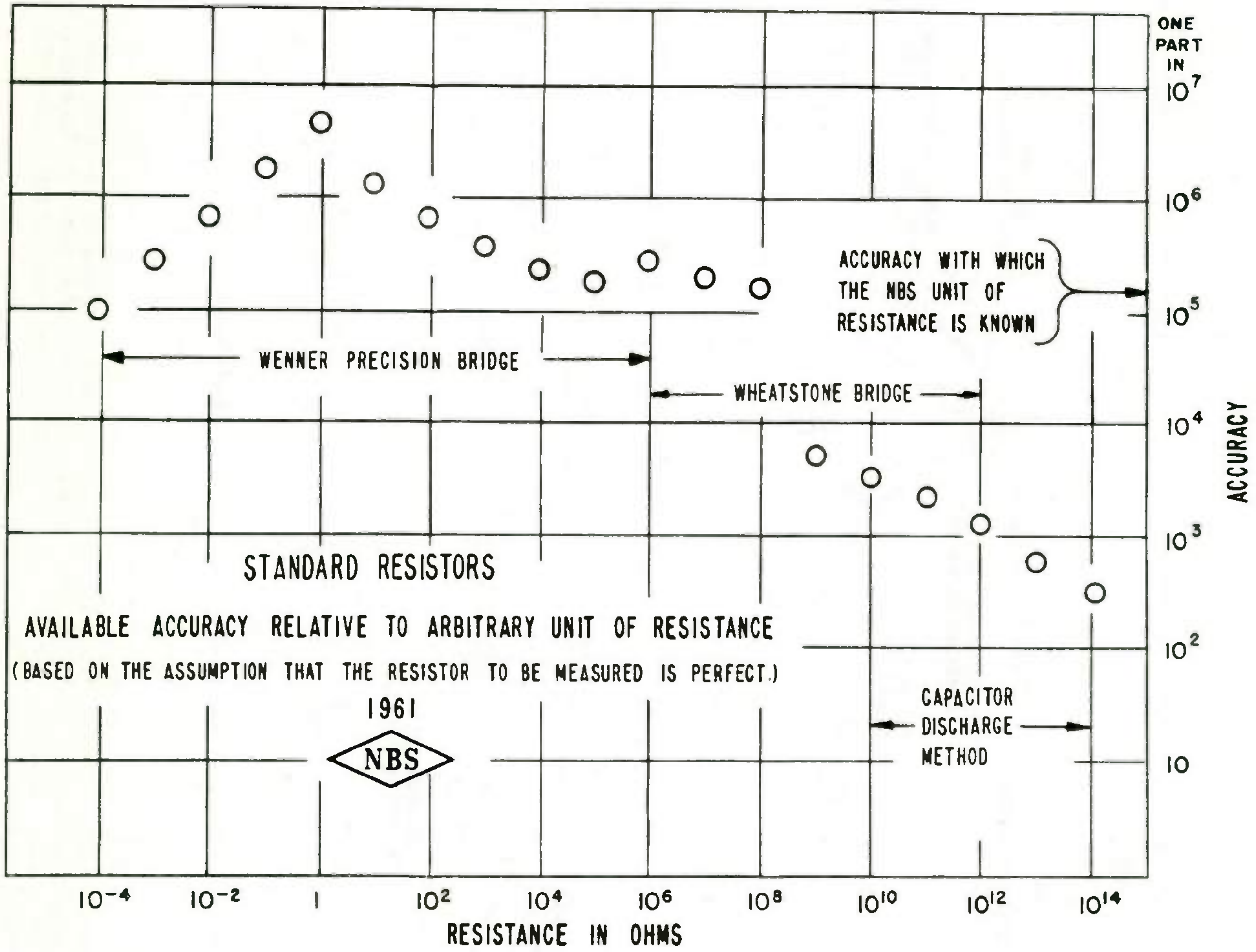


Fig. 4. The state of the art in electrical resistance measurements. Along the abscissa are displayed increasing values of resistance and along the ordinate increasing accuracies expressed in terms of the ratio of the uncertainty to the total resistance values.

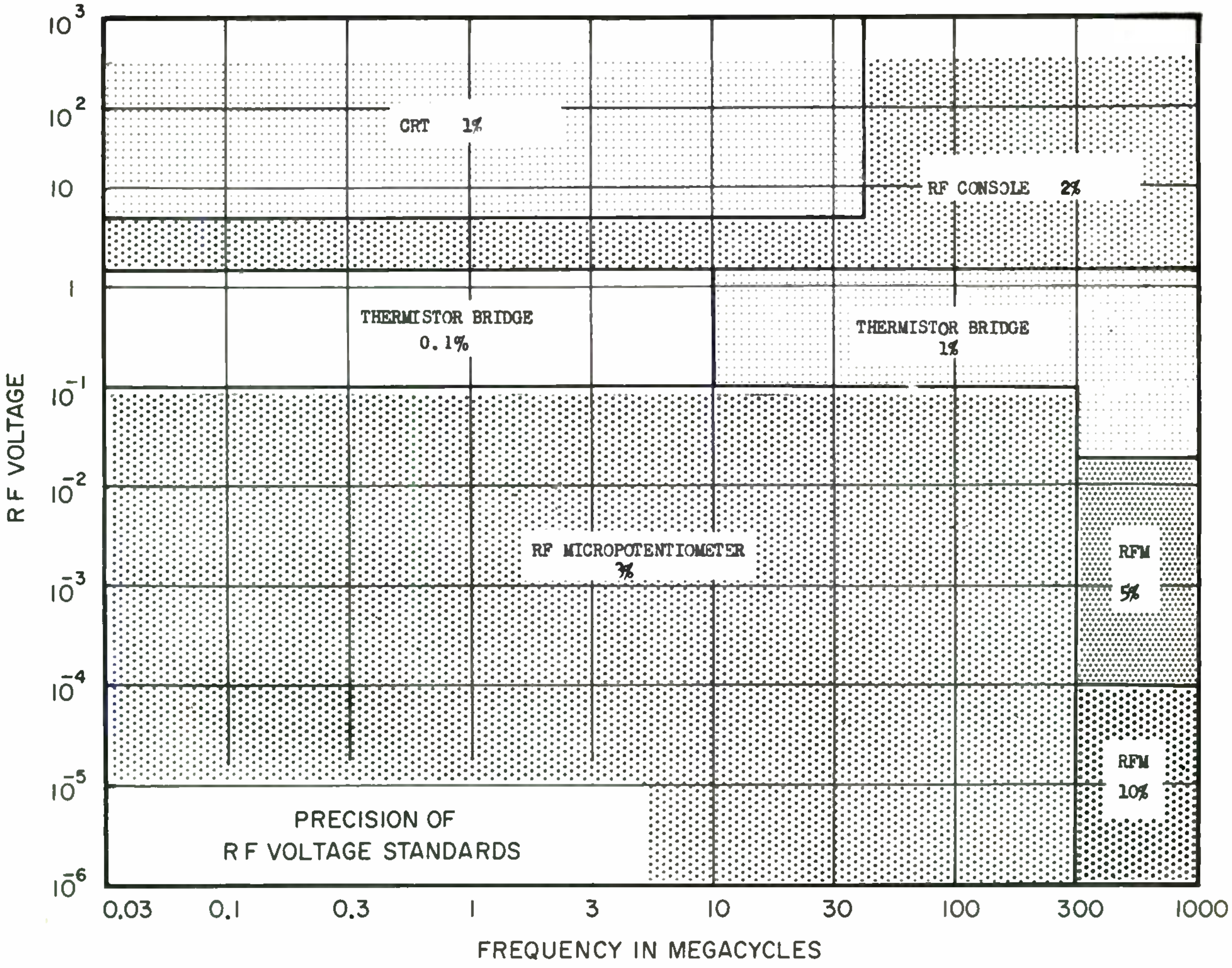


Fig. 5. State of the art for the precision of radio frequency voltage standards. Along the abscissa are represented increasing frequencies expressed in megacycles per second, and along the ordinate values of RF voltage. Areas of frequencies and voltage where a given precision is available are displayed with specific boundaries in the figure.

A FAST, HIGH-CURRENT PHOTOMULTIPLIER

L. F. Wouters, A. E. Villaire, V. E. Wheeler, and R. Kalibjian
Lawrence Radiation Laboratory, University of California
Livermore, California

A fast (subnanosecond) photomultiplier with high output current (>2 amp) is being developed for study of very fast decay processes. Estimated transit time per stage is about 3×10^{-10} sec, with a spread of only about 10^{-11} sec. This tube differs from other fast photomultipliers in having a very short acceleration-deceleration region (giving fast transit time and small transit-time dispersion). Preliminary results from a twelve-stage dynode tube show good focusing from dynode to dynode. Development of the dynode section of the tube is the principal subject of this paper.

Introduction

Ever since the invention of the scintillation detector in 1947-1948, certain limitations in suitable commercial photoelectric tubes have restricted the performance. In photomultipliers, these limitations relate to the characteristics of the photocathodes and the multiplying structures. Significant improvements in time response and in space charge limitations would be useful, and greater flexibility in gain adjustment, gating, and electrode utilization are also desirable.

It was realized many years ago that a dynode configuration using centrally located, high-voltage accelerating electrodes might introduce significant improvements. Developments of this nature have been carried out by various groups. In 1956, Allen and Megill[1] described an experimental dynode structure of this type, and RCA presently markets a tube using centrally located accelerating electrodes.[2]

A somewhat different electron-multiplying geometry is under development at the Lawrence Radiation Laboratory. The aim is to develop a practical tube having the following improved characteristics:

1) Significantly higher linear output current capability (>2 amp).
2) Significantly shorter transit time and less transit-time dispersion.
3) Configuration that lends itself to extracting the signal information from any stage.
4) A geometry in which the focusing and space charge conditions would be fairly independent of individual-stage gain adjustment.

These considerations pointed towards a configuration based on "total" application of the acceleration-deceleration concept, thus permitting the electrons to traverse almost the entire structure at maximum velocity.

The design of a photomultiplier tube involves three parts: (1) the input-photocathode section; (2) the dynode structure; and (3) the output section. While all three parts must have equally good characteristics in order to realize the above objectives, in this report attention will be centered on the dynode structure.

Certain characteristics of "medium vacuum" plasmas and material surfaces, as well as constructional limitations, impose constraints on the choice of dimensions and operating conditions. For example, at about 300 ev impact energy one obtains optimum multiplication. On the basis of experience with breakdown fields in complex tube geometries, approximately 3 kev was chosen as the maximum transit energy.

To avoid ion and x-ray feedback effects, direct vacuum paths between surfaces at extreme voltage differences must be suitably closed. For this reason deflecting surfaces overlap in the lateral direction – but this overlap must not be so great that the beam aperture begins to interfere with trajectories. Fringe field regions also need to be protected to prevent penetration of undesired exterior fields. Such protection can be devised to inhibit feedback along exterior paths.

Analog computations were carried out to evaluate the geometrical concepts, i.e., to check the first-order design calculations summarized below and to make second-order refinements in focusing and path accommodation. The analog method consists of a configured electrolytic tank with field-sensing probe (which traces the electron trajectory) and a computer which imposes the instantaneous integrals of the equations of motion into the probe by servo methods. The electrolytic tank is mounted on a modified Electronics Associates x-y plotter and the computer is an Electronics Associates analog machine. This system, as shown in Fig. 1, greatly reduces the computational labor for particle trajectory calculations. It also permits a greater latitude in design exploration of configurations and operating conditions.

Evolution of the Geometry

A "reflection" geometry, as shown in Fig. 2A, provides for a high traversal velocity as well as for a low impact energy. In this arrangement, the accelerating system has the appearance of a simplified electron gun. The deflection function operates very much like that in an oscilloscope tube. The transverse dimension of the structure will have to be increased in order to provide more latitude to the beam paths.

The elementary configuration of Fig. 2B makes use of the adjacent potential differences to exert a transverse force on the electrons over almost the entire path. The strongest field (hence greatest effect on trajectory) is applied after the electron leaves the dynode surface. Trajectory runs on the computer revealed some problems with this early geometry. For example, it was anticipated that some degree of focusing by curving the dynodes would be needed; however, it proved difficult to geometrically accommodate trajectories having adequate focusing. These trajectories are controlled almost entirely by electrode curvature in combination with the transverse field.

The compromise geometry shown in Fig. 3 was then developed. It uses a combination of shaped-electrode focusing, transverse electrostatic deflection, and electrode tilting. We call this geometry the guided-beam, constant-gradient multiplier. (Note that the grids are actually curved wires lying in the plane of the illustration.) In this design, the grids accelerate secondary electrons from the dynodes to a high velocity in a very short distance. These high-energy electrons then drift rapidly towards the succeeding dynode. Upon passing through the grid of the succeeding dynode, they are rapidly decelerated to the low velocity required to give a maximum secondary emission ratio. This tube differs from other fast photomultipliers in that the acceleration-deceleration region is very short; consequently, transit-time dispersion is expected to be held to a very small value, and the allowable space-charge density is also expected to be quite large.

In the decelerating region the primary electrons are injected at an angle θ with respect to the electric field lines. These electrons could suffer a total reflection, in which case they will not reach the dynode surface. In order for these electrons to strike the dynode, the energy which corresponds to the vector velocity in the direction of the electric field lines must be greater than the retarding field energy. The limiting relationship which permits the primary electron to graze the dynode is

$$V_1 = V_0 \sin^2 \theta$$

where eV_0 is the initial energy of the primary electron, eV_1 is the impact energy at the dynode, and θ is the angle of incidence as shown in Fig. 4.

Initially, trajectory runs were made for the geometry of Fig. 3 for an initial energy eV_0 of 3.3 kev and an impact energy eV_1 of 300 ev; however, for this arrangement the primary electrons suffered total reflection. Decreasing the initial energy to 3.0 kev and increasing the impact energy to 600 ev enabled the primary electrons to impinge upon the dynode as shown in Fig. 5 for a typical set of analog trajectory runs. Focusing is achieved almost entirely by electrode curvature. The focus diagram of Fig. 6 shows the electrons converging at the center of the dynode. The electrons from a point source on the dynode essentially fall on a line focus on the succeeding dynode. A possible solution to electron end loss may be a toroidal dynode structure, in which a cross-section of the axial plane would have essentially the geometry of Fig. 3.

Satisfactory deflection control is obtained by tilting the electrodes and limiting the transverse field to that naturally imposed by successive stage potentials. The effect of electrode tilt can be expressed in terms of the best choice of ratio of dynode spacing to dynode gap, b/a. This is shown as a function of lateral motion contributions in Table I. The function is not severe, and the final choice of 9° was made primarily to permit maximum accommodation of dispersive trajectories.

Table I. Electrode tilt effects.

Tilt/field ratio	Optimum b/a
0	$\sqrt{V_y/V_x}$
1/2	$\sqrt{3/2}\,\sqrt{V_y/V_x}$
1	$\sqrt{2}\,\sqrt{V_y/V_x}$
2	$\sqrt{3}\,\sqrt{V_y/V_x}$

The estimated transit time for the geometry is about 3×10^{-10} sec per stage with an acceleration-deceleration increment of about 6×10^{-11} sec. Various contributions to the time dispersion indicate an accountable transit time spread of 10^{-11} sec per stage (full-width) as summarized in Table II.

Table II. Estimated transit time dispersion contributions (per stage).

Initial electron energy dispersion - - - - - - - - - - -	2×10^{-12} sec
Field penetration (through grids) - - - - - - - -	2×10^{-12} sec
Extreme geometrical trajectory differences - - - -	10^{-11} sec
Overall - - - - -	$\sim 10^{-11}$ sec

The primary electron impact energy of 600 ev at the dynode unfortunately imposes an interdynode potential of 600 volts which could contribute to a very large total voltage for a cascade of 10 or 12 stages. At present, we are investigating the possibilities of decreasing both the impact energy and the initial energy (as governed by the grazing relationship); consequently, this would mean an increase in transit time dispersion. For the case where the impact energy is decreased from 600 to 300 ev (with corresponding decrease in initial energy as governed by the grazing relationship), the transit time dispersion in the acceleration-deceleration regions can be minimized by decreasing the gap between dynode and grid; also, the transit time between grids will increase by no more than a factor of 2.

The ultimate output current would be generated by a defocused incident beam, "spraying" onto its target dynode. The resultant dynode emission would be space-charge-limited to 30 amp per cm^2. Important deviation from linearity would certainly start at a lower value.

Laboratory Model Studies

Scaled-Up (5×) Electron Model

In order to demonstrate the electron optics of the dynode structure, as determined in the electrolytic tank, a 5× scaled electron model was constructed and tested in a demountable vacuum system as shown in Fig. 7. A scaled-up model permits rapid checkout of a concept with relaxed tolerances in structural dimensions.

The model's 90% transparent grid consisted of 0.005-inch-diameter molybdenum wire spot-welded to a nickel rod frame. No attempt was made at a perfectly formed grid; actually the grid had more of a triangular cross section than a cylindrical one. An electron gun was used to supply the primary electrons to the first dynode. The dynodes were dusted with a phosphor powder in order to observe the luminescent patterns of the impinging electrons. Proper focusing between dynodes was observed for a dynode tilt of about 9°. The electron spread was limited to the center portion of the dynode.

Full-Scale, Three-Stage Model

Having verified the electron focusing between dynodes in the 5× scaled model, we next built a full-scale, three-stage model. This model was primarily intended as a prototype for future cascaded systems. We wanted to study (1) the fabrication and assembly of the electrodes, (2) the operation of the multiplier with very high electric fields (75,000 volts per inch), and (3) the overall performance.

The dynode is 0.316 × $1\frac{1}{2}$ inches and it has a concave cylindrical surface of 3/4-inch radius. A grid conforms to this surface with a gap of 0.040 inch as shown in Fig. 8.. The grid consists of parallel 0.002-inch-diameter nickel wires spaced 0.020 inch apart. These grids are formed by brazing the wires onto a 0.020-inch-thick photo-etched nickel frame. The brazing filler is allowed to flow in a wet hydrogen furnace to prevent the grid from sticking to the oxidized stainless steel mandrel. Possibility of burnout of the nickel wire at very high currents is slight. Since the intended use of the multiplier is for a duty cycle of about 10^{-4}%, the average power dissipated in the grid is very low (less than 1 milliwatt for 1 ampere of current).

All of the insulating mounts for the multiplier are positioned on the back side of the structure. This leaves the sides open so as to inhibit charging effects. As shown in Fig. 8, the structure is held together with multiform glass rods and Kovar beaded glass spacers. The three-stage tube is shown in Fig. 9. An electron gun provides the source of primary electrons to the first dynode. In order to examine focusing effects visually, the electrodes of the multiplying structures have been dusted with a phosphor powder.

Breakdown of the electric field (3 kv across 0.040 inch) is not observed in the dynode-grid assembly during operation. It is not possible to determine dark current due to the high field in a three-stage tube. The patterns of the collected secondary electrons on the arrival dynode, corresponding to a point source of secondary electrons leaving the departure dynode at various positions, are shown in Fig. 10. The image from a point source has an appreciable spread in the longitudinal direction due to the diverging lens effect in both the acceleration and deceleration regions. The parallel wires of the grid minimize the divergence in the lateral direction; the electron spread in this direction could be attributed to scattered primaries having higher component velocities of about 100 ev.

The gain of the multiplier is low since the dynodes are made of stainless steel. Current interception by the grids and shield plates appears to be important. The pair of 90% transparent grids intercept about 20% of the electrons. In addition, there is 15% interception by the shield plates; hence, the total transmission per stage is about 70%.

Twelve-Stage Model

We next built a 12-stage dynode tube. The construction was simplified somewhat by making two separate halves as shown in Fig. 11. The shield plates and dynodes are carefully jigged and held together with multiform glass rods. The dynode surface is thus directly accessible for the placement of the fragile grids. The 2 half-sections are assembled together and mounted on a 23-lead glass stem as shown in Fig. 12. The assembled tube with an electron gun input section is shown in Fig. 13. The tube is presently being evaluated for its electrical characteristics.

Input and Output Sections

As previously mentioned, the input and output sections of the photomultiplier must have as good characteristics as the dynode structure in order to realize the small transit-time dispersion. To preserve the wide bandwidth capabilities of the dynode structure, the processing of information will be performed inside the tube. The input photocathode section will be a spherical surface with an accelerating grid (parallel spaced wires) spaced close to it. In principle, the operation will be very similar to that of the cylindrical dynode-grid structure.

The output section of the tube differs from the standard photomultiplier in that the output current from the last dynode is formed into a beam and focused upon a phosphor screen. Two sets of deflection plates are included in order to provide for a circular sweep as shown in the schematic of the tube in Fig. 14. For example, in the study of very fast decay processes in scintillators, the time constant is determined by increasing the period of the circular sweep until a luminescent trace of less than 360° is recorded.

At present, characteristics of various types of lenses are being studied for the beam-forming electrodes. The rectangular cross section of the beam from the last dynode is to be formed into a small circular cross section. In order for the efficiency of the lens to remain high, the defining aperture can intercept only a small fraction of the beam current. The aberration in the lens must be small to keep the transit time dispersion also small in this section.

Acknowledgments

This work was done under the auspices of the U. S. Atomic Energy Commission. The authors wish to thank Mr. L. Mancebo for technical discussions, and Mr. J. P. Henz, Mr.

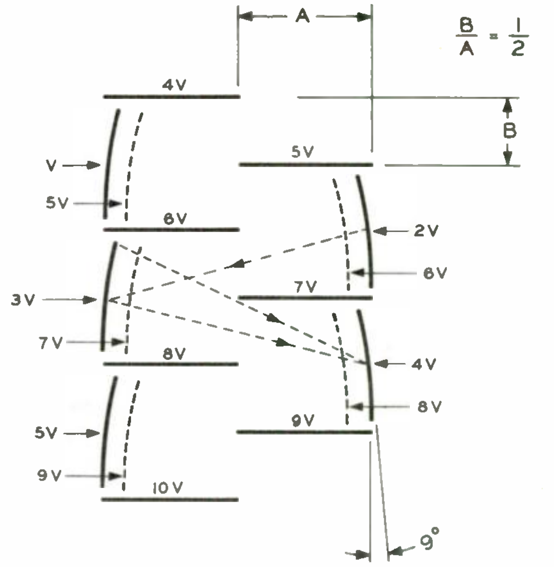
C. H. Gillespie, and Mr. W. C. Grayson for the fabrication and assembly of the various models of the multiplier tube.

References

1. J. S. Allen and L. R. Megill, IRE Trans. on Nuclear Sci. NS-3, 112 (Nov. 1956).

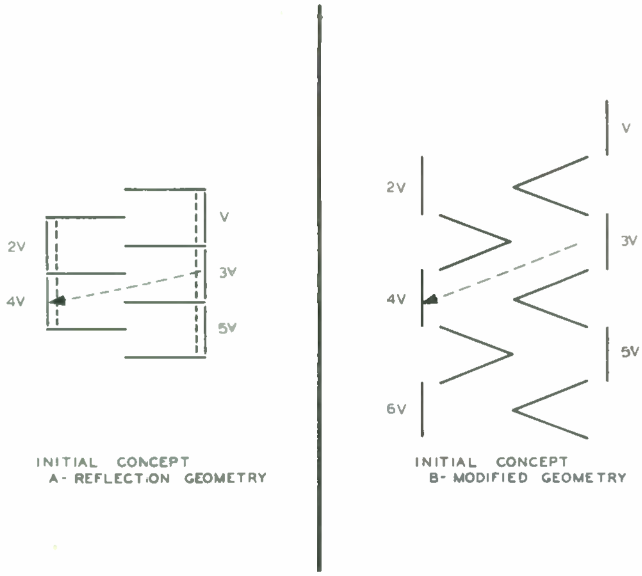
2. G. A. Morton, R. M. Matheson, and M. H. Greenblatt, IRE Trans. on Nuclear Sci. NS-5, 98 (Dec. 1958).

Fig. 1.



GUIDED BEAM CONSTANT GRADIENT MULTIPLIER

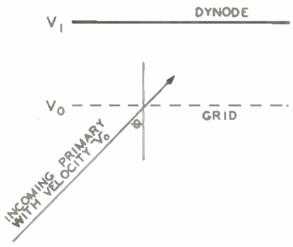
Fig. 3.



INITIAL CONCEPT
A- REFLECTION GEOMETRY

INITIAL CONCEPT
B- MODIFIED GEOMETRY

Fig. 2.



LIMITING CONDITION FOR AN ELECTRON TO GRAZE THE DYNODE IN A RETARDING FIELD - $V_1 < V_0$

$V_1 = V_0 \sin^2 \theta$

Fig. 4.

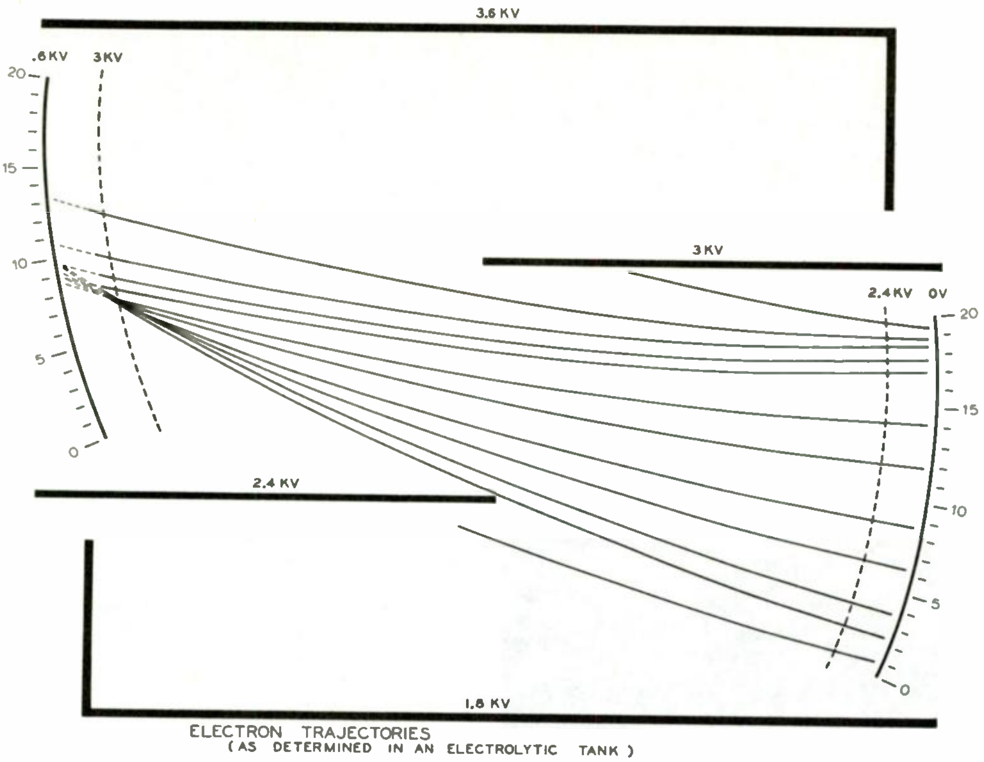


Fig. 5.

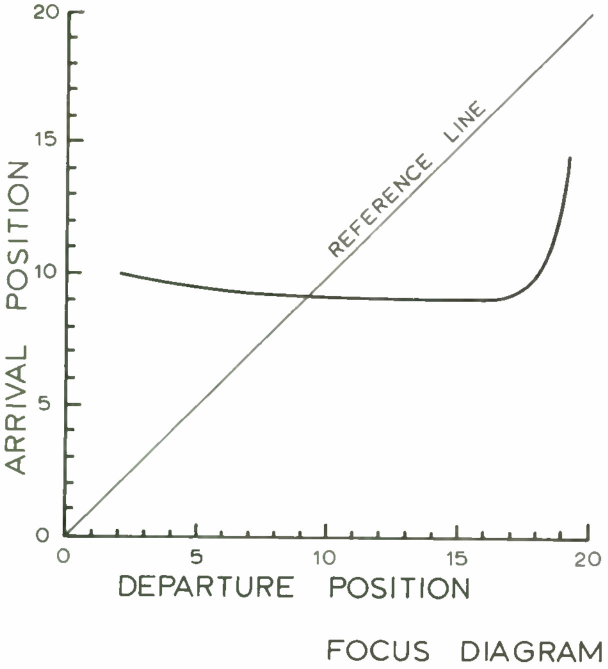


Fig. 6.

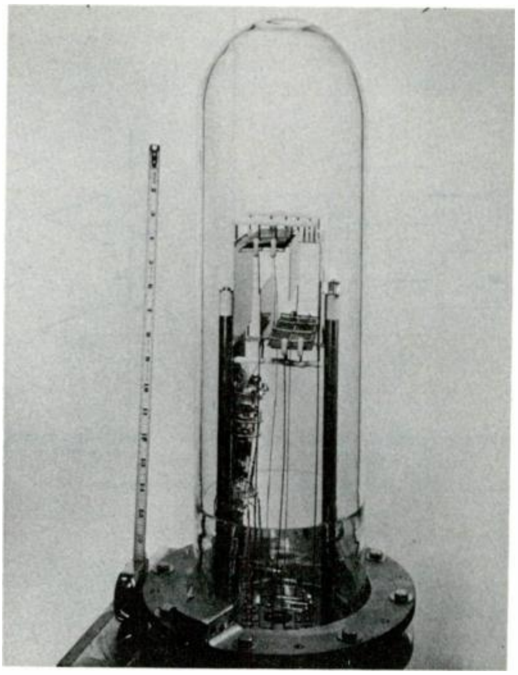


Fig. 7.

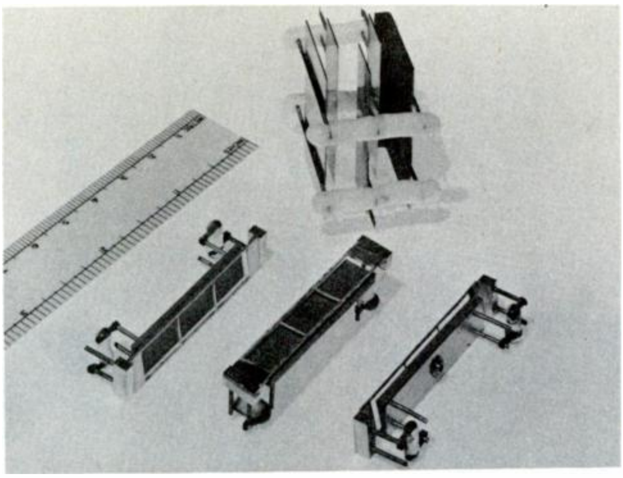


Fig. 8.

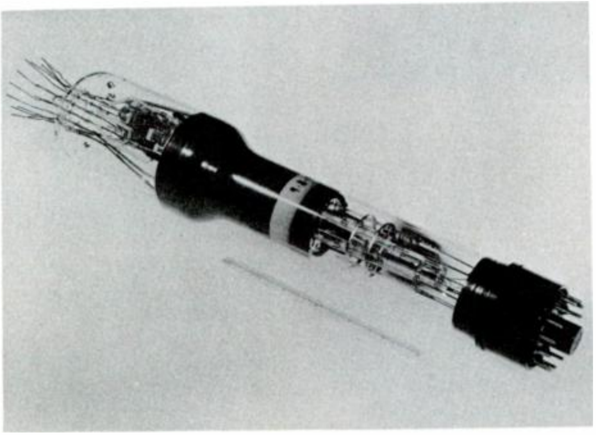


Fig. 9.

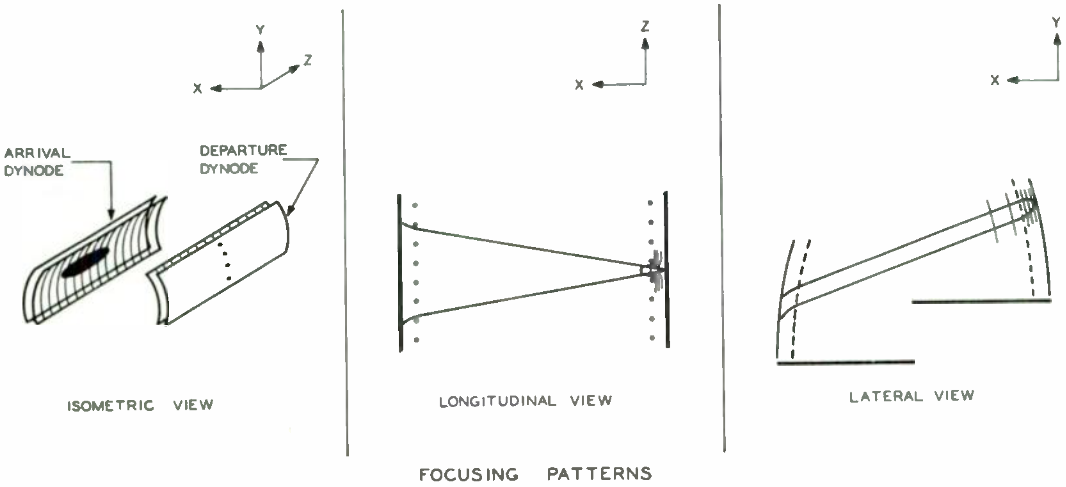


Fig. 10.

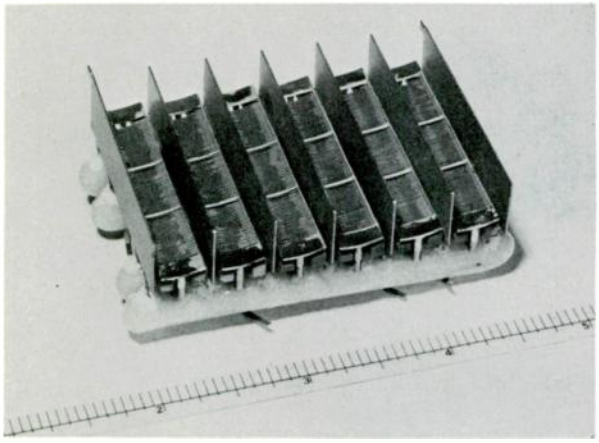


Fig. 11.

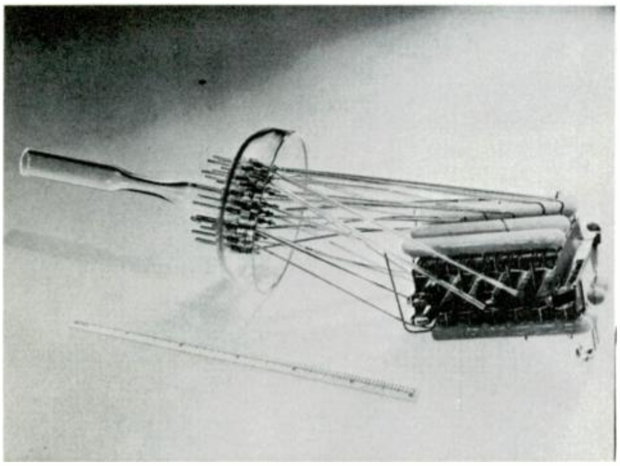


Fig. 12.

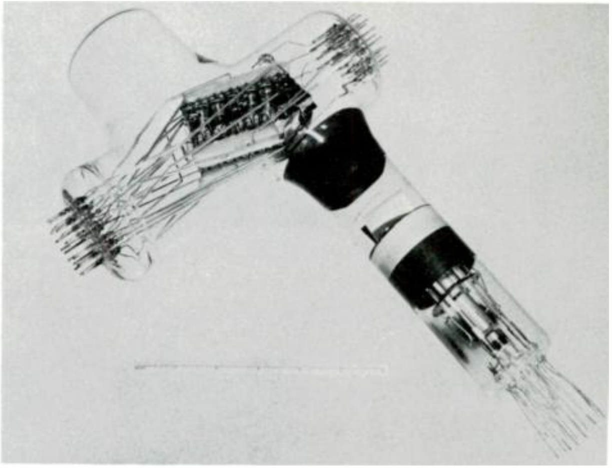


Fig. 13

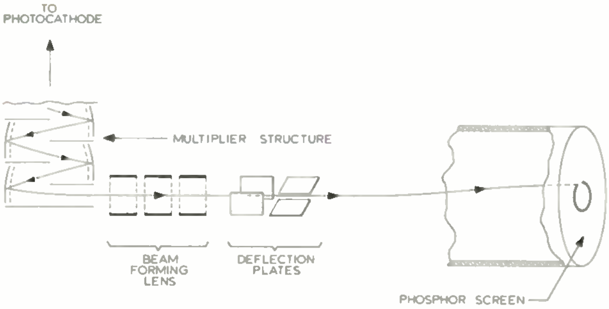


Fig. 14

NUCLEAR PULSE CURRENT-SENSITIVE AMPLIFIERS

Irving M. Meth
and
Robert T. Graveson
Health and Safety Laboratory
U. S. Atomic Energy Commission
New York, New York

Abstract

The development of current sensitive feedback amplifiers and their application to nuclear instrumentation is described and is exemplified by a non-overloading amplifier stage. Characteristics of this stage include a stage gain of 40, a loop gain of 250, a rise time of 0.2 μsec, and an input resistance of six ohms. Specific applications include a linear current discriminator, current to voltage converters, charge integrators and delay line shapers. A review of feedback theory and design equations is included.

Introduction

Current sensitive amplifiers have been used as nuclear pulse-amplifiers.[1,2] The pertinent characteristics of amplifiers used with nuclear detectors are charge integration and the random time and pulse height distribution of the signal. Idealized characteristics of current sensitive feedback amplifiers are:

1. zero input resistance,
2. infinite output resistance, and
3. stable current gain from input to output.

Current sensitive amplifiers bear a dual relationship [3,4] to the more conventional or voltage sensitive amplifiers.

These principles are illustrated by the development and a number of specific applications of a non-overloading amplifier stage. Characteristics of this amplifier include a stage gain of 40, a loop gain of 250, a rise time of 0.2 μsec, and an input resistance of six ohms.

Theory of Current Amplifiers

The equivalent circuit of a current sensitive amplifier is given in Figure 1, where:

A is the short-circuit current gain of the base amplifier,
R_i is the input resistance of the base amplifier,
R_o is the output resistance of the base amplifier, and
γ is the proportion of output current fed back to the input.

The ratio of output to input current A_{iF} is given by

$$A_{iF} = \frac{I_L}{I_E} = \frac{AR_o}{(R_o + R_L) + A\gamma R_o} \quad , \qquad (1)$$

which may be written as

$$A_{iF} = \frac{A}{1 + A\gamma} \quad \frac{R_o\,(1 + A\,\gamma)}{R_o\,(1 + A\,\gamma) + R_L} \quad . \; (2)$$

Equation (2) may be represented by a Norton's equivalent circuit with short-circuit gain given as

$$A_{iF}\Bigg|_{R_L=0} = \frac{A}{1 + A\,\gamma} \quad , \qquad (3)$$

shunted by an output resistance of value

$$R_{oF} = R_o \quad (1 + A\,\gamma) \qquad . \qquad (4)$$

Thus, for large values of loop gain, $A\gamma$, the current gain is approximately $\frac{1}{\gamma}$, independent of amplifier gain and load resistance.

The input voltage, V_{in}, required to support

I_E, is given by

$$\frac{V_{in}}{I_E} = \frac{R_i}{1 + A\gamma} = R_{iF} \quad . \qquad (5)$$

Thus, the feedback loop degenerates the input resistance and the reduction of input resistance is a direct measure of the loop gain.

Though the equivalent circuit indicates transformer coupling, it is more convenient to utilize current division at a node to derive the feedback current. This is shown in the equivalent circuit of Figure 2, which is derived from the circuit of Figure 1 by replacing the base amplifier with a common emitter stage, T_1. The division of emitter (or load) current, of the feedback node F, causes a fraction (γ) or I_L to be fed back to the input. By neglecting the small input resistance R_{iF} as compared to R_2, a good approximation is obtained for γ:

$$\gamma = \frac{I_F}{I_L} = \frac{R_1}{R_1 + R_2} \quad . \qquad (6)$$

The forward gain, A, to be utilized in relationships (1) to (5) includes the gains of T_1 and T_2. Thus,

$$A = \beta_1 (1 + \beta_2) \quad ,$$

and

$$I_L' = \alpha_2 \, I_L \quad ,$$

where α_2 and β_2 are the common base and common emitter gains of T_2, and β_1 is the common emitter gain of T_1. For large values of loop gain,

$$\frac{I_L'}{I_E} = \alpha_2 \frac{R_1 + R_2}{R_1} \qquad (7)$$

independent of the individual stage gains.

This amplifier cannot be used for a stable voltage amplifier since the voltage gain is not degenerated by the feedback loop. This may be seen by examining the expression for voltage gain, A_v:

$$A_v = A_{iF} \frac{R_L}{R_{iF}} \quad . \qquad (8)$$

The expressions for A_{iF} and R_{iF} both contain the loop gain in their denominators. Therefore,

$$A_v = A \frac{R_L}{R_i} \qquad (9)$$

and is not a stabilized quantity. Effectively, a voltage source at the input terminals shorts the feedback current to ground, producing open-loop operation.

Prototype Circuit

The principles described above have been embodied in the amplifier of Figure 3, which is designed to accept negative pulses. Adaptations of a cascode stage with a boot strap load, complementary symmetry emitter follower and a high gain, direct-coupled feedback loop have been incorporated into this amplifier.

A large forward current gain is achieved by utilizing a cascoded stage (X_1 and X_2), with a boot strap collector resistor, that is loaded by an emitter follower, X_3. Since the common base transistor (X_2), in the cascode stage, effectively short circuits the common emitter transistor (X_1), and since the output impedance of a cascode approximates the high impedance of a common base stage, then the short circuit signal current of X_1 is transferred to the emitter follower, X_3. Boot strapping prevents the signal current from being diverted from the high input impedance emitter follower to the collector resistor. The gain of this arrangement is the product of the individual betas of X_1 and X_3. X_2 is utilized as an impedance matching stage.

The feedback voltage used for boot strapping the collector resistor is coupled from the emitter follower, X_3, through a Zener diode, Z-1. The Zener voltage and collector resistance establish the quiescent value of the emitter current in the cascode stage. Since boot-strapping techniques are most effective if the emitter follower gain approaches unity, a second emitter follower, X_5, isolates the boot strapping and feedback nodes. Thus, low values of R_1, required for large current gains, do not reduce the open loop gain.

A complementary symmetry emitter follower (X_4 and X_5) establishes the feedback node, F. In the quiescent stage, X_4 is cut off and X_5 is biased for a known level of emitter current. With the application of a signal, X_5 is driven out of conduction. At cut-off, X_4 is driven into conduction maintaining a uniform loop gain. The maximum output capability of the amplifier is established by the quiescent value of the emitter current in X_5. Even under overload conditions, the response is that of a closed loop system.

Quiescent conditions and signal gain are established through a high gain, direct-coupled feedback loop. The absence of energy storage elements prevents "set-up" during overload with a resulting slow recovery period. The Zener diode is driven into increased conduction with signal and thus does not limit amplifier linearity. In fact, each stage is driven into increased conduction with signal. The limit of closed loop operation is bottoming of X_2. If the storage time of X_2 is objectionable, a pair of germanium and silicon diodes may be used to clamp the collector junction of X_2 to a small negative voltage.

Overload capability is limited by the maximum current handling capability and off-time of the NPN emitter follower, X_4. Under overload, X_4 is required to supply large peak currents without driving into excessive beta saturation. Time delay of the lagging edge under severe overload can be caused by two effects:

1. Since X_4 must be driven to cut-off before X_5 responds to the signal, the lagging edge of the output pulse appears delayed with respect to the corresponding edge of the input signal. This delay is dependent on the alpha cut-off frequency of X_4.

2. Since the return difference, I_{in}, drives X1 towards decreased conduction, X_1 may, for signals with a fast trailing edge, be driven beyond cut-off. As the loop opens, X_5 is forced towards conduction and X_4 towards cut-off, producing a rapid recovery from overload.

Applications

A number of applications is immediately suggested for this circuit, the most obvious being a stable current amplifier. Other applications include a linear current discriminator, a current-to-voltage converter, a charge integrator, and a delay line shaper.

Stable Current Amplifier

In the current amplifier of Figure 4, the gain is determined by the value of resistors R_1 and R_2. A gain of 40 per stage is found convenient. R_2 is set at 6800 ohms and R_1 at 180 ohms. Typical performance figures include an input resistance of six ohms, a rise and fall time of 0.2 μsec. The oscillograms of Figure 5 show, to a common time scale, the input and output waveforms for a stage gain of 40. The gain-bandwidth product of the amplifier approximates the figure of merit of the (2N1301) transistors within the amplifier.

The input resistance indicates a closed loop gain of approximately 250 and a forward gain of 10,000. Gain stability is dependent, primarily, on the stability of resistors R_1 and R_2. This amplifier can handle 100 times maximum linear input with negligible distortion and with a recovery time of 0.7 μsec for rectangular pulses; exponentially decaying pulses require a negligibly short time to recover from heavy overload. The oscillograms of Figure 6 show overload performance. The input and output signals for normal, X3, X10, X30, and X100 of maximum linear operation are given to a common time scale. The low value of input resistance permits these stages to be cascaded without use of impedance matching networks. However, a step-down transformer may be used as an interstage coupling network to achieve additional current gain between stages.

Linear Current Discriminator

The gain of the amplifier may be adjusted, maintaining fixed quiescent conditions, as is shown in Figure 7. The output of the amplifier is coupled to a tunnel diode discriminator that is followed by a PNP transistor. The tunnel diode is biased for zero current. If the signal exceeds the peak current of the tunnel diode, it will trigger to its high-voltage state driving the output transistor into saturation. Since the peak current is fixed, the value of input current required to exceed the tunnel diode threshold is determined by the gain setting of the amplifier. To maintain a variable discriminator level, which is linearly related to the input current, requires a stage gain that is hyperbolically related to its control. If R_1 is small compared to R_2, the gain varies inversely with R_1 as required. Thus, the triggering threshold varies linearly with R_1.

A typical discriminator characteristic is shown in Figure 8. Larger values of R_1 than those shown cause the threshold to vary parabolically; smaller values of R_1 cause the gain to become unstable. The maximum tunnel diode drive is the quiescent current of X_5, which can be arbitrarily adjusted by varying the value of resistance that returns R_1 to ground.

Current-to-Voltage Converter

The conversion resistance of this stage is determined by the values chosen for R_L and current gain. Two factors limit the maximum value of R_L. One factor is the capacitance across R_L that decreases the rise time of the output signal. This effect may be reduced by coupling from X_5 through an emitter follower, degenerating the load capacitance. The second limitation on R_L is the maximum "stable" gain of the amplifier. Since the input resistance is low, the voltage gain of the stage for large values of R_L will be high - 10,000 to 40,000 being typical figures. Parasitic oscillations can result with these large gains. Therefore, great care must be exercised in shielding the input and output leads.

Charge Integration

By shunting R_2 with a capacitance, the output can be made proportional to the integral of the input current. Thus, the circuit can be used as a charge-to-voltage or charge-to-current converter.

Delay Line Shaping

The amplifier can be used as a delay-line driver for shaping pulses with a slow trailing edge. The delay-line, with a length equal to one half the desired pulse width, is short circuited and the load resistance for the output transistor is set equal to the characteristic impedance of the delay-line. The output is derived from the input of the shorted delay-line. The oscillograms of Figure 9 give the input and output waveforms for a slowly decaying signal that is shaped into a 1 μsec rectangular pulse.

References

[1] Goulding, F. S. - Transistorization, Nucleonics, 17, 6, 64-71, June 1959.

[2] Graveson, R. T. and Sadowski, H. - Pulse Amplifiers Using Transistor Circuits, IRE Transactions on Nuclear Science, Vol. NS-5, No. 3, 179-182, December 1958.

[3] Wallace, R. L. and Raisbeck, G. - Duality as a Guide in Transistor Circuit Design, BSTJ, Vol. XXX, 381-418, April 1951.

[4] Waldhauer, F. D. - Wide-Band Feedback Amplifiers, IRE Transactions on Circuit Theory, Vol. CT-4, No. 3, 178-190, September 1957.

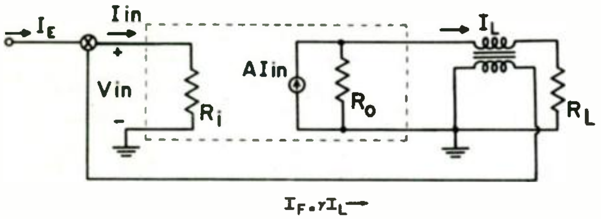


Fig. 1. Equivalent circuit of current sensitive amplifier.

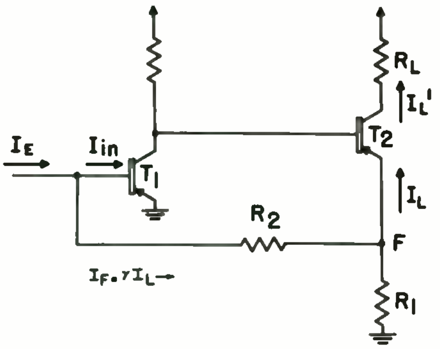


Fig. 2. Equivalent circuit of current sensitive amplifier using resistive current division.

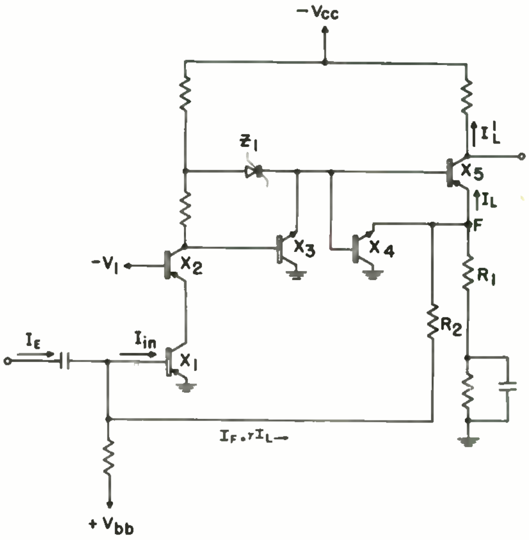


Fig. 3. Prototype of current sensitive amplifier.

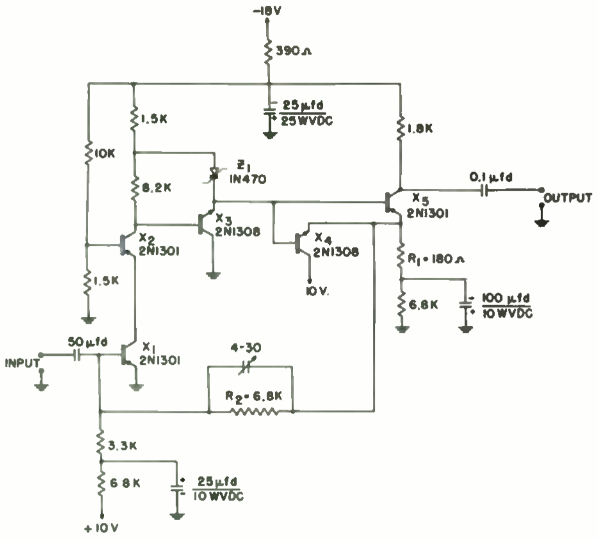


Fig. 4. Schematic of X40 current amplifier.

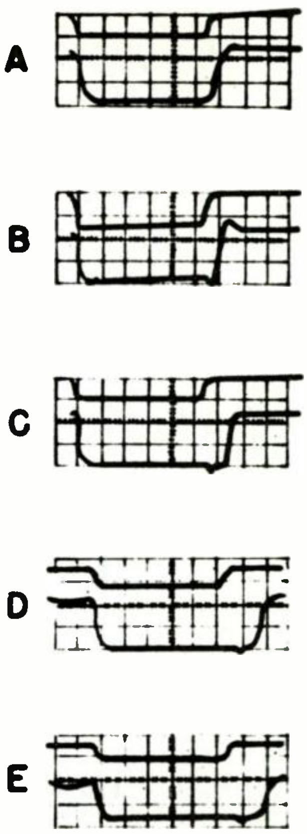


Fig. 6. Overload performance. Input and output waveforms for normal, X 3, X 10, X 30 and X 100 maximum linear operation. Time scale 0.5μ sec/cm.

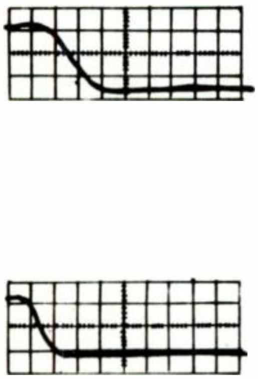


Fig. 5. Input and output waveforms of X40 current amplifier. Time scale 0.1μ sec/cm.

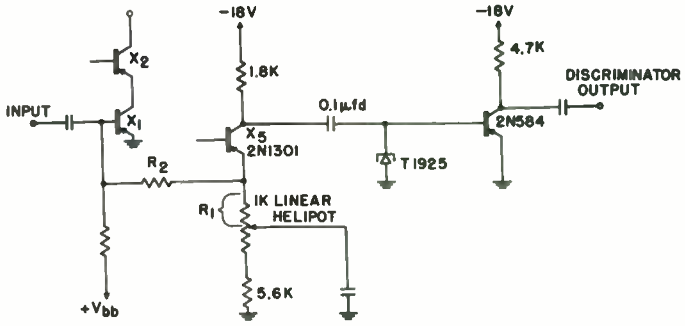


Fig. 7. Linear current discriminator.

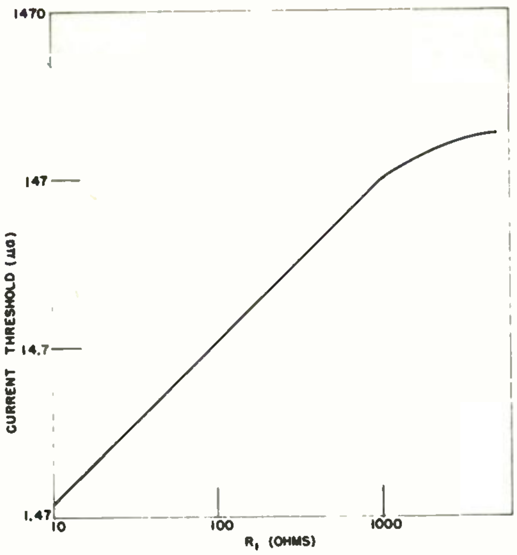


Fig. 8. Current-discriminator characteristics.

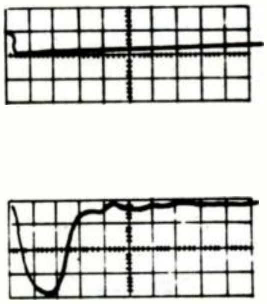


Fig. 9. Delay line shaping. Input and output waveforms.
Time scale 0.5μ sec/cm.

A LOW NOISE HIGH GAIN BANDWIDTH CHARGE SENSITIVE PREAMPLIFIER*

Jack Hahn, Ralph Mayer
Pegram Nuclear Physics Laboratories
Columbia University
New York 27, New York

Summary

The concept of the charge sensitive amplifier is discussed and a novel method for obtaining a high gain amplifier stage through the use of a positive current feedback dynamic plate load is described. It is shown that under certain conditions this type of feedback can provide increased bandwidth as well as increased low frequency gain. The schematic of a preamplifier embodying these principles is given and its performance is described. The measured noise, with no external capacitance, is 510 r.m.s. electrons for equal integration and differentiation time constants of 1 µsec. The measured risetimes for closed loop gains of 100 and 1000 are 25 and 96 nanoseconds respectively. Linearity and gain sensitivity data are also included.

Introduction

The nuclear physicist, when performing energy measurements, is generally interested in measuring the amount of charge delivered by a radiation detector. As long as the integration time constant associated with the detector output is long compared to the charge collection time, and as long as the total shunt capacitance associated with the integration time constant does not vary, the voltage developed by a fixed amount of charge will be constant, and will be given by the relationship V = Q/C. Thus even though the prime interest is in determining the value of Q, it is possible for the physicist to use a voltage amplifier, i.e. an amplifier whose output voltage is related to its input voltage by a constant, A.

With the advent of the solid state radiation detector the situation has changed. The capacitance of the solid state detector varies as a function of applied bias voltage. In order to take advantage of the inherently high energy resolution of this detector an amplifier is required whose output is truly a linear function of the collected charge. It is pointed out by Cottini, et al,[1] that an output-input relationship of this type can be obtained through the use of capacitive feedback. Amplifiers of this type, whose output voltage is proportional to the input charge have, for obvious reasons, been called "charge sensitive" amplifiers and a number have been described in the literature.[2]

This paper will describe a low noise, charge sensitive preamplifier, which has unusual gain bandwidth characteristics, good linearity, and low sensitivity to input capacitance variations.

The Charge Sensitive Amplifier

The charge sensitive amplifier configuration is shown in block form in Fig. 1a. The triangular block represents an amplifier of gain -A; C_f represents the feedback capacitor; C_D represents the capacitance associated with the detector, the amplifier input, and the wiring or cable capacity. The detector is represented by the charge source Q.

It will now be shown that the amplifier output is given approximately by $-Q/C_f$. It can be shown by simple circuit analysis that the effective input capacity obtained by placing C_f across the amplifier, from input to output, is $C_f(1 + A)$. This is indicated in Fig. 1b. Thus the effective input voltage, e_{in} is given by $Q/[C_D + C_f(1 + A)]$. If C_fA is much larger than C_D, e_{in} is given approximately by Q/C_fA. Since e_o is $-Ae_{in}$ it follows that $e_o \sim -Q/C_f$.

In the absence of the amplifier the detector output voltage is given by Q/C_D. With the amplifier present the voltage output is $-Q/C_f$. Thus the voltage gain due to the amplifier is $-C_D/C_f$.

The design of a charge sensitive amplifier presents problems not usually met in the design of the more conventional voltage sensitive amplifier. The typical voltage amplifier used in the nuclear field is a two stage non-inverting amplifier with negative feedback from the output to the cathode of the input tube. In order to optimize the transient response a phase lead is placed in the resistive negative feedback network so that the over-all loop behaves essentially as two lags plus a lead.

In the charge sensitive amplifier, however, the high frequency negative feedback network is a capacitive voltage di-

* Work partially supported by U. S. Atomic Energy Commission.

vider, $C_f/(C_f + C_D)$, and is frequency independent. Thus even a two stage amplifier offers design problems since the usually desirable phase lead cannot be obtained from the feedback network.

For the above reasons the charge sensitive amplifier has typically been designed about a single stage amplifier. This has meant that the open loop gain, -A, has not been as high as is desired. In order to circumvent this, most designers have resorted to a bootstrapped or dynamic plate load configuration.[3] This type of configuration makes use of positive feedback to raise the effective dynamic plate load and thereby raise the gain. The approach used in the amplifier being described is related in that positive feedback is also used to obtain an extremely high plate load. The method to be described, however, provides a higher gain and also affords a means of inserting a phase lead in the positive feedback loop so that under certain conditions the effective gain bandwidth is considerably enhanced.

A Comparison of Two Types of Dynamic Plate Load

Fig. 2a shows a simplified schematic diagram of the conventional bootstrapped plate load amplifier used in conjunction with a low noise cascode input section. The feedback provided by T_3 causes T_2, the upper tube of the cascode section, to see a dynamic impedance of approximately $\mu_3 R_L$ ohms.

Fig. 2b shows a simplified schematic diagram of the circuit we have been using. Again T_1 and T_2 are a conventional cascode input section. Now, however, the dynamic impedance seen by T_2 can be made extremely large and can in fact be made to approach infinity.

In Fig. 2a, T_3 in conjunction with R_L forms a dynamic plate load. Since the point X in Fig. 2a is a high impedance or voltage summing point we will refer to this circuit as the Positive Voltage Feedback Dynamic Plate Load (or PVF circuit). In Fig. 2b, the point X is a low impedance or current summing point. We will therefore refer to the circuit of Fig. 2b as the Positive Current Feedback Dynamic Plate Load (or PCF circuit).*

In order to see that the impedance looking upward at point X in Fig. 2b can approach infinity, we can reason as follows: let us break the circuit at point X and inject one ampere of current upward. This current will see two parallel paths. Into the cathode of T_3 it sees an impedance of $(R_L + r_{p3})/(\mu_3 + 1)$. Into R_f it sees an impedance of R_f in series with $r_{p4}/(\mu_4 + 1)$. Since $R_f >> (R_L + r_{p3})/(\mu_3 + 1)$ essentially all of the current flows into the cathode of T_3 and generates a voltage of R_L volts at the plate of T_3. Assuming that T_4 is a perfect cathode follower of unity gain, R_L volts appear at the cathode of T_4 and cause R_L/R_f amperes to flow through R_f and return to point X. If R_f equals R_L, then under the assumptions we have made, one ampere of current returns to point X. If we assume that this one ampere of current flows back into the source, it exactly cancels the original ampere which had been initially injected. In other words we have created a finite voltage at point X but have drawn no current from the source. This, in effect, says that we must have an infinite impedance at this point. If, in fact, the assumptions made above, namely, that all the injected current flows into the cathode of T_3, and that the cathode follower T_4 has unity voltage gain, are not valid, the impedance can still be made infinite by merely making R_f appropriately smaller than R_L. This will be elaborated upon below.

A more careful analysis of the resistance presented by the PCF plate load shows that it has the form of $R = \dfrac{r}{1 - L}$ where r is the impedance seen looking upward from point X in the absence of the positive feedback, and L is the loop gain of the positive feedback loop formed by T_3 and T_4. r is given approximately by the impedance seen looking into the cathode of T_3, that is $(R_L + r_{p3})/(\mu_3 + 1)$ while L is given approximately by

$$\frac{R_L}{R_f + \dfrac{R_f + r_{p_4}}{\mu_4} + \dfrac{R_L + r_{p_3}}{\mu_3}}$$

Earlier when we injected one ampere of current into point X and calculated the current which would be fed back, we were operationally defining what we mean by the value of L. At that time we found the loop gain to be approximately R_L/R_f. We now see that a more detailed analysis results in a value which differs only by the second order terms $(R_f + r_{p4})/\mu_4$ and $(R_L + r_{p3})/\mu_3$.

* Low impedance and high impedance as used here are relative terms, and refer to the impedance at point X in the absence of feedback. After feedback both points rise in impedance level in direct proportion to the amount of positive feedback used.

It can be seen that by properly choosing R_L and R_f, the value of L can in fact be made very close to unity, and thus $R = \frac{r}{1-L}$ can indeed be made to approach infinity.

It has been assumed in this discussion, and it will be assumed in the rest of this paper, unless otherwise indicated, that L is close to but never greater than unity. If L is greater than unity a negative impedance results. The subject of negative impedance loads is a complex one and no attempt will be made to deal with it in this paper other than to indicate that under certain conditions this can be a stable mode of operation. Some of the implications of this will be touched on briefly briefly below.

We shall now derive approximate gain expressions for both the PVF and PCF plate load amplifiers. In both circuits the open circuit voltage gain of the cascode amplifier T_1 and T_2 is approximately $-\mu_1\mu_2$. In the PVF amplifier, the load seen by the cascode is approximately $\mu_3 R_L$. By simple voltage divider action the voltage appearing at the plate of T_2 is approximately $\frac{-\mu_1\mu_2\mu_3 R_L}{r_{p1}\mu_2 + R_L\mu_3}$. The gain from the plate of T_2 to the output point is essentially unity. If we assume that identical tubes are used for T_2 and T_3 we may assume that $\mu_3 = \mu_2$. If this is true then the gain of the PVF plate load amplifier is given approximately by $\frac{-\mu_1\mu_2 R_L}{r_{p3} + R_L}$.

In the PCF amplifier (Fig. 2b) if the load presented to the cascode amplifier T_1 and T_2 approaches infinity the voltage appearing at the plate of T_2 is the input voltage multiplied by the open circuit voltage gain of the cascode amplifier. Thus, the voltage appearing at the plate of T_2 in the PCF case is approximately $-\mu_1\mu_2$ times the input voltage. The gain from the plate of T_2 to the cathode of T_4, the output point, is approximately $\frac{\mu_3 R_L}{r_{p3} + R_L}$. Thus the overall voltage gain of the PCF amplifier used with a cascode input is given by $A \cong \frac{-\mu_1\mu_2\mu_3 R_L}{r_{p3} + R_L}$. As R_L approaches infinity the PVF cascode combination approaches a limiting gain of $-\mu_1\mu_2$ and the PCF cascode combination approaches a limiting gain of $-\mu_1\mu_2\mu_3$. If, in the PCF amplifier L is made slightly larger than unity, then, as we have indicated, T_2 sees a negative load impedance. When a generator is terminated in a negative impedance, it is capable of supplying an output voltage greater than its open circuit output voltage. Under these conditions the gain is no longer limited to $-\mu_1\mu_2\mu_3$. The gain can now be made to approach infinity.

The High Frequency Gain of the PCF Amplifier

The discussion so far has dealt only with the low frequency gain of the PCF amplifier. Positive current feedback where the feedback element is a pure resistance, R_f, affects only the low frequency gain of the circuit. In Fig. 3 curve 1 represents the frequency response of the amplifier of Fig. 2b in the absence of the positive feedback ($R_f = \infty$). Curve 2 shows the frequency response of the same amplifier with a more typical value of R_f. It can be seen that the high frequency behavior is identical under the two conditions of operation. Obviously, as the gain of the positive feedback loop falls at higher frequencies, due to the unavoidable capacitive loading of the circuit, the positive feedback becomes ineffective.

There exist numerous capacitances in the PCF circuit but the dominant one is the capacitance from the plate of T_3 to ground. This capacitance implies that the plate load of T_3, R_L, is in reality an impedance $Z_L = R_L/(1+j\omega R_L C_L)$. Since R_L appears in the numerator of the expression for the loop gain, and R_f appears as the dominant term in the denominator it is apparent that by making R_f frequency dependent as well, the entire expression for L can be made less sensitive to frequency. Thus the high frequency loop gain of the positive feedback loop can be improved by shunting R_f with a capacitor C_o (See Fig. 2). When this is done typical frequency response curves such as 3, 4, and 5 in Fig. 3 are obtained as C_o is increased. It is to be noted that at extremely high frequencies the gain has not been affected. However, at medium frequencies the gain is improved considerably and under certain conditions the usable gain bandwidth has been extended.

It should be noted that the open loop gain at which the -20 db. per decade slope changes to a -40 db. per decade slope increases for each successively higher curve. The gain at which this slope change occurs is a critical one in describing the closed loop behavior of the amplifier, i.e., the behavior of the amplifier after negative feedback is applied around it. At the point where the slope change occurs, the open loop amplifier has approximately 135° phase shift. If negative feedback is used to obtain a closed loop gain equal to the open loop gain at the point of change of slope the amplifier will be stable but its transient response will exhibit considerable ringing. To obtain a more desirable transient response a closed loop gain larger than the gain at the point of change of slope must be used. This reduces the loop phase shift and results in improved transient response.

With this transient response limitation in mind, Fig. 3 shows that while each successively higher curve corresponds to a successively higher effective gain bandwidth each successively higher curve also dictates a higher value of closed loop gain.

As far as the user is concerned, Fig. 3 shows that as the value of C_D/C_f, the closed loop gain, is increased, the value of C_o can also be increased, resulting in operation on one of the higher curves and consequently in improved high frequency response. This will become more obvious later when risetime is plotted against closed loop gain. For most conventional feedback amplifiers the gain bandwidth product is a constant and the closed loop risetime varies directly as the closed loop gain. In the PCF amplifier the effective gain bandwidth can be increased with increasing closed loop gain, consequently the risetime increases more slowly than the gain.

Circuit Details and Performance

A complete schematic diagram of our most recent amplifier including a White cathode follower output stage is given in Fig. 4.

The following features should be noted: Negative D.C. feedback is used around the PCF amplifier for operating point stabilization. The two 2N652A transistors effectively provide a +27 volt power supply to which the cathode of T_1 is referenced.

T_3 conducts approximately 7.5 ma of current, and T_4 approximately 8 ma of current. T_1 and T_2 conduct the total current of 15.5 ma. Two halves of a 7308 are used in parallel for T_2 so that the 15.5 ma may be passed with only 68 volts across the tube. A somewhat minor advantage of the configuration used is that a high current flows through T_1, thus enabling its transconductance to be high but only half of the current flows through T_3 and R_L (and T_4 and R_f) thus making it possible to keep the total supply voltage moderate despite the fact that three tubes are stacked one upon another.

The open loop low frequency gain is critically dependent upon the ratio of R_L to R_F. R_L is a 9K wirewound, Ohmite 3 watt axial lead resistor. R_f is a matched 9K resistor of the same type paralleled by a 1%, 91K metal film resistor. The Ohmite resistors are remarkably uniform and can be readily matched, using a Wheatstone bridge, to within 10 ohms or less. Typically out of 10 resistors we have been able to find 3 pairs of resistors matched to within 10 ohms.

The amplifier input capacitance is approximately 35 pf. The value of C_f used depends upon the desired closed loop gain. The ratio of the total capacitance at the amplifier input to C_f should be kept greater than 15 or the amplifier will oscillate. We have typically used a value of C_f of 1 pf for small detectors. For values of C_f up to approximately 2.5pf the amplifier is stable even with no detector connected to the amplifier input, i.e., the amplifier input capacitance itself insures stability.

The noise characteristics of the amplifier are determined by the type of input tube used and the operating point at which this tube is used. The equivalent noise charge of the amplifier in RMS electrons is plotted against external capacitance in Fig. 5. Equal integrating and differentiating time constants, τ_1 and τ_2, of 10^{-6} sec were used. The charge sensitivity was determined by feeding in a known voltage through the 1 pf test input capacitor. The noise output was measured with a Hewlett-Packard type 400H VTVM. A full description of this measuring technique has been given by Fairstein.[4]

For a particular detector, and specific values of τ_1 and τ_2 it may be desirable to modify the operating point of the input tube T_1 or to substitute another tube for T_1. If it is desired that only the current through T_1 be modified and that all other operating points remain fixed this can be accomplished by placing either a current source or sink at the plate of T_1. A current source can simply be a resistor going to B+, and a current sink can be a resistor going to ground. As long as the value of this resistor is large compared to the plate resistance of T_1 the open loop gain of the amplifier will not be seriously affected. It may also be necessary to modify the 2.7K, and 3K resistors in the 27 volt transistor power supply.

The computed d.c. open loop gain,

assuming L $\sim$ 1 was $\frac{-\mu_1\mu_2\mu_3 R_L}{r_{p_3} + R_L}$. For the tubes and load resistors used (μ_1 = 56, $\mu_2 = \mu_3$ = 28, r_{p_3} = 3K, R_L = 9K)[1] a gain of 33,000 is predicted. Using resistors matched as indicated above typical values of measured low frequency gain varied from 15,000 to 30,000 as tubes were changed. For the values indicated the circuit was always low frequency stable. When the value of R_f was decreased in small increments d.c. gains as high as 260,000 were measured before low frequency instability occurred. The reason why gains larger than 33,000 can be obtained was indicated earlier. When R_f is decreased in small increments L can be made slightly larger than unity. The dynamic load impedance becomes negative and the low frequency gain can now approach infinity.

Fig. 6 presents a plot of closed loop risetime versus closed loop gain. The feedback capacitor, C_f, was fixed at 1 pf. The desired closed loop gain was obtained by placing the appropriate value of capacitance across the amplifier input, and the capacitor, C_o, in the positive feedback loop was adjusted so that the output waveform had maximum risetime with zero overshoot. It should be noted that the risetime increases more slowly than the gain.

Table 1 presents data on the amplifier "reserve gain." The open loop gain at 1 µsec and at 10 µsec is tabulated as a function of the setting of C_o. For each pair of measurements C_o was set at the value corresponding to maximum risetime without overshoot for the closed loop gain listed in the first column of the table. It can be seen that for a setting of C_o corresponding to a closed loop gain of 1,000 the gain at 1 µsec is 3 times the value corresponding to a setting of C_o for a closed loop gain of 35.

The linearity of the amplifier was measured by using a mercury pulser and a 256 channel multichannel analyzer. The multi-channel analyzer was used so that each channel width was 0.1% of the input. A pulse was fed into the preamplifier test input. It was amplified and then shaped by equal integration and differentiation time constants of 1 µsec. The shaped signal was then fed into the multi-channel analyzer and the centroid of the resulting pulse amplitude distribution was computed. The pulser was then moved to the preamplifier output and readjusted until the pulse once again fell into approximately the same channel. The new centroid was computed. The ratio of the two pulser settings corrected for the centroid offset was taken as the amplifier gain. When used as a closed loop gain of 100 the gain for output voltages up to 2.2 volts varied less than 0.1%. This is the order of the expected measurement errors.

Table 2 gives data on the sensitivity of the closed loop amplifier gain to variations in filament voltage, B+ voltage, and detector capacitance. A fixed charge was fed in through the 1 pf capacitor. C_f was also 1 pf. The gain was measured as in the linearity tests described above. A reference gain was measured with a B+ voltage of 350 volts, filament voltages of 6.8 v.a.c. and a total input capacitance of 100 pf resulting in an effective closed loop gain of 100. The amplifier output was maintained at a nominal 2 volts. The table lists deviations from the reference gain in percent.

It can be seen that for independent variations of the a.c. filament voltage from 7.5 to 5.1 volts and for B+ variations of approximately ± 15% the maximum gain change was 0.15%.

When the total input capacitance was increased from 100 to 200 pf the gain decreased approximately 0.8 percent.

The output voltage sensitivity to input capacitance variation is perhaps the most significant criterion of a charge sensitive amplifier since it measures directly the effect the amplifier was designed to minimize. It can be shown that even though the amplifier output is passed through a 1 µsec differentiating network, it is nevertheless the low frequency gain of the amplifier which is of importance in minimizing output voltage sensitivity to input capacitance variation. The output voltage of a charge sensitive amplifier is given by

$$\frac{-Q_{in} \quad A_o \qquad (1 - e^{-t/\alpha\tau})}{C_{in} + C_f\,(1 + A_o)}$$

where A_o is the low frequency gain of the open loop amplifier, τ is the amplifier open loop time constant (assuming a single time constant amplifier), and "α" is, for large A_o, approximately $\frac{C_{in}}{A_o C_f}$. For A_o approximately 25,000, C_{in}/C_f approximately 100 and τ approximately 25 µsec we find that $\alpha\tau$ is approximately 0.1 µsec. Thus at t = 1 µsec the term $e^{-t/\alpha\tau}$ is $e^{-10} \sim 5 \times 10^{-5}$ and is consequently negligible. As a result the gain is stabilized by the full low frequency gain A_o. We have experimentally investigated this as follows. Starting with an amplifier whose sensitivity is essentially like that described in Table 2, i.e., where the gain decreases for a 100 pf increase of input capacitance by about something less than 1% we have reduced R_f in small increments, thus increasing the low frequency gain. As R_f is decreased the output voltage change becomes smaller and smaller until finally a point is reached where the output voltage does not change at all when capacitance is shunted across the input. Apparently at this point the open loop gain has become infinite. If R_f is decreased still more, the output once more changes, but now the output increases when the capacitance is increased. This indicates that the amplifier is now open loop low frequency unstable. Attempting to operate the amplifier about the point of infinite open loop gain seems to have no disadvantage except for the fact that the amplifier may oscillate for low resistive impedances (10K or less) shunted across the input. For the latter reason we have chosen to operate the amplifier well within the range of finite gain.

Should extremely low sensitivity to input capacitance variations be necessary it is probably desirable to carefully pad R_f to obtain extremely high open loop gain.

Layout

Fig. 7 is a photograph of the bottom of the amplifier. The input UHF connector is at the left. The signal output, test signal input, and power leads all go through the large cannon connector at the right. For the most part, point-to-point wiring is used, with standoffs wherever convenient. Lead lengths were kept as short as possible, particularly at the plate of T_3 and the cathode of T_1. All grounds are made directly to a heavy bus wire, part of which may be seen in the photograph. Initially this bus was grounded only at the input UHF connector, however it was found that 60 cps noise spikes were minimized when the bus was grounded at the output connector as well.

A copper shield, at the left, is used to separate the input tube, T_1, from all the other tubes. The feedback capacitor C_f passes through a hole in this shield. This is done to minimize any capacitive feedback from output to input other than that through C_f. In the newer models, such as the one pictured, C_f is a Johanson type 1802 variable air capacitor.

The positive feedback loop trimmer capacitor, C_o, can barely be seen in the photograph. It is located to the right of the copper shield and to the left of the vertical bus wire. It is adjusted from the top of the chassis. In the lower portion of the photograph, a small printed circuit board may be seen. The two transistors and associated components indicated in Fig. 4 are mounted on this board. Below the transistors, and not visible in the photograph, are some screened ventilating holes. The newer models will have additional holes to further reduce heating. A perforated bottom plate having small rubber, bumper type, feet as well as the use of IERC heat dissipating tube shields have also been found effective in reducing chassis and tube temperature.

Acknowledgments

The authors would like to thank all the engineers in the Electronics Group at the Pegram Laboratory for many stimulating discussions regarding the PCF mode of operation. We would also like to thank Mr. Thomas Becker for taking many of the data, Mr. Ramon Padilla for doing the layout and construction work, and Mr. Jerry Packer for the printed circuit work.

References

1. C. Cottini, E. Gatti, G. Giannelli, and G. Rossi, "Minimum Noise Preamplifier for Fast Ionization Chambers," Il Nuovo Cimento (Ser. 10), Vol. 3, No. 2, pp. 473-483; Feb., 1956.

2. G. G. Kelly, "A New Amplifier for Pulse Spectrometry," IRE National Convention Record 1957, part 9. See also papers by J. L. Blankenship and C. J. Borkowski; T. L. Emmer; and R. L. Chase, W. A. Higinbotham and G. L. Miller; published in IRE Transactions on Nuclear Science, Vol. NS-8, Number 1, Jan. 1961, Proceedings of the Seventh Annual National Meeting, Solid State Radiation Detectors.

3. G. E. Valley, Jr., and H. Wallman, "Vacuum Tube Amplifiers," Volume 18, Radiation Laboratory Series, McGraw Hill, 1948, pp. 456-8.

4. E. Fairstein, "Considerations in the Design of Pulse Amplifiers for Use with Solid State Radiation Detectors," IRE Transactions on Nuclear Science, Vol. NS-8, Number 1, Jan. 1961, Proceedings of the Seventh Annual National Meeting, Solid State Radiation Detectors, pp. 129-39.

TABLE 1

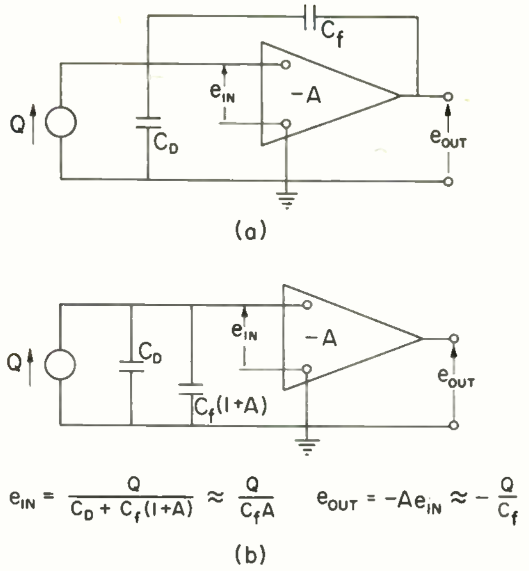
Open loop gain at 1 μsec and 10 μsec as a function of C_o

C_o set for closed loop gain of:	Open Loop Gain	
	1 μsec	10 μsec
35	4,250	23,800
50	4,250	23,800
100	5,260	26,300
200	7,140	29,800
500	10,200	32,200
1,000	13,500	33,000

TABLE 2

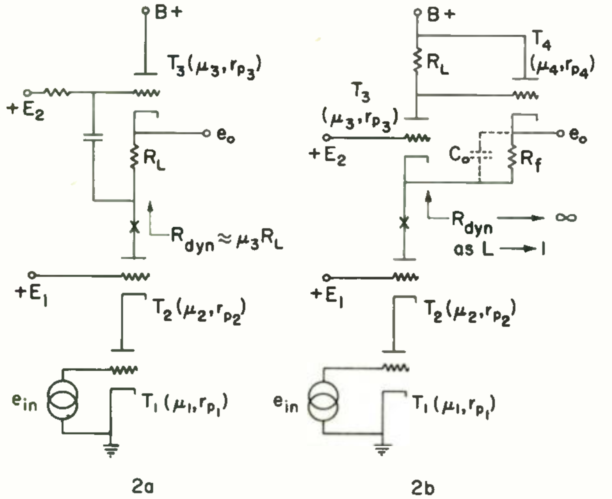
Closed loop gain variations as a function of changes in filament voltage, B+ voltage, and input capacitance (amplifier output 2v)

Filament Voltage	B+ Voltage	Total Input Capacitance	Per cent deviation from reference gain
6.8 v.a.c.	350 v.	100 pf	----
6.8 v.a.c.	300 v.	100 pf	- 0.15
6.8 v.a.c.	400 v.	100 pf	+ 0.05
5.1 v.a.c.	350 v.	100 pf	< 0.05
7.5 v.a.c.	350 v.	100 pf	< 0.05
6.8 v.a.c.	350 v.	200 pf	- 0.79
6.8 v.a.c.	350 v.	1,100 pf	- 4.35



$$e_{IN} = \frac{Q}{C_D + C_f(1+A)} \approx \frac{Q}{C_f A} \qquad e_{OUT} = -Ae_{IN} \approx -\frac{Q}{C_f}$$

Fig. 1. (a) Block diagram representation of charge sensitive amplifier. (b) Equivalent circuit of Fig. a.



Positive voltage feedback
Dynamic Plate Load

Positive current feedback
Dynamic Plate Load

Note: μ and r_p refer respectively to the amplification factor and plate resistance of the appropriate tubes.

Fig. 2. A comparison of two types of dynamic plate loads.

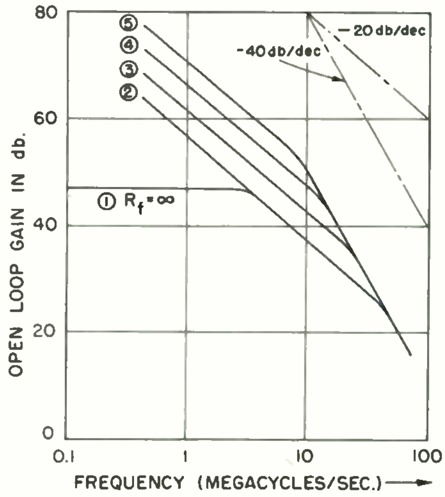
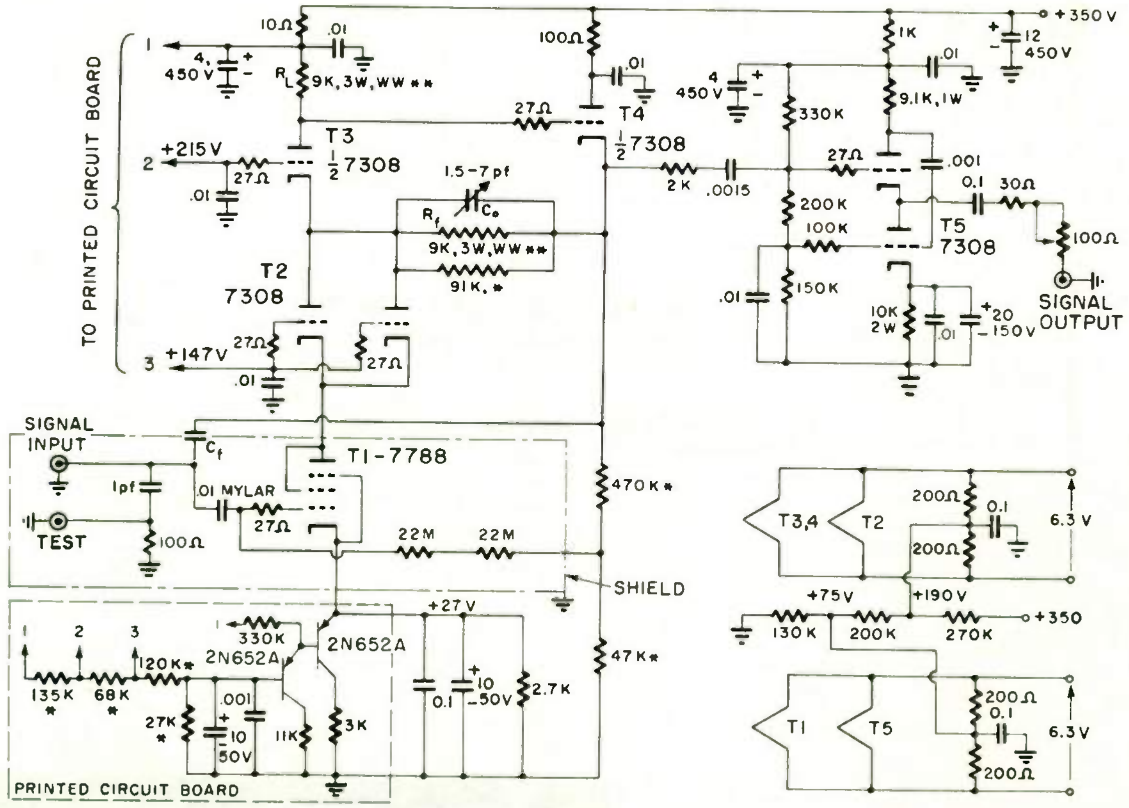


Fig. 3. Open loop gain vs. frequency for typical P.C.F. amplifier.



* 1% RESISTORS, CORNING TYPE S20

** MATCHED RESISTORS, SEE TEXT

ALL CAPACITORS IN μf UNLESS STATED

ALL RESISTORS 1/2W UNLESS STATED

Fig. 4. Schematic diagram of charge sensitive preamplifier.

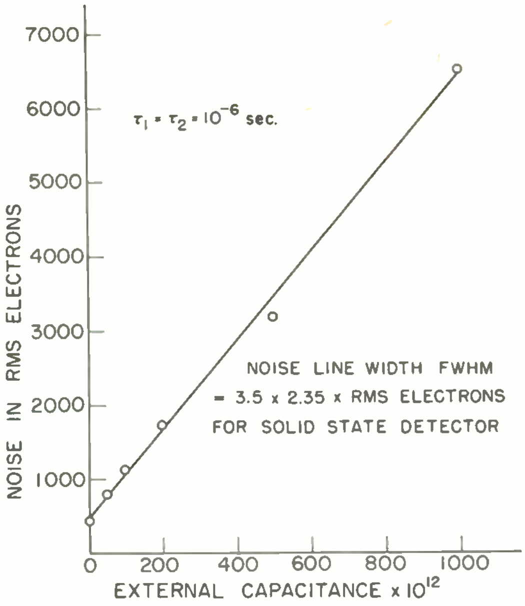


Fig. 5. Noise in RMS electrons vs. external capacitance.

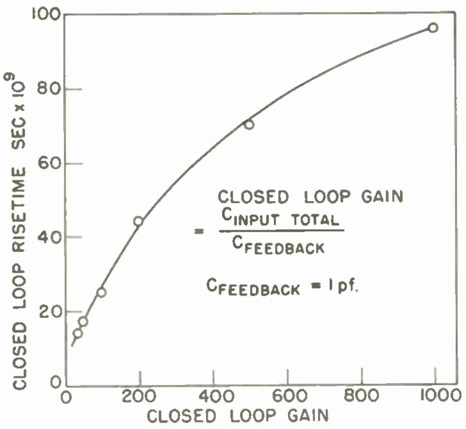


Fig. 6. Closed loop gain vs. risetime.

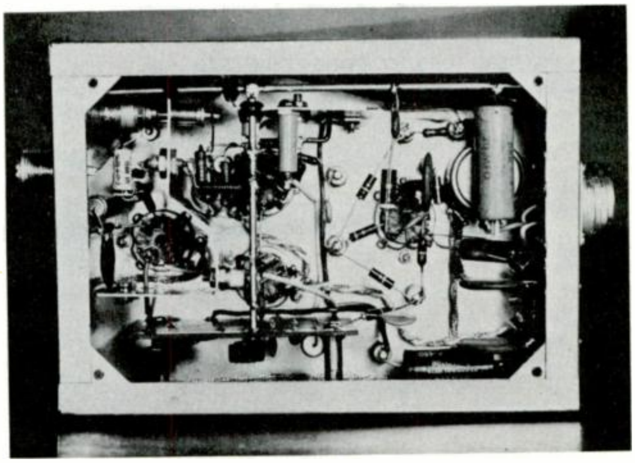


Fig. 7. Bottom view of amplifier chassis.

A TRANSISTORIZED REACTIVITY COMPUTER

R. F. Shea
General Electric Company
Knolls Atomic Power Laboratory
Schenectady, New York.

ABSTRACT

The usual reactor kinetics equations are shown to consist of three major components, a fast term; a product or quotient, and a summation of transient terms. These can be synthesized by using a number of simple R.C. networks having time constants related to the fast time constant and to the decay constants of the various groups of delayed neutron emitters, plus either an analog multiplier or divider. If an input signal proportional to neutron flux or reactor power is applied to this network, the output will be proportional to reactivity.

A circuit is shown which incorporates these principles, using a diode divider. Curves are presented of output reading, indicating transient reactivity, for input currents having constant positive period, as obtained from a motor-driven potentiometer. It is shown that these curves are in reasonable agreement with calculated curves of reactivity for such periods.

INTRODUCTION

Nuclear instrumentation generally takes the form of measurement of neutron flux and reactor period by means of such devices as proportional counters, ion chambers, logarithmic amplifiers, count rate meters, etc. It is assumed that the pulse rate or ion chamber current, as the case may be, are proportional to reactor power, this assumption thereby ignoring such effects as rod shadowing or temperature effects on coolant, etc.

The period meter is used to guard against too-rapid power excursions, however, by its very nature it has a transient response which is different from that of the reactor. Thus if a reactor were suddenly to assume an exponential rate of change from a previous steady state, the period meter would not present a step function but would respond with an exponential rise, requiring several periods before it would correctly indicate the true value.

Another factor to consider is that the reactivity is not proportional to inverse period, except approximately at very low values of reactivity as illustrated on Fig. 1. Therefore, period, or more rigorously, inverse period, is only an approximate indication of reactivity and a true reactivity computer should be of considerable value in observing the effects of rod motion or other factors affecting reactor power. Such a reactivity computer can be synthesized by a combination of time constant circuits and analog circuits (multiplier or divider) as shown below.

THE REACTOR KINETICS EQUATIONS

In a source-less reactor or one operating at high power where the effects of the source is negligible the reactor kinetics equations are given as follows:

$$\frac{dn}{dt} = \frac{n(\delta k_e - k_e \beta)}{\ell} + \sum_{i=1}^{6} \lambda_i C_i \qquad (1)$$

$$\frac{dC_i}{dt} = \frac{k_e \beta_i n}{\ell} - \lambda_i C_i \qquad (2)$$

where

n = neutrons per cc
k_e = effective multiplication
δk_e = $k_e - 1$
C_i = concentration of i^{th} group of delayed neutron emitters ($i = 1$-6)
λ_i = decay constant of i^{th} group
β = total fraction of delayed neutrons
β_i = total fraction of delayed neutrons of i^{th} group
ℓ = mean effective lifetime of thermal neutrons

These equations may be normalized by use of alternative parameters τ_o, y_i, f_i and Δk where

$\tau_o = \frac{\ell}{k_e \beta} \cong \frac{\ell}{\beta}$ is the neutron time constant

$y_i = \frac{\ell C_i}{k_e \beta} \cong \frac{\ell C_i}{\beta}$ is the normalized value of delayed emitter

$f_i = \frac{\beta_i}{\beta}$ is the fraction of the i^{th} group

$\Delta k = \frac{\delta k_e}{k_e \beta}$ is excess multiplication in dollars

(Note: Δk is approximately equal to the reactivity $\rho = \frac{\delta k_e}{\beta}$ and will be called reactivity in this paper).

Using these parameters the original kinetics equations become

$$\tau_o \frac{dn}{dt} = n\,(\Delta k - 1) + \sum_{i=1}^{6} \lambda_i\, y_i \qquad (3)$$

$$\frac{dy_i}{dt} = f_i\, n - \lambda_i\, y_i \qquad (4)$$

Expressing eq. 4 in operational form, using $p = \frac{d}{dt}$:

$$p\, y_i = f_i\, n - \lambda_i\, y_i \qquad (5)$$

$$\text{whence } y_i = \frac{f_i\, n}{p + \lambda_i} \qquad (6)$$

$$\text{and } \lambda_i\, y_i = f_i\, n - \frac{p\, f_i\, n}{p + \lambda_i} \qquad (7)$$

The summation $\sum \lambda_i\, y_i$

$$\sum \lambda_i\, y_i = \sum f_i\, n - \sum \frac{p\, f_i\, n}{p + \lambda_i} \qquad (8)$$

$$= n \qquad - \sum \frac{p\, f_i\, n}{p + \lambda_i}$$

Substituting in (3):

$$\tau_o\, pn = n\,(\Delta k - 1) + n - \sum_{i=1}^{6} \frac{p\, f_i\, n}{p + \lambda_i}$$

$$= n\,\Delta k - \sum_{i=1}^{6} \frac{p\, f_i\, n}{p + \lambda_i} \qquad (9)$$

Δk may, therefore, be obtained as

$$\Delta k = \frac{1}{n}\left(\tau_o\, pn + \sum_{i=1}^{6} \frac{p\, f_i\, n}{p + \lambda_i} \right) \qquad (10)$$

Δk may be derived by taking a summation of a prompt term τ_o pn and a number of terms of form $\frac{p\, f_i\, n}{p + \lambda_i}$ and dividing by the factor n.

(Note: Alternatively Δk can be multiplied by n and the product equated to the summation. This requires a multiplier and operational amplifier. The output of the amplifier is Δk and is one of the inputs to the multiplier[2]).

SYNTHESIS OF THE SUMMATION TERMS

The term $\frac{p\, f_i\, n}{p + \lambda_i}$ has the same form as the output of a simple differentiating RC circuit with attenuator as shown in Fig. 2. For such a circuit the response function is

$$\frac{V_o}{V_i} = \frac{p\, RC}{1 + p\, RC} \qquad (11)$$

If RC is $= \frac{1}{\lambda_i}$ and R is tapped at a point corresponding to the fraction f_i the voltage ratio becomes

$$\frac{\frac{p\, f_i}{\lambda_i}}{1 + \frac{p}{\lambda_i}} = \frac{p\, f_i}{p + \lambda_i} \qquad (12)$$

and if V_i corresponds to neutron flux n we have $\frac{p\, f_i\, n}{p + \lambda_i}$ for the output, which is of the desired form. The reactor kinetics terms represented by the summation in equations (9) and (10) may, therefore, be synthesized by a mesh of RC circuits.

Referring again to our RC circuit, if the portion of R below the tap has a resistance R_o the current through it will be

$$I = \frac{p\, f_i\, n}{R_o(p + \lambda_i)} \qquad (13)$$

If we now arrange a number of RC circuits as shown in Fig. 3 (for a six-group simulation), and choose the constants such that $C_i\, R_i = \frac{1}{\lambda_i}$ and $f_1\, R_1 = f_2\, R_2 = f_i\, R_i = R_o$, then the output current I_o is a summation

$$\sum_{i=1}^{6} \frac{p\, f_i\, n}{p + \lambda_i}$$

and we have one of the terms of equations (9) and (10). The τ_o pn term may be closely approximated by adding one more RC mesh having a time constant τ_o and $R = R_o$ above. The overall operation may, therefore, be synthesized as shown in Fig. 4.

ELECTRONIC ANALOG DIVISION

From the foregoing, it is obvious that a vital portion of the reactivity computer is the divider. A number of techniques have been developed for performing this operation, however, an extremely simple one has been described by Smith and Prabhakar.[3] This is based on the exponential relationship between current and voltage for a diode at small signals. If $V_o = a \log \frac{I_o}{b}$ where V_o and I_o are D.C. values, and if an A.C. voltage V cos pt is also applied the current is $I_o \exp(\frac{V}{a} \cos pt)$. This may be expressed as a series:

$$I = I_o \left[1 + \frac{V}{a} \cos pt + \frac{1}{2} \left(\frac{V}{a} \cos pt\right)^2 + \cdots \right] \qquad (14)$$

If the A.C. amplitude is kept small the higher order terms are minor and the A.C. current through the diode becomes proportional to the product $I_o V$. Thus such a device may be used as a multiplier or divider. To use it as a multiplier the D.C. current is obtained through a high resistance, hence is essentially proportional to the applied D.C. voltage. The A.C. voltage is applied directly, through a large capacitor. In a practical application of this technique the diode becomes the input of a transistor and thus the collector current becomes a measure of the product.

When used as a divider the D.C. and A.C. currents are both applied through large values of resistance and in this application the A.C. voltage across the diode is proportional to the quotient $\frac{V_{AC}}{I_{DC}}$ (or $\frac{V_{AC}}{V_{DC}}$ since I_{DC} is proportional to V_{DC} due to the large resistor).

Fig. 5 shows the circuit of a diode divider and associated amplifier, with a typical response curve. In this case the diode was a type 1N457, the D.C. voltage was applied through a resistance of 100,000 ohms and the A.C. voltage was a 1000 cycle square wave, applied through one megohm and a 1 microfarad capacitor. The output A.C. voltage appearing at the collector of the 2N338 had an amplitude given by the equation

$$V_o = \frac{V_{AC}}{2V_{DC}}$$

V_{AC} could range from 1 to 20 volts, V_{DC} from 2-20^v. The limitations are imposed by the diode characteristic and transistor saturation. For values of V_{DC} below 2 volts the curve of V_o vs. $1/V_{DC}$ becomes non-linear. The above ranges indicate that this circuit should be suitable for a reactivity computer which operates in the power range and where the power is maintained between 5 and 100%.

TRANSIENT REACTIVITY COMPUTER USING DIODE DIVIDER

The basic building blocks described above have been assembled into a complete transient reactivity computer, the schematic of which is shown on Fig. 6. Operation of the various portions of this circuit is as follows:

The signal corresponding to flux is applied to the input of the first 2N336. This signal should be in the form of a current rather than voltage, hence proper excitation would be from an ion chamber. Maximum sensitivity is 14 microamperes full scale and the sensitivity control can reduce this as desired.

The amplified current from the 2N336 emitter follower is applied to the emitter of the 2N336A common-base stage and acts to vary its collector current and thus the voltage output. The bias on the base of the 2N336A is adjusted so that all the 30 volt supply is dropped in the 22K resistor when the input current is zero, thus making the output voltage zero for this condition. (Actually the output voltage is set for -0.4^v for I_{in}=0 to insure linear relationship between V_o and I_{in} over a major portion of the range - see Fig. 7). The meter is shunted to read 1 ma for I_{in}=0 and about 0.1 ma at full input signal.

The output voltage from this two-transistor preamplifier is applied to the kinetics network and also to the diode divider through 100,000 ohms. The kinetics network is designed to provide the equivalent of the two-group parameters, plus the τ_o term and produces an output across the 2700 ohm resistor proportional to all terms in equation (10) except the $\frac{1}{n}$ factor.

The output of the network is converted to A.C. by a diode capacitance modulator. This consists basically of two diodes and two resistors. The diodes are switched by application of a 15 KC square wave, however, the bridge balance results in no A.C. being imposed on the A.C. amplifier unless there is a D.C. (or low-frequency transient) signal applied across the diodes. Application of this signal produces unequal values of diode capacitance and thus an A.C. output results. The six ohm potentiometer provides the null adjustment (as noted one of the arms should contain a variable portion to center the null on the 6 ohm potentiometer). The 25 pf capacitor was added to improve the quality of the null.

The variable inductor has a range of about 0.1-0.5 henries and resonates the bridge capacitance, plus that of the amplifier, to produce maximum A.C. amplitude, when the bridge is unbalanced by imposition of a signal. The A.C. amplifier consists of two cascaded common-emitter stages with D.C. bias feedback for stabilization.

The output of the A.C. amplifier is applied to the diode divider through the 1 microfarad capacitor and 1 megohm resistor. The A.C. voltage developed across the diode is proportional to this applied A.C. signal and inversely to D.C. input, which represents flux. Thus this A.C. signal is proportional to Δk. It is converted to D.C. by means of the diode ring demodulator and finally amplified by a two-stage amplifier consisting of two 2N335's in a Darlington arrangement. The two 1N462 diodes serve to set the output voltage at zero when the input voltage is zero, also compensate for temperature effects on the transistor emitter-to-base voltages.

In setting up this circuit the bridge is first balanced so that no A.C. appears at the output of the A.C. amplifier for steady-state conditions. (Actually harmonics will be present and the null is obtained by viewing the output on an oscilloscope and adjusting for a minimum of the fundamental). The preamplifier is adjusted as stated previously to obtain linear proportionality between input current and output voltage (-0.4^v for I_{in}=0). The output amplifier is adjusted to provide zero output for zero input volts. The only adjustment remaining is that of phase relationship between the two signals applied to the demodulator so that the output zero will correspond to the zero of A.C. amplifier output. In the instrument this was accomplished by synchronizing a double-beam oscilloscope from the voltage across one of the oscillator base diodes and looking at the oscillator excitation to the demodulator and the output of the audio transformer simultaneously on the two beams. The 0.02 mfd capacitor across one winding of the oscillator transformer was used to bring these two waveforms into approximate synchronization.

As constructed the reactivity meter indicates positive reactivity with a ratio of meter scale reading to Δk of about 1.5 : 1, i.e. a reactivity of 40 cents produces about 0.6 ma. The zero can be moved upscale so that negative reactivity can also be read on the meter.

EVALUATION OF REACTIVITY METER

In order to provide a means of evaluating this and other forms of reactivity meter a source of exponentially-increasing current or voltage was constructed. This consists of an exponential potentiometer connected to a battery, together with additional fixed resistances to set up the initial conditions. For a current signal a series resistance of several hundred thousand ohms is used. The potentiometer is driven continuously by a variable-speed D.C. motor through a worm and pinion. The motor speed is adjusted by varying its applied voltage to obtain desired periods. The circuit is as shown in Fig. 8.

The switch S is used to hold the input signal at a fixed level, corresponding to the initial part of the exponential rise, until the transient is to be recorded. A recorder may be connected across the 100 ohm resistor to provide a record of actual current input to the amplifier. The battery voltage V, the series resistor R_1 and the minimum-setting resistor R_2 are adjusted to obtain the proper range of current (or voltage, in which case R_1 is omitted) and also proper exponential curve shape. In this connection it should be remembered that transistor amplifiers have input currents and thus some current will flow even without the battery V. It usually requires running quite a few curves, varying V_1, R_1 and R_2, until the proper curve is obtained. Once this has been done the relationship between period of the exponential signal and the motor speed can be determined and different periods can be chosen at will by varying the motor speed. (Motor speed is obtained by timing gear revolutions). In operation the motor speed is set for the desired value, switch S is held on position 1 until all transients have subsided, then when the potentiometer has reached the fixed contact at the bottom of its exponential rise (as indicated by a mark on the pinion gear) the switch is thrown to position 2 and the input now proceeds through an exponential excursion. Fig. 9 shows a tracing of a recording of the input current supplied to the circuit of Fig. 6 for an input period of about 10 seconds.

If the two-transistor preamplifier is operating in its linear range, its output voltage will be a replica of the input current. Fig. 10 shows a tracing of the output voltage curve obtained under the same conditions as Fig. 9.

Fig. 11 shows the output obtained at the recorder terminals for the input signal shown in Figs. 9 and 10. It will be noted that this voltage rises more or less exponentially toward a final assymptotic value which will be related to the period, but not linearly. Also shown on this figure is a calculated curve. This curve is a plot of the solution to the kinetic equation (eq. 10) for an input flux having the form

$$\phi = \phi_o e^{at}$$

a being the inverse period. The solution for this case is

$$\Delta k = \frac{a}{a+\frac{1}{\tau_o}}\left[1-e^{-(a+\frac{1}{\tau_o})t}\right] + a\sum_{i=1}^{2-6}\frac{f_i}{a+\lambda_i}\left[1-e^{-(a+\lambda_i)t}\right] \qquad (16)$$

It will be seen that the fit of the observed and calculated curves is reasonably good, considering the inaccuracy within which the period is maintained constant. The dip in the observed curve is presumably due to a discontinuity in the potentiometer where sections are joined together. Its presence can barely be noticed on the input curves but it becomes very pronounced when operated upon by the Δk circuit.

Figs. 11 through 14 show tracings of additional curves where the period was readjusted to approximately 20, 40 and 60 seconds, respectively. Also shown on Fig. 14 is a calculated curve for this period and again the agreement is reasonably good.

One problem in operating a device such as the Δk meter is that of testing for zero balance or drift. This is simple when it is known that the input is following a steady-state, either some constant value or zero. However, where this is not known, as for example when monitoring a reactor, some technique becomes necessary for ascertaining that balance corresponds to a true static condition. This may best be done by grounding the capacitors in the kinetics network where they connect to their respective resistors. This removes the transient input terms and allows a balance of the zero adjustment.

CONCLUSION

The above tests have indicated that it is feasible to construct a relatively simple Δk computer based on the diode divider, which will produce a reasonably accurate indication of transient reactivity from initiation to completion, as long as the input level remains within the range of linear operation of the divider and amplifiers (approximately from 1 to 13 microamperes in this design). The degree of accuracy is difficult to determine by the techniques employed as it is impossible to accurately calculate the transient period from a recording such as Fig. 9. This is evident from the tracings of Δk versus time, where the almost imperceptible discontinuity in input current shows up as a distinct dip and overshoot in Δk.

REFERENCES

1. Nuclear Reactor Physics, RL Murray, Prentice-

Hall, N. J. 1957.

2. Design and Use of the Reactivity Computer, Stubbs, G.S., Trans. IRE, NS-4, 1, pp. 40-48 (March 1957).

3. Multipliers and Dividers in A.C. Computers, Smith, C.H., and Prabhakar, A., Electronic Engineering, 32, 393, pp. 714-716 (Nov. 1960).

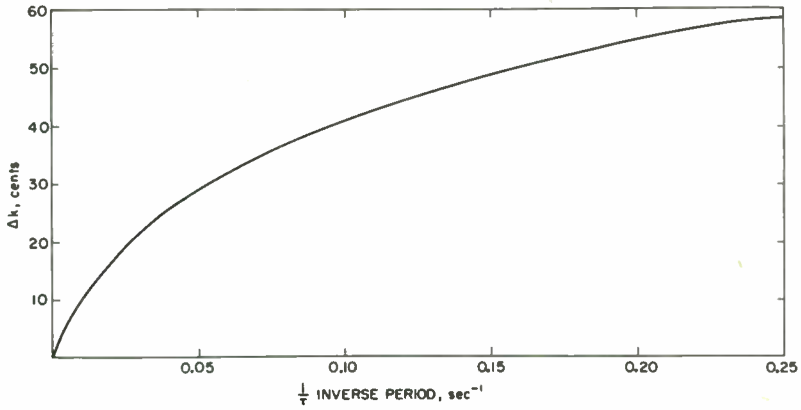


Fig. 1.

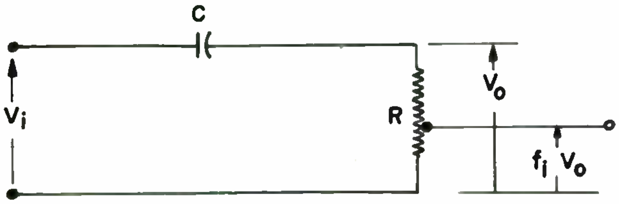


Fig. 2.

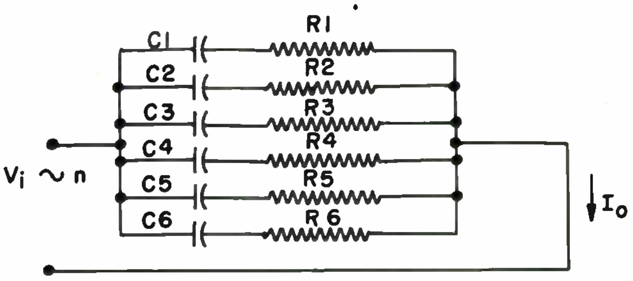


Fig. 3.

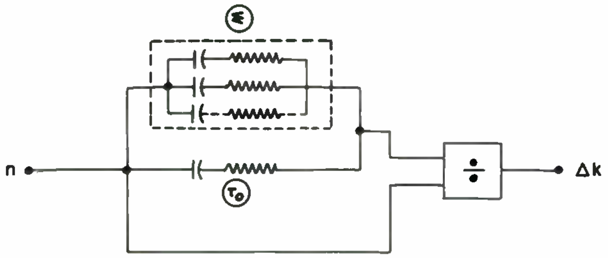


Fig. 4.

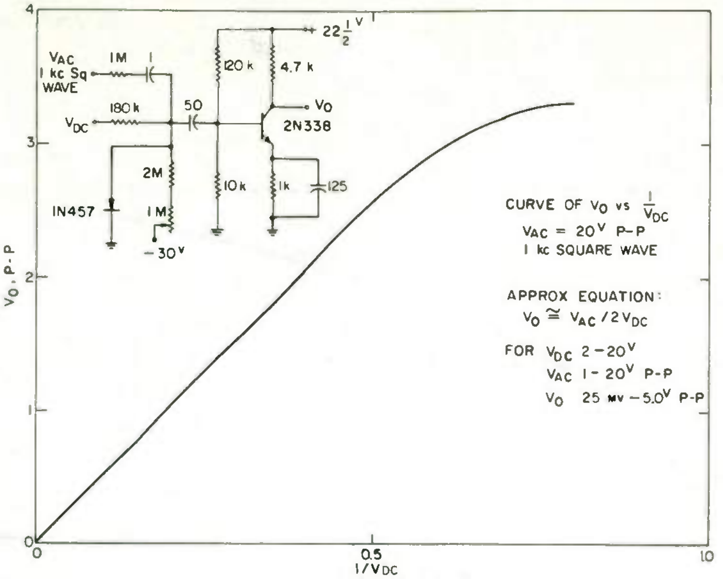
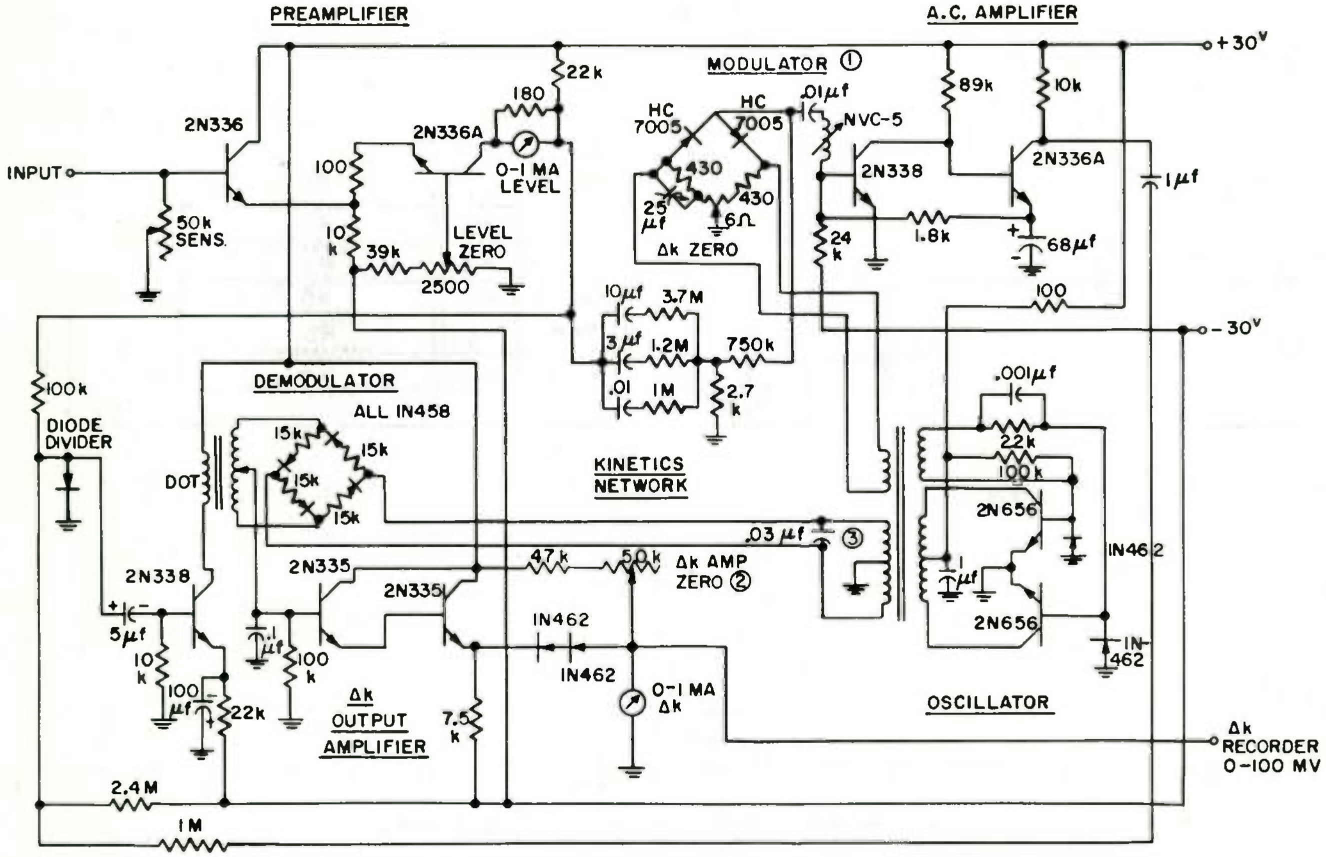


Fig. 5.



NOTES: ① CAPACITOR ACROSS BRIDGE ARM TO IMPROVE BALANCE. ONE OF 430Ω RESISTORS SHOULD BE VARIABLE TO CENTER ZERO POSITION ON 6Ω POT.

② ADJUST FOR ZERO OUTPUT FOR ZERO OUTPUT FROM DEMODULATOR.

③ CAPACITOR SELECTED TO PROVIDE CORRECT PHASE SHIFT.

④ METER READS FULL SCALE FOR ZERO INPUT CURRENT.

Fig. 6.

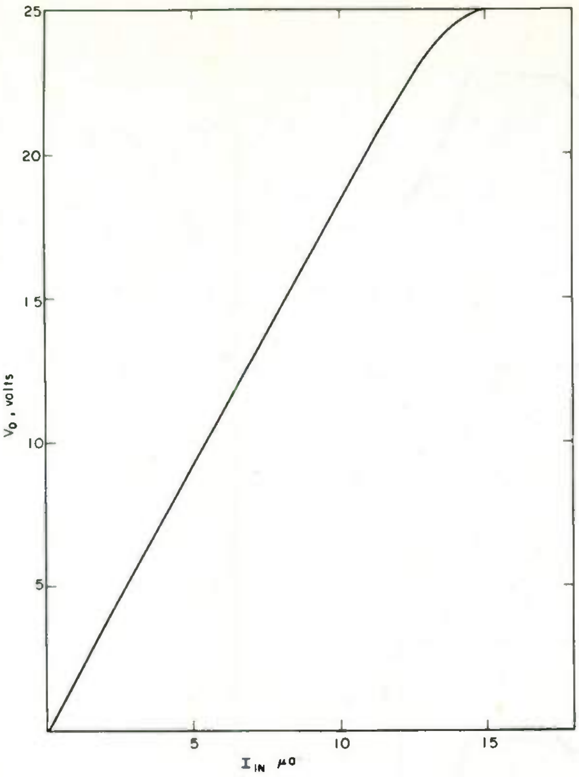


Fig. 7.

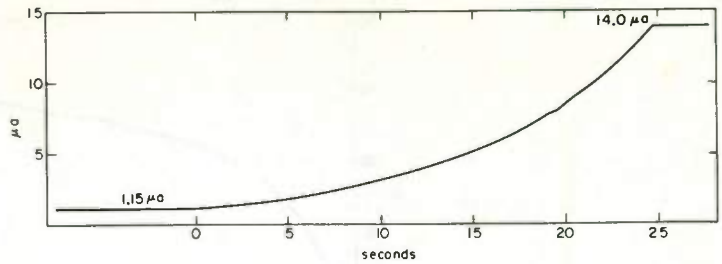


Fig. 9.

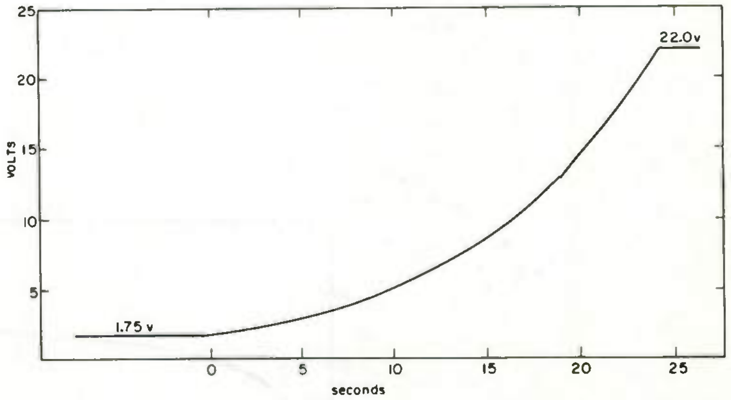


Fig. 10.

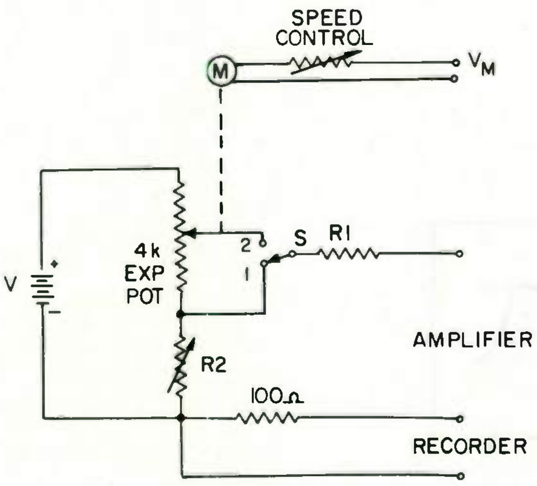


Fig. 8.

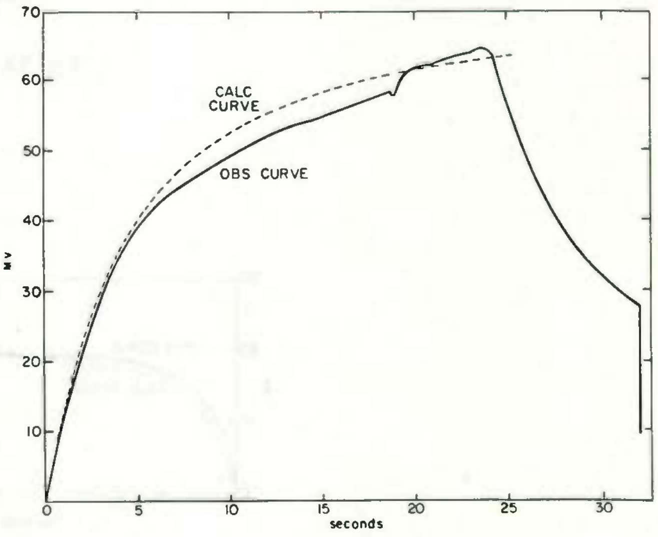


Fig. 11.

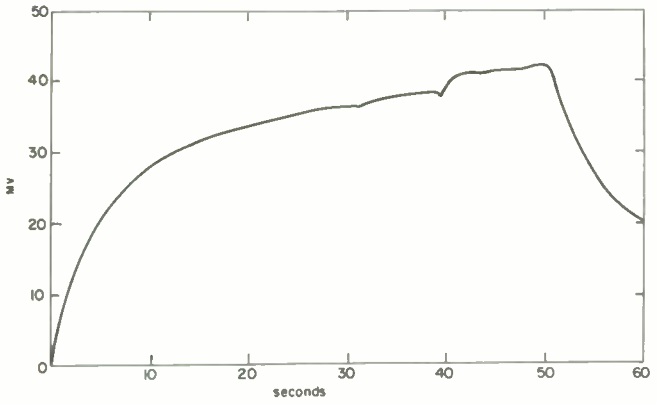


Fig. 12.

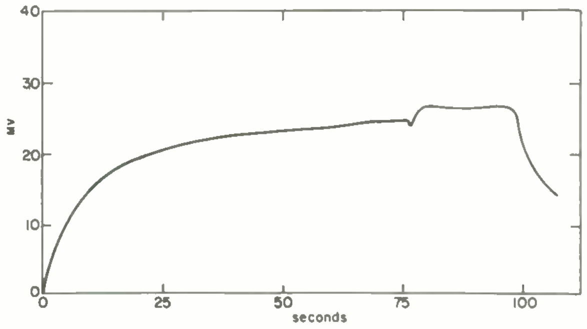


Fig. 13

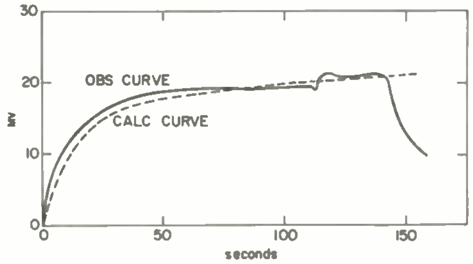


Fig. 14.

POSITRON SCANNER FOR LOCATING BRAIN TUMORS*

S. Rankowitz, J. S. Robertson, W. A. Higinbotham and
M. J. Rosenblum
Brookhaven National Laboratory
Upton, N.Y.

Summary

A system will be described which makes use of positron emitting isotopes for locating brain tumors based on the method developed by Sweet, Brownell and Aranow.[1,2,3] This system inherently provides more information about the distribution of radioactivity in the head in less time than existing scanners which use one or two detectors. A stationary circular array of 32 scintillation detectors scans a horizontal layer of the head from many directions simultaneously. The data, consisting of the number of counts in all possible coincidence pairs, is coded and stored in the memory of a Two-Dimensional Pulse-Height Analyzer.[4] A unique method of displaying and interpreting the data will be described which enables rapid approximate analysis of complex source distribution patterns.

Introduction

For a number of compounds, the rate of uptake by brain tissue is slower than the rate of uptake by muscle tissue or by certain types of tumor tissue. Thus, after intravenous injection of labeled compounds, the radioactivity may be higher temporarily in brain tumors than in surrounding healthy brain tissue. This phenomenon has been used to locate tumors in the head. The technique is difficult because the tracer level in the blood stream and in muscle tissue is comparable to that in the tumors; the activity ratios are small and the effect is transitory. The scanning has heretofore been done with one or two detectors so that it has taken a long time to accumulate data even when relatively large amounts of tracer have been injected.

Sweet, Brownell and Aranow[1,2,3] have used positron emitters as the radioactive tracers. An emitted positron gives rise, through an annihilation reaction, to two 0.51 Mev gamma rays which are emitted in opposite directions. When detectors are placed on opposing sides of the source, a coincidence indicates activity within the cylindrical space connecting the detectors. This has several advantages compared with the use of a single gamma collimated detector, e.g. improved resolution and insensitivity to background single gamma radiation without the necessity of large, heavy shielding and collimating structures. The current work is an extension of this technique, making use of a multiplicity of detectors. This report covers four aspects of this development: 1) factors affecting arrangement of detectors, 2) electronic circuits for counting coincidences, 3) relating output data to activity distribution in the source, 4) preliminary results. The instrument is built and has been tested with artificially produced radiation patterns. It has yet to be tried on human patients, and the best method of analyzing and presenting the data has not been determined.

Arrangement of Detectors

Positron coincidence scanning with two detectors is illustrated in Fig. 1. The detector pair is scanned slowly up and down and forward and back registering the coincidence counts produced by positron annihilation gamma rays on a projection map of the head. A serious disadvantage of this kind of projection representation, also used with the single gamma scanning technique, is the distortion of the recorded radioactivity pattern which can result. In cases of

* This work was done under the auspices of the U. S. Atomic Energy Commission.

multiple tumors, the projections can be misleading even when two projections from different angles are taken.

Additional detectors are desirable to provide more information about the radioactivity pattern with greater speed, i.e. by simultaneously scanning from many directions. One possible configuration is shown in Fig. 2 with the detectors arranged in two planes on each side of the head. The number of coincidences equals $\left(\frac{n}{2}\right)^2$, where n is the total number of detectors. This arrangement is more sensitive to counts at the center of the pattern than at the top and bottom, and some of the detectors are farther from the center than others. It was finally decided to arrange the detectors in a circle as shown in Fig. 3. The system has 32 detectors with 1 1/4 inch diameter by 1 inch deep NaI crystals arranged in a 15 inch diameter circle. Figure 3 shows the sensitive areas which intersect the head, for the coincidences between detector No. 1 and the opposing detectors. The head is, therefore, "scanned" simultaneously from each detector resulting in 9 to 11 paths through the head producing coincidences for each detector. An average head will produce about 170 to 180 coincidence combinations. This ring will take data in one plane and will be moved in 10 steps along the vertical axis of the head to obtain three-dimensional data.

Coincidence Pair Counting System

In addition to the detector counting assembly the system includes: 1) coincidence circuitry which counts coincident pulses between any pair of detectors and codes the data in a form suitable for storage in 2) a Two-Dimensional Pulse Height Analyzer memory.[4]

The detectors are numbered from 1 to 32 sequentially around the circumference. The coincidence counting system, using standard Computer Control Co. "S-Pacs"[5] and compatible laboratory-built circuitry, examines each event, rejecting single counts and multiple coincident counts of three or greater, and finally codes the two detectors in the pair for memory storage. The lower number detector is coded in a train of pulses equal to its assigned number and stored in the X memory of the Two-Dimensional Analyzer. The higher number detector is simultaneously coded into another pulse train equal to its assigned number and stored in the Y memory.

Referring to the system block diagram (Fig. 4), each of the 32 detector amplifier outputs drives one input of a 2 input NAND circuit which in turn sets one stage of the 32 stage shift register. The other input is a gate common to all 32 input NANDS which can prevent setting of the shift register during analysis of an event.

When the input gate is open, any single or multiple coincident event$^{(A)}$ can set the appropriate shift register stages in parallel. The Clock Gate Flip-Flop$^{(B)}$ is set by the switching of any shift register stage through the 32 fold "OR" circuit. The input NAND gate closes immediately$^{(R)}$ preventing acceptance of any more detector pulses until completion of the analysis. The Clock Gate Flip-Flop triggers Pulse Shaper$^{(C)}$ which inhibits for 10μsec the start of the gate$^{(F)}$ to the Clock Multivibrator$^{(G)}$. During this delay period, the states of all 32 stages of shift register are examined in parallel to determine whether the event was a single or a multiple coincidence. Since only the coincidence pair of 0.51 Mev annihilation gammas is of interest in this scan, the analysis of singles and all other multiple coincidences are prevented. The three standardized outputs from the coincidence analyzer, indicating the number of switched shift register stages, are converted in the logic NAND-NOR circuitry to a standard one or zero level voltage$^{(D)}$; logic zero (0 volts) represents a coincidence pair, and logic one (-6 volts) represents a single or triple or higher order coincidence. This voltage is gated in a 2 input NAND by a standard 0.6 μsec pulse(E) at the end of the 10 μsec clock gate inhibit pulse. A coincidence pair will result in a closed gate. Any other condition opens the gate and permits the pulse(E) to trigger a delay multivibrator which generates a 10 μsec pulse(L) to reset the shift register, all flip-flops, and open the input gate.

The coding of the 2 detectors involved in a coincident pair proceeds by gating "ON" the free-running Clock Multivibrator after the 10 μsec logic delay period. Shift pulses(G) are generated at about 600 kc. These shift pulses appear simultaneously at the X and Y outputs until the X Gate Scale-of-two is set(K). A 0.6 μsec delay in triggering this scaler permits the last X pulse to appear at M. The number of shift pulses required to change the state of the No. 1 shift register equals the lower detector number involved in the coincidence pair. The No. 1 shift register stage(H) gates through a clock pulse(I) to trigger the X Gate Scale of 2. The shift pulses appear simultaneously at the X and Y outputs until the X Gate scaler is set, inhibiting the X output. The shift pulses continue until the No. 1 shift register stage again gates through a second clock pulse to trigger the X Gate Scaler, resetting it. Returning the scaler to its reset state resets the Clock Gate Flip-Flop and stops the clock. Therefore, the number of pulses in the Y train corresponds to the second detector number involved in the coincidence pair. Resetting the Clock Gate Flip-Flop triggers pulse shaping circuits; one(L) resets the shift register, and a second one generates a "STORE" pulse(P) to the Two-Dimensional Pulse-Height Analyzer Memory.

The "STORE" pulse transfers the X and Y register counts in the Pulse-Height Analyzer into the proper location of its magnetic drum memory. During this memory storage time, the Scanner input NAND's remain blocked(R). As a safety measure, to prevent opening the input gates for the few microseconds between resetting the Clock Gate Flip-Flop and starting the "Dead Time" pulse, the 10 μsec shift register reset pulse is applied to NOR(R). Upon completion of the "Dead Time" pulse the Scanner is able to accept a new input from the detectors.

Data Presentation and Analysis

The stored data in the Two-Dimensional Analyzer consisting of the number of coincident counts recorded in each pair of detectors, sorted into X (horizontal) and Y (vertical) coordinates, can be displayed on the analyzer oscilloscope as an intensity modulated two-dimensional map or as count magnitude curves of the X channels for any selected Y. The curves of any selected group of Y channels may be simultaneously displayed vertically displaced.

The most useful presentation for analyzing the data and locating the tumor is the X-Y map. The curves are used to determine amplitudes and channel groupings more accurately in doubtful regions, when necessary. The display map intensity modulation can be varied with the scale factor switch in binary steps to permit setting the threshold of visibility at different levels.

Although actual tumors behave as extended sources of radiation in an absorbing medium containing varying amounts of lower level radiation, the technique of locating tumors may be illustrated by first observing the coincidence data patterns produced by a point source. Refer to Fig. 5.

1. Point source in center: It is obvious that a point source in the center of the detector circle will produce an equal number of coincident counts in all opposite detector pairs. Since the opposite pairs are 16 detectors apart, the coincidence pattern will be a straight line of constant difference, Y-X = 16, plotted on Fig. 6 as curve 1.

2. Point source 1.5 inches from center on diameter 9-25: If all coincidences are plotted on X-Y coordinates, Curve 2 results. This curve intersects Curve 1, the line of constant differences at only one point, X = 9, Y = 25. Curve 3 results from the source at 3 inches from center on the same diameter. For both curves:

$$\text{if } X < 9,\ (Y-X) < 16$$
$$\text{and } X > 9,\ (Y-X) > 16$$

The diameter on which the source is located is determined therefore by locating the counter pair for which Y-X = 16. The distance of the source from the center of the circle may be determined by observing the coordinates X, Y at which (Y-X) differs most from 16. The location of a point source in the scanned area is then determined by the intersection of the radial line for (Y-X) = 16 and a line defined by any pair of coincidence detectors obtained from the coincidence curve. The coincidence curve as shown here appears on the display scope of the Two-Dimensional Analyzer.

The X,Y coincidence data provides a great deal of information about the location and size and shape of the tumor. A point source can clearly be accurately located. An extended tumor can be located by a similar analysis as long as the background level caused by radiation from surrounding brain tissue and blood is low enough to produce a contrast. A tumor of finite extent will produce a band on the display similar to that from a point source (Fig. 7) whose width is approximately proportional to the average diameter of the tumor. This band will intersect the reference curve Y-X = 16 over some region of coordinates. The center of this intersection will determine the angle of the tumor from the center of the head, and the boundaries can be resolved by observing the outer coordinates along the band and plotting these coincidence lines between the detectors.

The following figures (8a to e) are Polaroid photographs of the Two-Dimensional Analyzer map display for some artificially produced patterns of radioactivity. Each figure contains three photographs of the same display, but with count scales of 2^7, 2^8, 2^9. The positron emitting isotope used as the source of tumor and background radiation is Na^{22}.

Figure (8a) shows the map display for three small (1/4 inch diameter) sources located on the diameter between detectors 9 and 25 at the center and at distances of 3 inches on either side of center with no background radiation. These correspond to curves plotted in Fig. 6.

The straight line, Y-X = 1, which appears on all photographs for the low count scale results from coincidences between adjacent detectors caused by the high energy single gammas emitted by Na^{22} and the proximity of the detectors. This well defined straight line, together with the zero axes, outlines the useful data area of the display. These coincidences are not useful analytically since they tell nothing about the source distribution.

It should be noted that data for Y = 32 does not appear on these photographs because of the selection of display scales available.

Figure (8b) shows the map display for the same 1/4 inch diameter source in the center and a larger source (about 1 1/2 inches diameter) located 2 1/4 inches from center toward detector 9 on the same diameter as in Fig. (8a), with no background.

Figure (8c) shows the same conditions as Fig. (8b) but with a uniform background radiation throughout the head area.

Figure (8d) shows the same conditions as Fig. (8c), but with a third 1/4 inch diameter source (about 1/3 the intensity of the larger one), located on a diameter from 1 to 17 (front to back midline of head), at a distance of 3 inches from center toward detector 1. The curve for this source intersects the reference constant difference line at the XY coordinates 1, 17. Since it is less active than the larger source, it is not visible on the highest count scale.

Figure (8e) illustrates the kind of deductions that can be made about the shape as well as location of the tumor. Four Polaroid photographs of the coincidence data produced by an approximately elliptical tumor between the center and point 3 of Fig. 5 with scale ranges of 2^7 through 2^{10} are shown in a uniform background. One edge of the band of X-Y data coincides with the straight line Y-X = 16, indicating on which side of center the source is located. The band crosses the reference line at (X, Y = 9, 25) indicating the major axis. Since the width of the band is definitely smallest in the region of (X, Y = 9, 25), the source dimension perpendicular to this diameter is smaller than the dimension along the diameter. The width of the X-Y data band at the extremes indicates the approximate length of the source along the diameter (9, 25); the number of coincident pairs in the region of (9, 25) indicates the dimension perpendicular to this diameter.

Thus, by examining the characteristics of the coincidence data on the intensity modulated X-Y coordinate map, a great deal of information about the location, size and shape of the source of radioactivity can be readily obtained despite a high level of background.

A study program is being conducted to determine whether a mathematical solution of the problem of converting the coincidence data to a map of the radioactivity distribution with reasonable errors is feasible. One such method treats each of the N areas within the head (where N equals number of coincidence combinations caused by the radioactivity within the head) as though it were the component of an N dimensional vector. The mapping from the space of such vectors on to the space of coincidence counts is linear and consequently can be represented by a N x N matrix. If this matrix has a stable inverse it is then possible to convert the coincidence data unambiguously to a map of the radioactivity distribution using a simple computer program. The inverse operation is degraded by statistical and system variations in the coincidence counts. It is too soon to tell whether this mathematical approach will prove to be practical. It is also possible that other modes of display may be advantageous.

References

1. G. L. Brownell and W. H. Sweet, "Localization of Brain Tumors with Positron Emitters," Nucleonics, Vol. 11, No. 11, pp. 40-45, (1953)

2. W. H. Sweet and G. L. Brownell, "Localization of Intracranial Lesions by Scanning with Positron-Emitting Arsenic," J.A.M.A. 157, 1183 (1955)

3. S. Aranow and G. L. Brownell, "An Apparatus for Brain Tumor Localization Using Positron-Emitting Isotopes," IRE Convention Record, Vol. 4, Part 9:8, (1956)

4. R. L. Chase, "A Two-Dimensional Kicksorter with Magnetic Drum Storage," IRE Convention Record, Part 9, p. 196, (1959)

5. Computer Control Co., Framingham, Mass.

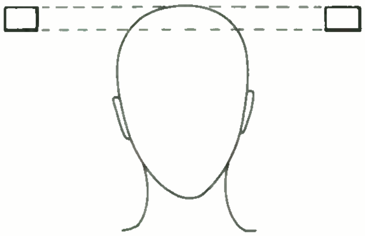


Fig. 1. Positron coincidence scanning with two detectors.

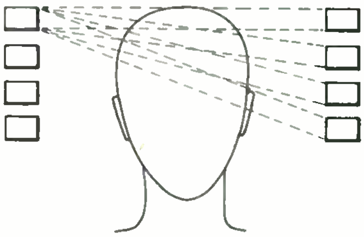


Fig. 2. Positron coincidence scanning with detectors arranged in two planes on each side of the head.

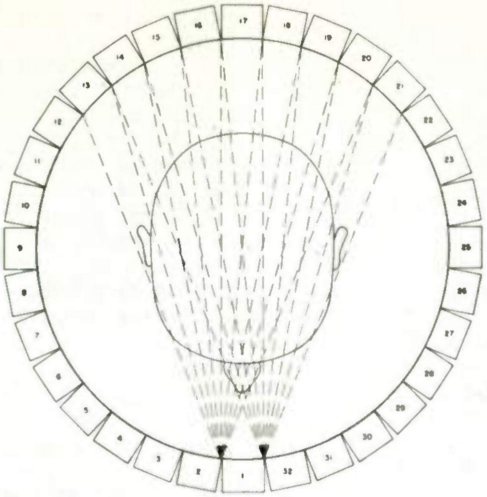


Fig. 3. Positron coindence scanning with 32 detectors arranged in a circle.

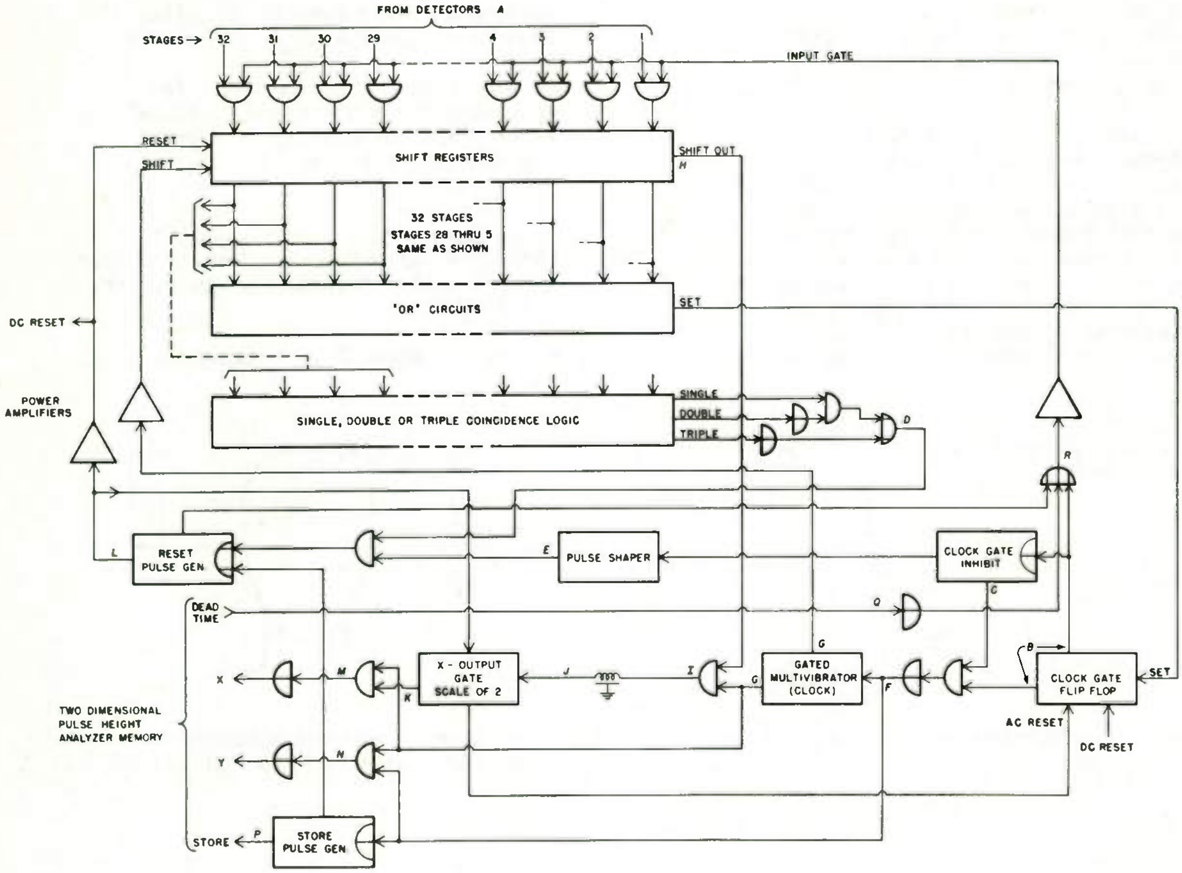


Fig. 4. Block diagram of coincidence detector pair coding system.

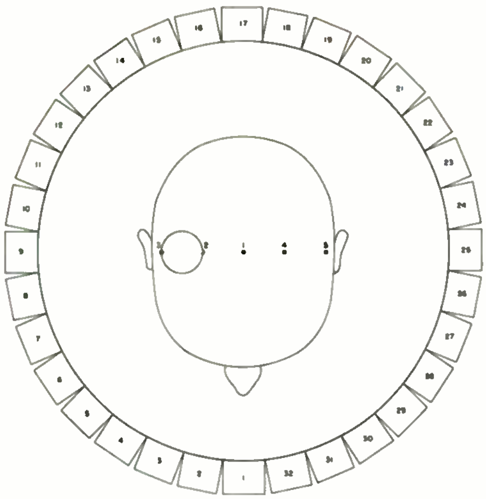


Fig. 5. Positrons of test sources for Figs. 6 and 7.

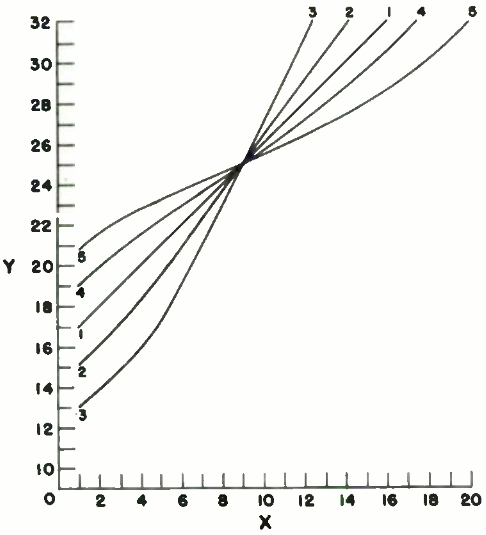


Fig. 6. X-Y coincidence patterns derived from point sources at positions 1 through 5 on Fig. 5.

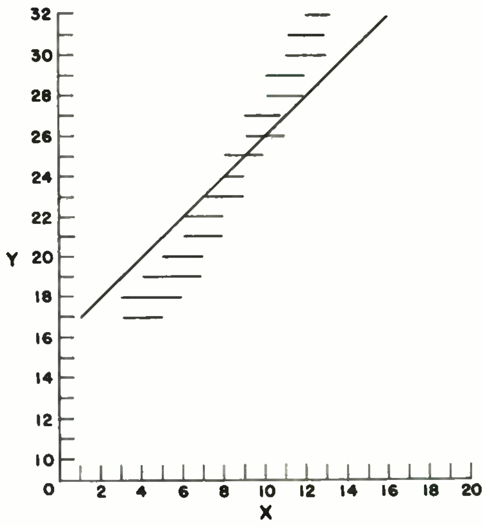


Fig. 7. X-Y coincidence pattern for point source at center (1) and for 1.5 inch diameter source located between points 2 and 3 on Fig. 5.

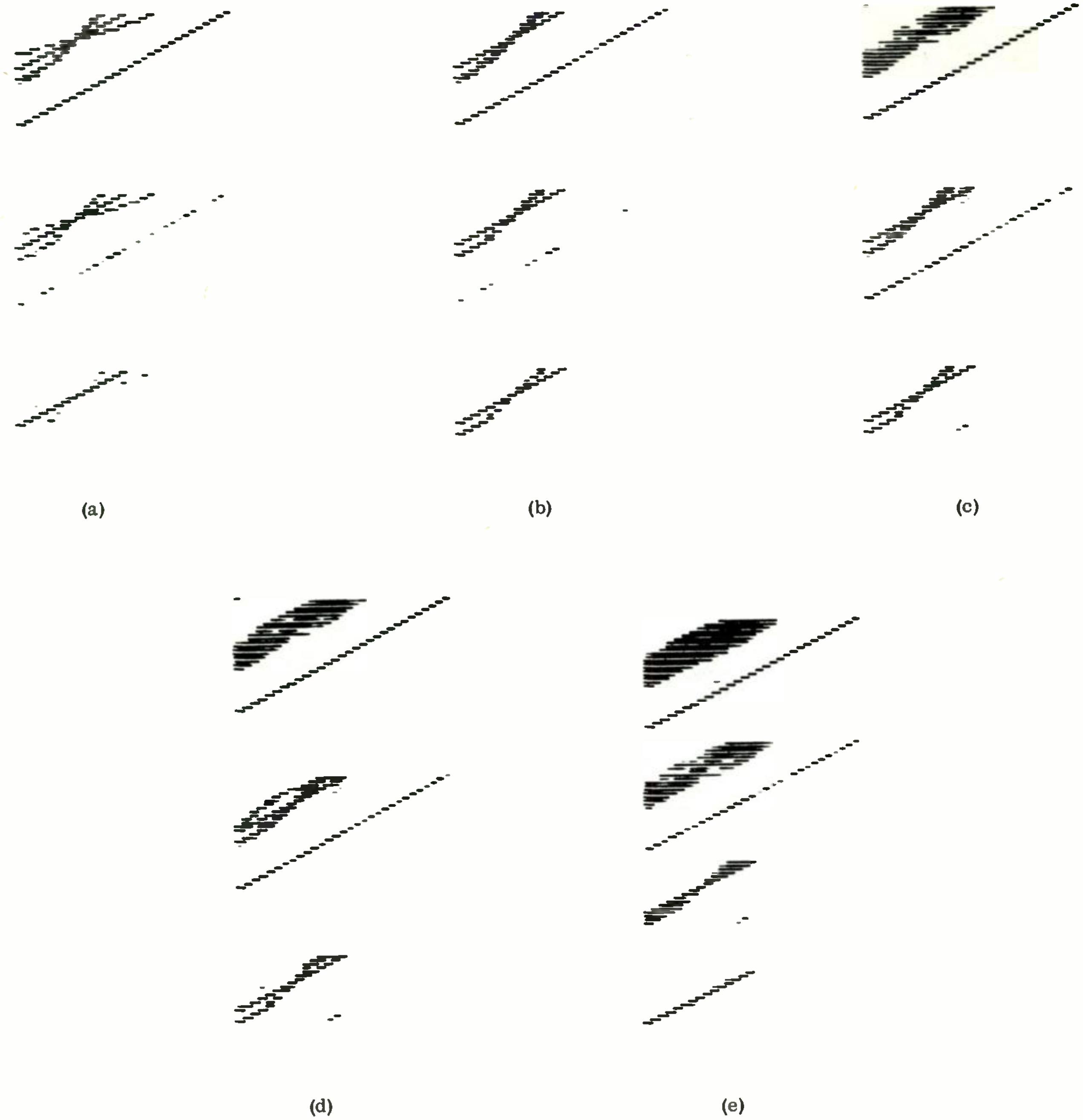


Fig. 8. Photographs of Two-Dimensional Analyzer map display. Count scales (top to bottom) are 2^7, 2^8, 2^9. Point source used is 1/4 inch diameter. (a) Three point sources located on diameter between detectors 9 and 25, at points 1, 3 and 5 of Fig. 5. (b) Point source at 1 and larger source between Points 2 and 3 of Fig. 5. (c) Same as (b) but with uniform head background. (d) Same as (c) but with third source (point) located three inches toward front of head on front to back centerline. (e) Point source at 1, and approximately elliptical shaped uniform source in field of uniform head background; long axis of source is on 9, 25 diameter. Fourth photo taken with 2^{10} counts full scale.

PANEL: AEROSPACE NUCLEAR PROPULSION AND POWER

Chairmen: R. E. Finnigan and P. M. Uthe
Stanford Res. Inst.
Menlo Park, Calif.

Abstract

This session reported recent advances in the instrumentation, control systems and engine dynamics aspects of the aerospace nuclear propulsion and nuclear auxiliary power fields. In addition, it discussed on nuclear radiation effects to controls and instruments. It reported technical progress on portions of the following specific nuclear programs:
ROVER (Nuclear Rocket)
PLUTO (Nuclear Ramjet)
SNAP (Systems for Nuclear Auxiliary Power).

BLACK AND WHITE, OR GREY?

Gerhardt Dorn
Applied Bionics Laboratory - Advanced Development
Westinghouse Electric Corporation
Baltimore 3, Maryland

Summary

Areas of activity in bioelectronics reported in the IRE Professional Group Transactions on Bio-Medical Electronics and in the Digest of the 14th (1961) Annual Conference on Electrical Techniques in Medicine and Biology are examined.

Problems in artificial intelligence and neural information processing models are selected as typical of this field and used as vehicles for discussing the present status of the art, in which current developments have not matched earlier expectations.

Oversimplification of models, caused by the difficulty of reconciling differing frames of reference characteristic of the interdisciplinary approach in problem solving, is suggested as the major factor contributing to this unexpected lack of progress.

I. Publication Survey of Bio-Medical Electronics Group

As the Institute of Radio Engineers Professional Group on Bio-Medical Electronics nears the end of its first decade, it is appropriate to survey the extent and variety of the articles published by this group during the period 1953 through 1961.

Figure 1 shows the number of papers published in the Transactions of the Professional Group on Bio-Medical Electronics by years, beginning with the first issue, Volume MR-1, November, 1953, and ending with Volume BME-8, Number 4, October, 1961. This represents a total of 273 papers published in the official organ of this group. It should be mentioned that six papers on medical electronics were published in the 1952 IRE National Convention Record, a year before the group became officially chartered. Note the rapid increase in the number of papers published, peaking in 1959 and then declining somewhat.

This decrease is understandable in light of the editorial comment appearing in the July, 1961, Transactions on Bio-Medical Electronics [1] and reading in part as follows: "Considering the number of papers presented at the [13th Annual] Conference [on Electrical Techniques in Medicine and Biology] , it is distressing to note that a relatively small percentage were recommended . . . for publication . . . More specifically, if one examines the majority of the papers presented during the Conference, it will be noted that with the exceptions they follow well-established lines. In a field such as bio-medical engineering, this is an alarming situation."

Table 1 summarizes the articles published by the group in terms of percentages of effort for areas in a broad classification of 11 fields of activity. These percentages are compared with a similar distribution of about 300 papers presented during the 14th (1961) Annual Conference on Electrical Techniques in Medicine and Biology.

Table 1. Percentage of Papers by Subject Areas

SUBJECT AREA	TRANSACTIONS OF IRE PGBME (%)	14th ANNUAL CONFERENCE (%)
Circulation	18	8
Biological Data, Instrumentation, Recording, Telemetry	16	30
Heart	16	15
Computers, Diagnosis, Models, Data Processing	14	9
Miscellaneous	14	6
Radiology	10	12

Table 1. Percentage of Papers by Subject Areas (Continued)

SUBJECT AREA	TRANSACTIONS OF IRE PGBME (%)	14th ANNUAL CONFERENCE (%)
Biological Control Systems	0	8
Biological Potentials	5	4
Artificial Organs	5	0
Gastroenterology	0	5
Nerves	2	3

II. Assessment of Current State of Development of Bioelectronics

In addition to surveying the extent and variety of publications in this field, it is also appropriate to try to evaluate the present state of development of bioelectronics.

It is not possible here to consider all of even the most important fields of interest. Consequently, artificial intelligence and neural information processing models are selected as typical of the more challenging and difficult problems of current bioelectronics activity and used as vehicles for a preliminary assessing of the current state of biomedical electronics.

Artificial Intelligence

The basic problem in artificial intelligence studies is concerned with the question of properties to be distinguished and manipulated by a machine.

It is considered necessary to classify problems and situations into manageable and useful categories.[2] To each situation is associated the name of its category. This name is called its "character," and its assignment to one of these categories is called its "characterization."

A frequently-encountered approach to the treatment of properties includes "property-list" methods, in which an object is subjected to a sequence of tests, each one of which detects some property considered to be important for heuristic reasons. These properties are required to be invariant under ordinary distortions or transformations. Related problems which emerge at this point are the invention of new and useful properties and the combination of many properties to form an identification system.

A property or attribute is defined as a two-valued function dividing objects to be identified into two classes or categories. An object does or does not have the property in question according to whether the value of the function is 1 or 0. Given n properties to be used for distinguishing, 2^n subclasses can be defined by the intersections of these properties. By combining these properties with AND's and OR's, 2^{2^n} patterns are obtained. For example, for the three properties rectilinear, connected and cyclic, there are eight subclasses (2^3) defined by their intersections and 256 patterns (2^8). For this sequence of properties (rectilinear, connected, and cyclic) a square can be characterized by the vector (1, 1, 1) and a circle by the vector (0, 1, 1).

For a variety of problems, characters such as these can be used not only as names for categories, but also as elements for defining patterns.[3]

The foregoing train of thought reflects substantial progress indeed in breeching the difficulties inherent in those basic problems of artificial intelligence, properties and categories.

Nevertheless, it is necessary to continue consideration of the concepts of property (or attribute) and category in a more generalized and penetrating fashion.

The restriction of a property to a two-valued function results in classification which reflects the Aristotelian distinction between essential (defining) and accidental (non-defining) attributes. This kind of logic is not very well suited to categorizing behavior. More suitable than this over-simplified dichotomy is a continuum of relative "criteriality" of an attribute or property for a class or category. When some value of a property is used as a basis for inferring the class membership of an object or event, then that property is to some degree criterial for the classification. On the other hand, to the degree that a property can change in value without affecting classification judgments, it is not criterial for the categories in question.[4]

It is useful to distinguish between "actual criteriality" and "potential criteriality." Actual criteriality describes the use of a property in some individual's categorizing behavior. Potential criteriality is concerned entirely with existent conditions rather than the behavior of a particular person.

Non-criterial properties are not all of the same kind. A property that has more than one value for the members of one category, and which assumes these same values for another category, is non-defining and "noisy." Thus, the quality of paper is a "noisy" property when used to distinguish books from magazines.

When a property has the same value for all the members of several categories, it is not only non-defining for the categories in question but also "quiet." In distinguishing books from magazines, the quality of being bound on one side and open on the other is "quiet."

Consequently, noisy properties impede, while quiet properties aid, the formation of categories.

Criteriality of a property for a category is always relative to some other category or categories from which the first is to be distinguished.

Further, any set of events, objects, or situations is susceptible of many alternative categorizations. In brief, a category is a human construction imposed on some population or array of events, objects, or situations.

Principles of grouping generate different types of categories, including unitary, conjunctive, disjunctive, and relational. In a unitary category, all members have some single property which is found only within the category. The joint presence of the appropriate values of several properties is a conjunctive category. In a disjunctive category, the members have no properties in common. A relational category is one in which a definitive relationship (topological rather than metric) exists between properties.

One additional distinction is important in this connection. The properties of most categories can be considered to be either formal or functional. Formal properties include such traits as mass, color, shape, and the like, while functional properties refer to the possible utility of the members of a category.[5]

One comment is necessary at this point. Without categorization it is impossible to form expectancies, since there can be no recurrences under this constraint. In other words, an event, even in all of its discriminable detail, never repeats. In this case it is useless to remember that B followed A because A will never recur. However, with categorization, events do recur; it is then profitable to presume that A led to B because there can be other instances of A and B.

In concluding this part of the discussion, it must be mentioned that these considerations bear on that important alter ego of artificial intelligence, the simulation of behavior. Behavior is, of course, notoriously non-linear. This diffifulty is compounded by the fact that the machine does not simulate behavior as such, but only some record or measurement of the behavior. Consequently, the machine theorist is at the mercy of the person deciding what aspects and properties of behavior are worth recording, and both are constrained by their respective definitions of the concept of "property."[6]

The degree to which an organism and a machine can be interchanged without altering the specified aspects of the situation determines the degree of success of the simulation. That aspect of an organism's behavior which an experimenter chooses to record must remain invariant for the model under consideration to be successful.

Neural Models

Progress in the construction of meaningful models of neural information processing is clouded by a too ready acceptance of the "all-or-none" principle to explain the functioning of neurons. A common assumption holds that a neuron sends messages to other neurons or to effectors in the form of pulses (action potential spikes). In addition it is assumed as a first approximation that the pulses of a given neuron are all alike. Thus, as far as information transmission by the neuron is concerned, only the time intervals between pulses differentiate neuron behavior and so carry information. However, if the neuron is considered to be an information relaying device (one whose differentiated states contribute to the structuring of events at the next higher echelon in the system), it is possible that aspects other than these pulses can carry information.

Since the actual coding systems of neural elements are not yet known, one is forced to examine the time courses of physiological states of neurons, even though the treatment of these time courses is purely descriptive. Physiological literature reflects the vast effort which physiologists have devoted to minute descriptions of how neurons respond to a great variety of stimuli, how responsive states are distributed over neurons, what their time courses are, the process of recovery, and the like.[7]

Current neural doctrine holds that the neuron comprises components of quite differing functions, including those of processing varieties of inputs and determining outputs.

The four principles discussed below are generally accepted today in neurophysiology. [8]

a. The neuron is a functional unit; as long as it has one output path or axon it "speaks with one voice." However, some neurons have two axons and can simultaneously deliver two nonidentical pulse-coded outputs. [9] In any case, the "all-or-none" pulse is characteristic solely of the axon. Other regions of the neuron are capable of graded responses.

b. The responses of many parts of neurons contribute to the initiation of the pulse response in a critical region at the base of the axon.

c. Although the axon-conducted pulse response transmits without decrement, responses of other parts of the neuron experience decrement during propogation. Because of this, parts of the neuron sufficiently far from the region where the pulse response is initiated may not contribute to this initiation. Some dendrites are so fine and long that it is doubtful whether they can directly influence pulse initiation. (See figures 2 and 3.) It is quite possible that much of the activity of dendrites consists of influencing other neurons. It also appears likely that brain waves are (1) the synchronized subthreshold dendritic potentials of many neurons and (2) causal agents rather than by-products.

d. Integrative and labile processes at the level of the single neuron are not confined to the synapse. [10] At least four such loci probably exist and are integrative in the sense of exercising labile evaluative functions on whatever comes to them.

Forms of activity in the neuron include the pulse and local potentials; these are considered to be responses to prior activity in the neuron itself. On the contrary, the generator potentials (sensory stimuli) of receptor neurons and the synaptic potentials (junctional transmitters) are responses of cells to impinging events external to the neuron. Also, there are spontaneous prepotentials which occur during steady state conditions. These types of potentials do not exhaust the possible states since there exist changes in state not inferable from potentials.

Figure 4 shows several presynaptic pathways converging from diffent sources: inhibiting (1), exciting other followers (2), exciting pacemakers (3), and accelerating (4). These produce synaptic potentials in their various special locations. Restricted regions also initiate spontaneous activity (pacemaker regions, shown schematically), local potentials (shown in only one place but possibly repeating elsewhere), and propagated impulses (arbitrarily located spike initiation). Only the axon supports "all-or-none" activity. Terminal ramifications presumably act by graded local potentials. Integration occurs at sites of confluence or transition from one event to the next. [7]

In summary, the pulse is only one of several forms of nerve activity. Excitation of one part of a neuron does not necessarily involve the entire neuron, nor is a particular region of the neuron the only site of a specific function.

This departure from the rigid digital point of view reflects a quiet but fundamental revolution in neural functional concepts.

This apparent wealth of complexity at the neuronal level may well be illusory. One has only to consider the possibility that glia cells may be intimately related to neuron functioning to realize that the complexity of brain activity as envisaged today is perhaps relatively simple. [11]

III. Conclusions

The problems of artificial intelligence and neural models discussed above are typical of bioelectronics in that they require interdisciplinary solutions. This approach, however, brings together not only the advantages of applying multidisciplines but also the disadvantages of reconciling the differing in individual frames of reference of these same disciplines. This presents a rather formidable difficulty when one considers the fact that in bioelectronic problems biological systems (which are generally anisotropic) are frequently joined with or transformed into physical systems (which are generally isotropic).

It is no surprise that a major consequence of this situation is the oversimplification of models. Even a cursory examination of the many proposals for modeling neural nets indicates a universal predilection for assumming that the "all-or-none" principle is the only mode of functioning of a neuron. Likewise, a too casual assumption of two-valued logic may lead to serious difficulties in the conceptual problems of artificial intelligence.

This general encroachment of oversimplification in the field of bioelectronics has produced a state of the art in which current developments have not matched earlier expectations.

This unexpected lack of progress can be corrected by appropriate university programs, perhaps primarily at the graduate level. Several such programs are now in progress and their impact should be felt within a few years.

In conclusion, a possible threat to the future growth and development of bioelectronics must be mentioned. This concerns the distressing (and, hopefully, waning) friction between disciplinary and vocational workers. This issue became overt during the conference (sponsored by the Professional Group on Bio-Medical Electronics and held in Omaha on October 26-27, 1961) studying the role of biomedical engineering in universities, and hospitals. This situation is far from being merely academic - one can only hope that the protagonists in question will resolve this obstacle to progress.

References

1. IRE Trans. on Bio-Medical Electronics, vol. BME-8, July, 1961.
2. M. L. Minsky, "Heuristic Aspects of the Artificial Intelligence Problem," Lincoln Laboratory, MIT. Group Report 34-55, 17 December 1956.
3. M. L. Minsky, "Steps Toward Artificial Intelligence," Proc. IRE, vol. 49, pp. 8-30, January, 1961.
4. J. S. Bruner, J. J. Goodnow, and G. A. Austin, "A Study of Thinking," John Wiley and Sons, New York, 1956.
5. R. Brown, "Words and Things," The Free Press, Glencoe, Illinois, 1959.
6. G. A. Miller, E. Galanter, and K. H. Pribram, "Plans and the Structure of Behavior," Henry Holt and Co., New York, 1960.
7. A. Rapoport and W. J. Horvath, "Information Processing in Neurones and Small Nets," WADD Technical Report 60-652, December, 1960 (AD 252897).
8. T. H. Bullock, "Neuron doctrine and electrophysiology," Science, vol. 129, pp. 997-1002, 1959.
9. T. H. Bullock and C. A. Terzuolo, "Diverse forms of activity in the somata of spontaneous and integrating Ganglion cells," J. Physiol., vol. 138, pp. 341-364, 1957.
10. T. H. Bullock, "Parameters of integrative action of the nervous system at the neuronal level," Exp. Cell. Res., Suppl. 5, pp. 323-337, 1958.
11. R. Galambos, "A glia-neural theory of brain function," Proc. N. A. S., vol. 47, pp. 129-136, 1961.

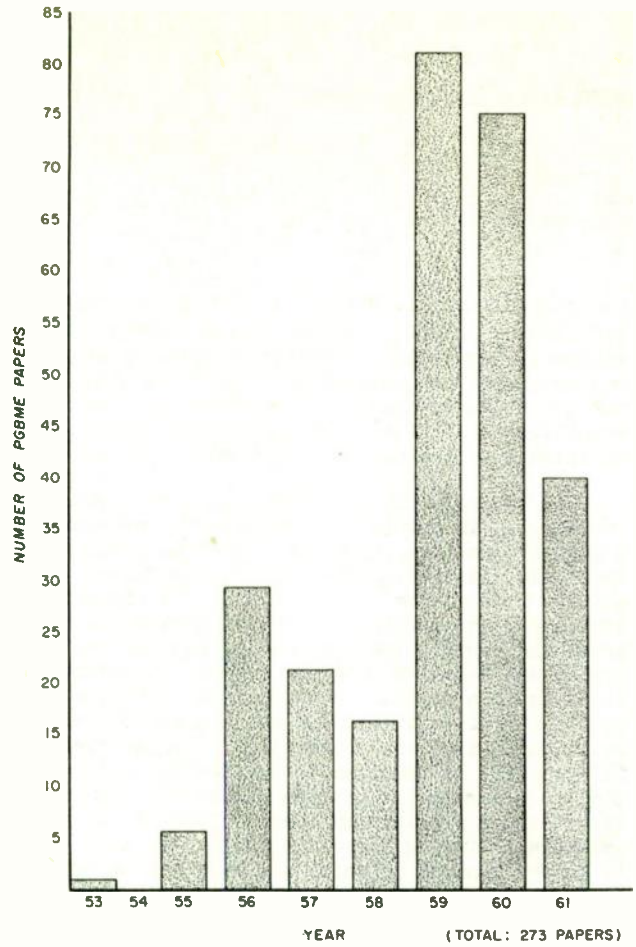


Fig. 1. Number of PGBME papers by year.

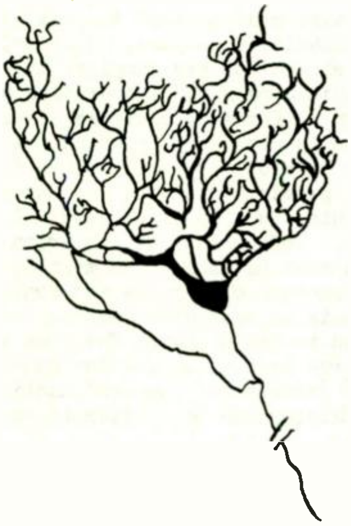


Fig. 2. Purkinje cell of the cerebellar cortex of man (after Cajal).

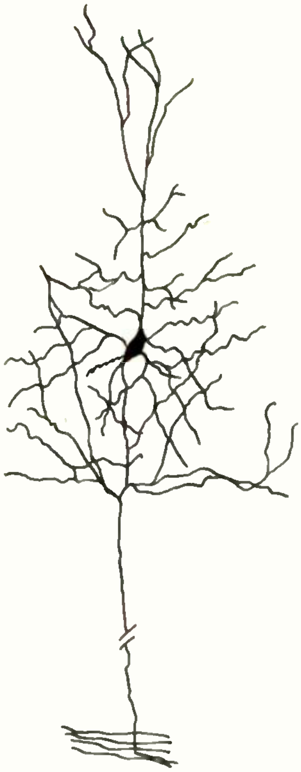


Fig. 3. Pyramidal cell of the cerebral cortex of man (after Cajal).

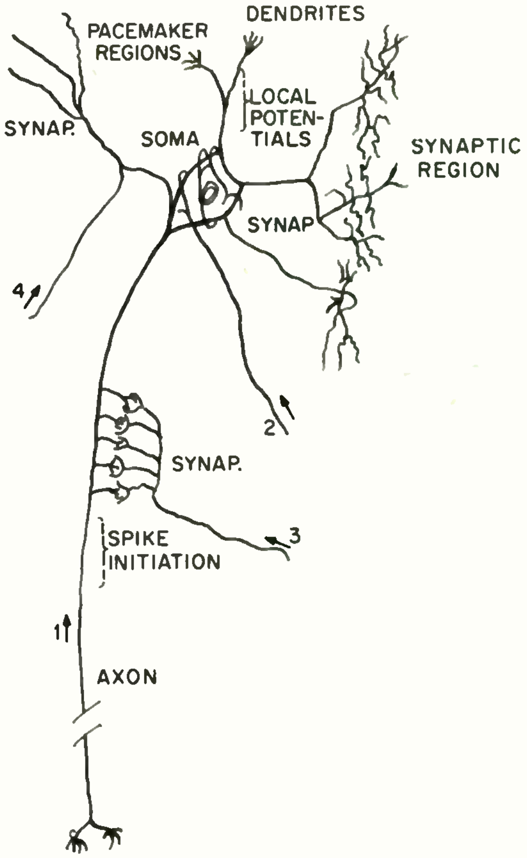


Fig. 4. Schematic representation of a neuron from the cardiac ganglion of a crab (after Rapoport and Horvath).

REPRESENTATION OF COMPLEX ELECTRONIC LEARNING SYSTEMS

E. B. Carne
Advanced Computer Laboratory
Melpar, Inc.

Summary

A learning system includes the learning network and the teaching circuitry. The teaching circuit must educate the learning network to perform the desired function by observing the output of the learning network and correcting its behavior by the application of reward and punishment. In simple situations the teacher will comprise a few elementary rules which can be applied immediately. In more complex situations, criteria must be learned which can then be used to control performance. System representations for this type of situation are given and an attempt is made to define degrees of learning and requirements for situations in which priority must be assigned to certain events. Prediction processes are also discussed.

Introduction

Essentially, a learning system comprises four sections: sensors, learning network, goal circuit and effectors as shown in Figure 1. The sensors observe the local environment and provide data concerning changes and conditions to the learning network. The learning network responds to this input and develops an output which actuates the effectors. This action is evaluated by the goal circuit which comprises some statement of the system objective and a means of evaluating the system response on the basis of the prescribed objective. A reward or punishment signal is applied to the learning network depending on whether the action is deemed to be good or bad. In the untrained state (i.e. initially), the probability of the correct action being developed by the learning network is extremely small. Progressive trial and error affords the opportunity for the goal circuit to influence the system by reward or punishment and to change the statistics of the learning network so as to improve its performance and to reinforce the states of those elements which contribute to the correct (desired) behavior.

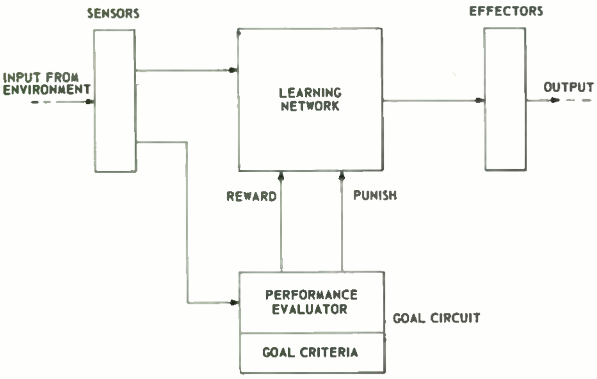


Fig. 1 *Block Diagram of Self-Organizing System*

It is essential to realize that the learning network is completely organized by external signals to perform any given function within the intrinsic capacity of the network. Problem solving is a matter of determining the correct states for the statistical learning elements by trial and error rather than being provided with a complete detailed program or list of instructions as in a digital computer. The relative simplicity of this training regimen and of its realization implies however that a longer time (in general) is required to develop the solution to a particular problem and that in many situations a preliminary training period is essential. During this time the system must be in an environment in which mistakes can be tolerated. This may imply some period of comparative isolation during which the machine is trained with certain classes of problems.

Learned Criteria

The tremendous advantage of a simple training procedure can be seriously compromised if the initial statement of the goal criteria becomes exceedingly complex. However, since most real situations are complex, it is difficult to achieve meaningful operation without complicated performance criteria. Accordingly, a mechanism is necessary by which the learning system can develop its own criteria for coping with these problems starting with much simpler models. This has a distinct parallel in human experience. Before a child can write intelligently, he must be taught the meaning of words and the structure of phrases and sentences. Prior to this he must have learned to recognize letters and to write them. The more elementary actions serve as a basis for the construction of a whole universe of complex functions. This type of training can be applied to machine intelligence.

Consider a maze runner in a simple maze as shown in Figure 2. Suppose the runner is given the primary instruction (primary goal) to keep moving. Further, suppose that the runner has a switch on its nose which indicates the presence of a wall or an obstacle. In addition to this elementary sensor, the runner has a number of more complex devices mounted on it which respond to different stimuli. One of these could be a

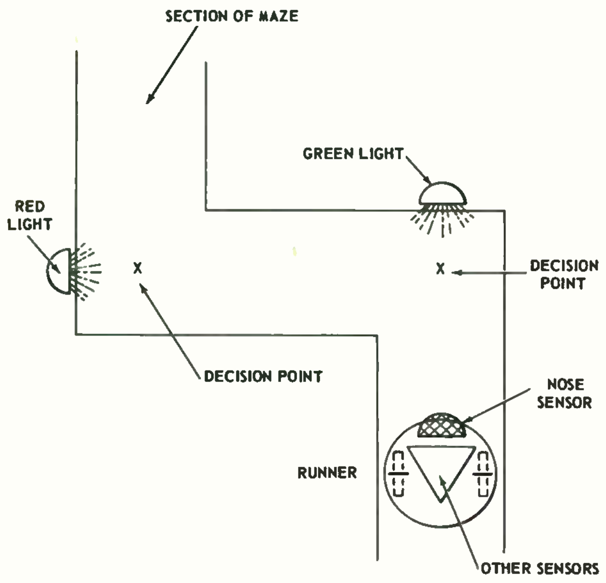


Figure 2. Diagram Illustrating Hypothetical Maze Problem

pair of photo-cells which respond to red and green lights. To implement the primary goal, the runner must make a decision whenever it bumps its nose against a wall. Initially it will turn right or left in a disorganized manner, depending on the initial settings of the statistical switches in the control net. For the sake of simplicity, assume that the maze is constructed without traps or re-entrant areas and that the geometry is such that if a wrong turn is made, the runner immediately contacts the adjacent wall and a second pulse is generated from the nose sensor. Fixed criteria can be implemented to recognize the presence of this second pulse as a wrong decision which will generate a punishment signal to the learning network. Similarly, the absence of the second pulse will be recognized as a correct decision and will occasion a reward signal. A system to implement this situation is shown in the upper half of Figure 3.

Suppose now that red and green lights are added at the corners of the maze so that red is always associated with a point at which the runner should turn right, and green is always associated with a point at which the runner should turn left. A subsidiary learning network can be added to the system as shown in the lower half of Figure 3, having inputs from the red and green sensors on the runner and the left or right drive instructions from the primary learning network. The output of this subsidiary network can then be trained, by comparison with the output of the fixed circuits, to give reward and punishment signals which will allow the maze runner to run the maze on the basis of the red and green signals only. Decisions can then be implemented by using the learned criteria in the subsidiary learning network rather than the fixed criteria originally programmed; the fixed criteria box could be removed from the net. Before removing the fixed criteria box, it is essential that the learned criteria be completely learned, and that the learning elements in the subsidiary learning network are fixed, i.e. no longer statistical. The runner may now be lifted from the elementary maze, and it will run other more complicated courses on the basis of red and green lights only. The fixed criteria box acts in the capacity of a "teacher" which can be discarded once the lesson has been learned. Of course, the recognition of red and green lights is not a great deal more complex than the use of a nose sensor, and has been used for the purposes of a simple illustration. The argument remains good if the direction information is contained in writing scrawled on the wall, or in the direction of an arrow. In these circumstances, the inputs to the learned criteria box can be very complex and the proposed method for organizing the appropriate criteria would be a very real advantage.

It cannot be overemphasized that successful application of this principle requires that the learned criteria network completely learn the function it is to perform. These learned criteria are in fact performance evaluators which must be cognizant of the whole gamut of possible choices in order to provide the proper

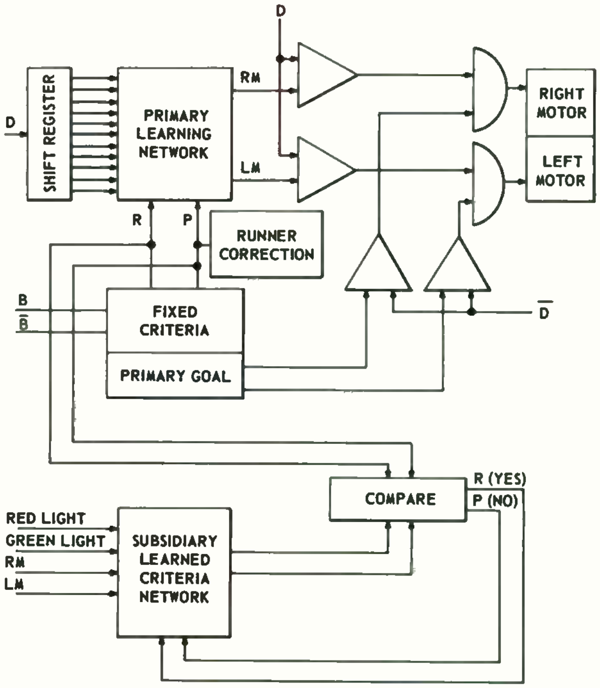


Figure 3. System Diagram of Connections Necessary to Train Runner on the Basis of "Bumps on the Nose" and Organize Additional Learned Criteria Network.

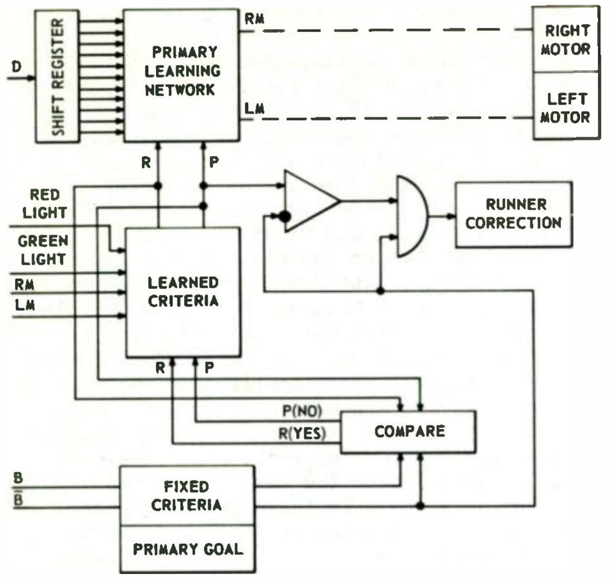


Figure 4. Series Training of Runner and Learned Criteria Network

direction by means of reward or punishment. If the evaluation does not lead to a distinct dichotomous choice, then the network will be confused by conflicting signals and will be unable to perform its basic function. The learned criteria network in this case represents a half-learned lesson and, as with humans, is of little use as a building block for more complex actions. In some cases, of course, the limitation may not be the criteria network itself, but rather the sensors and their inability to distinguish between separate instructions. In this case the appropriate portions of the network will stabilize at some level reflecting the probability of determining the correct state. An alternate scheme for teaching the runner is shown in Figure 4. Here, the fixed criteria are used to organize the learned criteria, which, in turn, organize the primary net. This scheme can be extended to other secondary sensors as shown in Figure 5 where the criteria associated with each sensor are learned in parallel under the training of the fixed criteria. The subsidiary learning networks implement specific learned criteria for each independent sensor which can, through use of the gating structure shown, be connected to the primary network in a priority order, by inhibiting the reward or punish signals from all nets except the one having highest priority. If two or more sensors are activated at the same time, the sensor with the higher priority will be the only one to reward or punish the primary net.

The above scheme assumes that some a priori priority has been established between the sensors by their particular nature or function. While this is undoubtedly true in a great number of common situations, it is also true that in other situations action is controlled by the mode in which the subject feels best qualified. This type of behavior corresponds to determining which of the learned criteria circuits having simultaneous inputs is best organized and allowing this circuit to reward or punish the primary net. In the case of networks implemented with statistical switches a quantitative measure of learning can be obtained directly from the switch states in the learned criteria circuits. Simulation of this type of behavior is therefore a matter of sampling the total number of switches in each circuit, summing their differences from the mid-point, and affording priority to the circuit having the greatest sum.

The schemes developed show how fixed criteria programmed to implement the primary goal for an elementary sensor can be used to establish learned criteria for higher order sensors without having to specify these criteria. The representation of this type system will require a large number of statistical elements distributed between the main learning network and the various learned criteria positions. Under certain circumstances it is possible that the learned criteria circuits may require more statistical elements than the primary learning net, because of the complexity of the associated sensing equipment.

Prediction Processes

The principle of the development of learned criteria can also be applied to simple prediction processes. Consider the equation

$$f(E_{n+1} \cdots E_{n+p}) =$$

$$F(E_{n-m} \cdots\cdot E_n, x_1 \cdots\cdot x_r)$$

The term on the left hand side represents a sequence of events $E_{n+1} \cdots E_{n+p}$ which will

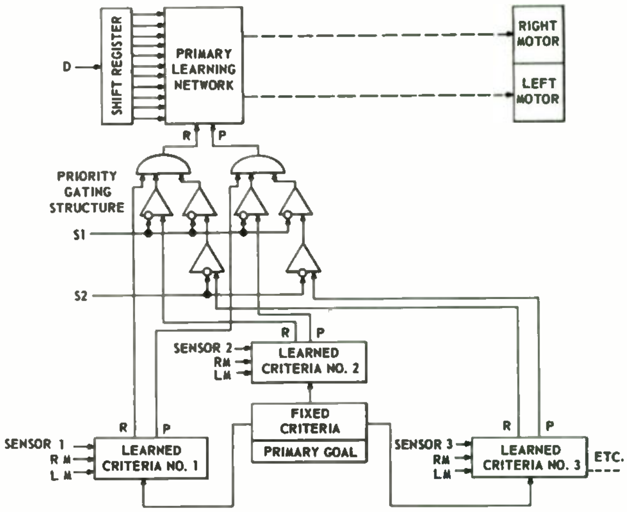


Figure 5. Extension of System to Multi-Sensor Situations

occur at some time in the future. The sequence E_{n-m} ....E_n represents a series of events which have already occurred and on which the prediction of future events must be based. The terms x_1 ....x_r represent functions such as correlation between past events. The process of prediction requires the discovery of the functional relationship F which will satisfy this equation. Provided this functional relationship is relatively simple and the quantities x_1 ....x_r represent logical connectives, this equation can be implemented by a system such as shown in Figure 6. Here successive events in the series are presented to the learning network and it is required to develop the next term. The necessary connective is developed under the influence of the learned criteria boxes indicated. These boxes have been organized by successive application of a known series of events so that they develop the connectives between the second, third, fourth, etc. previous events and the present event. Since there will undoubtedly be some disagreement between them (in general), some form of weighting must be applied to the separate reward or punish signals developed by these criteria. Summation of the weighted signals will develop an overall reward or punishment which can be applied to the learning network. Such a system can be applied with considerable facility to number series and probably to time-ordered phenomena. In theory it will also perform with mathematical progressions, however the practical considerations connected with the representation of large numbers may preclude this.

Remarks and Conclusions

The definition and development of general goal criteria is one of the more complex problems in learning system theory. The foregoing discussion is an attempt to delineate some basic principles which might be used to accommodate complex situations and allow the development of criteria from simple training situations. If the relative simplicity of teaching a learning system is to be preserved, it is essential that some such mechanism be developed.

Bibliography

1. "A Study of Generalized Machine Learning", Final Report, AF33(616)-7682, February 1962.

2. "Electronic Realization of Functional Nerve Nets", Final Report, AF33(616)-7834, March 1962

3. "A Self-Organizing Binary Logical Network", E. B. Carne, E. M. Connelly, P. H. Halpern and B. A. Logan, included in "Biological Prototypes and Synthetic Systems" (book), edited by E. E. Bernard and M. R. Kare, Plenum Press, Inc., New York, N. Y., 1962

4. "Self-Organizing Models - Theory and Techniques", E. B. Carne, Proceedings, National Aerospace Electronics Conference (NAECON), 1962

5. "Electronics Learns from Biology", A. Cerneretto, Electronic Design, Sept. 14, 1960

6. "Self-Learning Machines to use ARTRONs", Aviation Week, Feb. 20, 1961

7. "Melpar Learner uses Statistical Switch", A Cerneretto, Electronic Design, Sept. 13, 1961.

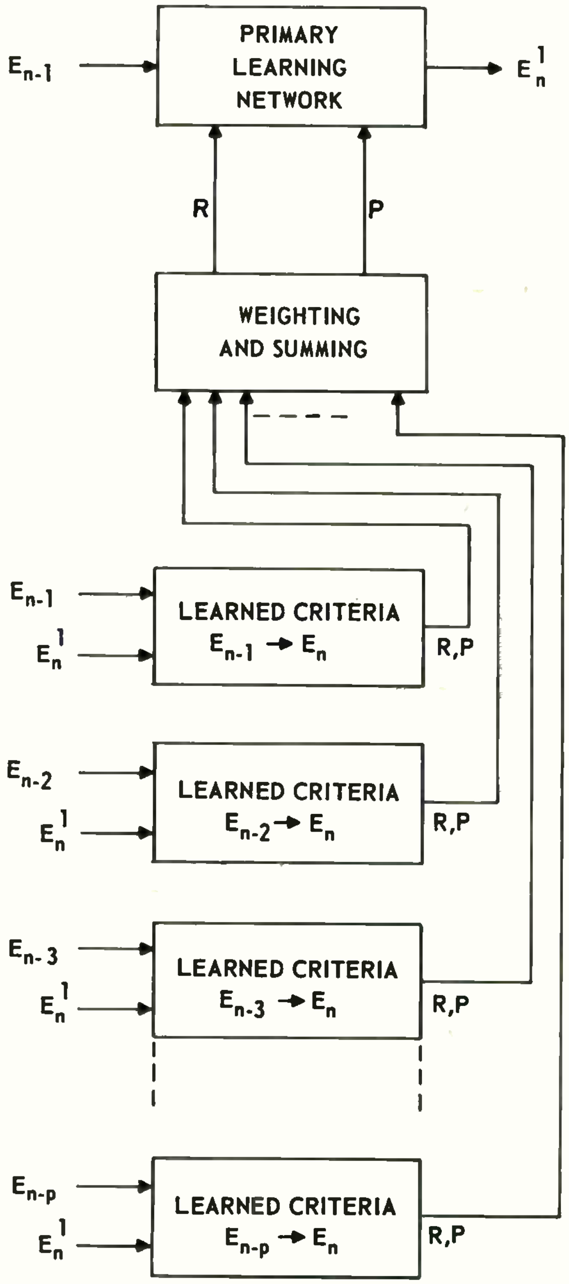


Figure 6. Extension of System to Prediction of E_n Given E_{n-1}, E_{n-2}, E_{n-3} ..... E_{n-p}

BIOLOGICAL ENERGY AS A POWER SOURCE FOR A PHYSIOLOGICAL TELEMETERING SYSTEM

Francis M. Long
University of Wyoming
Laramie, Wyoming

Summary

A brief study of three biological energy sources, biological potentials and chemical gradients, blood pressure and flow, and muscular activity and motion, revealed that the first two possibilities presented difficult problems in electrode and tissue reactions and that the third possibility might have more immediate application. A theoretical study of an accelerometer system, utilizing relative motion, indicated that several milliwatts could be delivered to the damping mechanism. A test model employing a piezoelectric crystal as the mechanical to electrical converter and a tunnel diode oscillator was successfully operated at power levels of approximately one microwatt.

Introduction

One of the major problems confronting the researcher who wishes to obtain physiological data from internally implanted radio transmitters is that of obtaining a relatively long-life power source for the transmitter. The presently available miniature batteries have an energy capability which limits either the distance of transmission or the time duration, factors which often limit the choice of experiments available. For some applications, this problem has been attacked by using low-duty-cycle pulse signals[1] and by using rechargeable batteries.[2] This paper presents some results of an investigation into another possible energy source, the use of biological energy derived directly from the animal to power the transmitter.

A general requirement placed upon such a system was that all energy required by the transmitter should be supplied by the subject and that this energy should be available for the useful life of the subject or for the duration of the test period. The following three subdivisions of possible biological energy sources were considered: 1) existing biological potentials and chemical gradients; 2) blood pressure and blood flow; and 3) muscular activity and motion. From consideration of a minimally acceptable telemetry system, a power output of at least one microwatt by the power source was deemed necessary. Studies of the tolerance of average sized dogs to surgically implanted packages gave the following specifications on the volume and the weight of such packages: maximum weight, 200 grams; maximum dimensions, 5 cm x 5 cm x 2 cm.

Preliminary Investigations

With the specifications of the power source determined, a study of the three subdivisions of biological energy chosen was begun. The preliminary study was concerned with attempting to estimate the energy available from the possible sources, with the means available, in a non-specific way, for converting such energy to a usable form, and with possible difficulties in utilizing the energy from these sources.

Existing Potentials and Chemical Gradients

Biological potentials and chemical gradients were grouped together because a device utilizing either would have many components and problems similar to a device utilizing the other and because both would be extracting energy from processes which are essentially chemical in nature. Two such sources studied were the nerve potentials and the stomach potentials.

The resting potentials and action potentials of nerves have been studied by physiologists for many years. A summary of current pertinent knowledge concerning these potentials can be found in Stacy, et al.[3] With regard to utilizing nerve potentials as a possible electrical power source, three difficulties were immediately evident. They were: 1) the voltage magnitude apparently rarely exceeds 100 millivolts, which is a small value to be directly usable; 2) the quantity of energy available would be small since it must be replaced by the normal metabolic processes of the body; and 3) the direct extraction of electrical current (electrons) from such sources appears to be a formidable problem.

The stomach potentials are potentials which exist between the mucosa and serosa of the stomach and which are thought to be in some way connected to the stomach's production of hydrochloric acid. Rehm[4], investigating this potential in the dog, found that, associated with the production of HCl, an energy equivalent to approximately 9 microwatts per square centimeter of stomach tissue area was produced. The difficulties to be overcome in order to utilize this energy are: 1) the reported potential of approximately 65 millivolts is a small value to be directly usable; 2) the equivalent source internal impedance appears to be variable; 3) suitable electrode systems would apparently be very complex and very large in size in terms of known components and techniques; and 4) permanent electrode contact with body tissues under these conditions presents problems in tissue reaction.

The presence of chemical energy gradients, such as the hydrogen ion concentration difference between the stomach and the bladder from which electrical energy might be extracted are also well known physiological occurences. However, the difficulties in the utilization of this energy are principally those already mentioned for potentials.

With regard to the direct utilization of body potentials and gradients, it appears that electrodes used must either enter directly into a chemical reaction and thereby eventually be destroyed or leave toxic residues, or electrodes which are inert may be used with the attendant problems of supplying or removing gases which, among other things, would disturb previously established contact conditions.

These electrode problems are not insurmountable. The primary consideration at present is the development of miniaturized electrode systems which can perform the necessary conversion from the physiological electron transport mechanism to the conduction type of electron flow in metallic conductors. Without such electrode systems, the true power which can be delivered by these physiological potentials cannot be accurately measured. Thus, the development of power supply systems using these potentials is also delayed.

Blood Pressure and Blood Flow

Considerable quantitative data are available concerning blood flow and blood pressure in various animals. According to Dukes[5], the approximate total power output of the heart may be calculated with the following equation:

$$\text{Energy per minute} = \frac{7}{6}\,QR + \frac{mV^2}{g} \quad \text{Kg-m}$$

where Q is the minute volume of the blood in liters

R is the resisting pressure in meters of blood

m is the mass of blood moved per minute in kilograms

V is the blood velocity in meters per second

g is the gravitational constant, 9.8 meters per second per second

With the appropriate parameter values, calculations show that the power output of the heart of a cow is about 12 watts, that of a man about 8 watts, and that of an average size dog about 0.6 watts. Thus the heart could easily supply the required power if suitable conversion methods could be found.

The Potter Instrument Company has constructed a device, the Potter Electroturbinometer, which can measure blood flow rate.[6] This device and its associated instrumentation was designed and used to measure blood velocities by having a small turbine, inserted into the blood stream, generate an electrical signal proportional to the blood velocity. Such a device might be converted to serve as an electrical power source. However, for permanent use, as the article states, the problems of blood clotting caused by foreign bodies in the blood stream, hemolysis, and damage to the platelets by mechanical abrasion still remain.

The pressure differential existing between the systolic and diastolic pressures might also be utilized by employing suitable pressure transducers. One of the major problems with such a method is the tendency of the arterial wall tissues to atrophy when their normal movement is constricted. Thus, encircling bands or surgical insertion of transducers in the arterial wall are presently of dubious value. However, there is reason to believe that this difficulty will be increasingly minimized as research continues.

The general sensitivity of the circulatory system to outside disturbances existing for any length of time presents serious difficulties. It is probable that new techniques and the modern anticoagulants such as heparin may yet lead to solutions which the circulatory system can tolerate.

Muscular Activity and Motion

Two possible methods of utilizing muscular activity and motion proposed were: 1) the use of a single muscle which could be detached at one end with no great inconvenience to the animal; and 2) the use of bodily motion, both total motion and internal relative motions. The study of muscles has been pursued in great detail for some time and the physiological and mechanical properties are well known and will not be repeated here.

The method of utilizing a single muscle detached at one end but intact at its other contact point for nutrition and nerve connection was studied. The investigation was not continued in detail when it became evident that a feedback stimulating system would be complex and when the recovery rate of a muscle from contraction was found to be very slow for continued stimulation.

A preliminary study of body motion indicated one important advantage this method possesses over those previously discussed. This was that no direct interference with body processes need be made. A device utilizing motions could be completely encased in physiologically inert materials and the only physiological problem would be that of the volume and weight tolerances of the animal carrier.

Power Available From Motion

Based upon the results of the preliminary investigations of the three possible energy sources, a more detailed study of motion was initiated. Two types of motion were considered, an approximately single frequency excitation such as that of the respiration system and an approximately random excitation such as that of the total body motion of an animal. An accelerometer method of extracting energy from these motions was chosen for study. The power available from such a system can be found by determining the power delivered to the damping. A simple mass, spring, and damping system is shown in Figure 1. Its describing equation is

$$m\frac{d^2y}{dt^2} + c\frac{dy}{dt} + ky = m\frac{d^2x}{dt^2}$$

Single Frequency Excitation

The instantaneous power delivered to the damping in such a system is

$$p(t) = c\frac{dy}{dt} \times \frac{dy}{dt} = c\left(\frac{dy}{dt}\right)^2.$$

Single frequency excitation can be described as $x(t) = X_m \sin \beta t$ where X_m is the amplitude of the excitation and β is the radian frequency of the excitation. Omitting the rather lengthy algebra, the average power delivered to the damper is, for critical damping:

$$P_{av} = \frac{\sqrt{km}\,\beta^6 X_m^2}{\left(\frac{k}{m} + \beta^2\right)^2}$$

Some typical values for these parameters might be:

X_m = 0.005 meters
m = 0.1 kilogram
= 3 radians per second
k = 10 meters per kilogram

Then,

$P_{av} \doteq$ 15 milliwatts.

Random Frequency Excitation

The problem of describing total body motion presented some difficulties. A preliminary study revealed that if an animal is subjected to random stimuli and is free to respond in any manner he selects, the animal's motion would, in all probability, also be random in nature. From studies of random signals such as those by Wiener[7], it was found that a suitable mathematical description could be made by specifying an autocorrelation function for the motion and, by analogy with the work cited, that an expression for the power available from the motion could be derived.

An autocorrelation function is defined as follows:

$$\phi(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} v(t)\, v(t+\tau)\, dt$$

where τ is the delay time, $v(t)$ is the time description of the function, and $v(t+\tau)$ is the time description of the function when delayed by the time τ. Thus, an autocorrelation function compares a given function with itself after it has been delayed and yields a quantitative measure of how well it does compare at these times. If the function is truly random, the $\phi(\tau)$ approaches zero as τ is allowed to increase, that is, there is no relation between the signal characteristics at two sufficiently displaced times.

Using the fact that the Fourier Transform of $\phi(\tau)$ is the power spectral density, or for purposes of this paper, the squared velocity spectra, $|V(j\omega)|^2$, the following expression for the average power in a random signal can be developed:[8]

$$P = \frac{c}{2\pi}\int_{-\infty}^{\infty} |V(j\omega)|^2\, d\omega\ .$$

Because the motion which actually delivers the power is not the total body motion but the motion of the accelerometer mass, this power integral must be modified to include a transfer function relating the motion of the mass to the total motion. This is a simple result from linear systems theory and yields the following power integral:

$$P = \frac{c}{2\pi}\int_{-\infty}^{\infty} |T(j\omega)|^2\, |V(j\omega)|^2\, d\omega$$

where $|T(j\omega)|^2$ is the required transfer function term.

A study of various autocorrelation functions resulted in the acceptance of the exponential form, a widely used form in work of this kind. That is:

$$\phi(\tau) = \sigma^2 e^{-\omega_1|\tau|}$$

where σ is the root-mean-square velocity and ω_1 is the so-called "cut-off" frequency of the frequency spectrum of the motion. Therefore,

$$|V(j\omega)|^2 = \frac{2\sigma^2\omega_1}{\omega_1^2 + \omega^2}$$

following from the previous discussion.

The linear transfer function for the accelerometer system in Figure 1 is found by Fourier methods and is

$$T(j\omega) = \frac{-\omega^2}{\left(\frac{k}{m} - \omega^2\right) + j\frac{\omega c}{m}}$$

or

$$\left|T(j\omega)\right| = \frac{\omega^4}{\left(\frac{k}{m} - \omega^2\right)^2 + \left(\frac{\omega c}{m}\right)^2}$$

For other accelerometer systems studied, similar expressions were obtained.

The resultant power integral for this system then becomes:

$$P = \frac{c}{2\pi}\int_{-\infty}^{\infty}\left[\frac{\omega^4}{\left(\frac{k}{m} - \omega^2\right)^2 + \left(\frac{\omega c}{m}\right)^2}\right]\left[\frac{2\sigma^2\omega_1}{\omega_1^2 + \omega^2}\right] d\omega$$

This integral is a type which can be evaluated by the method of residues and the solution is:

$$P = \sigma^2 m\,\omega_1\left[\frac{\omega_1\left(\frac{c}{m}\right) + \frac{k}{m}}{\omega_1^2 + \omega_1\left(\frac{c}{m}\right) + \frac{k}{m}}\right]$$

This result is especially interesting when compared with similar expressions developed for other accelerometer systems. For example, if the mass - spring system is replaced by a simple pendulum system, the power equation is:

$$P = \sigma^2 m\,\omega_1\left[\frac{\frac{B\omega_1}{m\ell^2} + \frac{g}{\ell}}{\omega_1^2 + \frac{B\omega_1}{m\ell^2} + \frac{g}{\ell}}\right]$$

where:

m is the mass of the pendulum bob

B is the rotational damping constant

ℓ is the length of the pendulum arm.

In these two expressions, and for others for somewhat more complex systems, the term $\sigma^2 m\omega_1$ appears as a multiplier for a bracketed expression which has a limiting value of 1.0. Thus, the maximum possible power derivable is determined by the expression: $P = \sigma^2 m\omega_1$. Some typical values for these parameters might be:

σ = 0.005 meters per second
m = 0.1 kilogram
ω_1 = 6.28 radians per second.

For these values, $P_{max} \doteq 58$ milliwatts.

The importance of this result, as for the single frequency case is that the magnitude of the power deliverable to the damping is of the order of tens of milliwatts. This is a very satisfactory value in terms of the specifications assigned to the problem.

Mechanical to Electrical Conversion

The problem of converting the available mechanical power into usable electrical power was approached by two methods. They were, 1) the use of a permanent magnet rotor a-c generator, and 2) the use of piezoelectric crystal. These two methods were to produce an alternating current which would then be rectified and the resulting direct current applied to a miniaturized transmitter.

The Permanent Magnet Generator

Several generator configurations were studied and many difficulties with this method were found. Among the more serious difficulties were 1) the reduction of the total weight and volume of the generator while maintaining a suitable magnetic flux path; 2) the "locking-in" of the pole faces of the rotor on those of the stator; and 3) the relative velocities of rotor and stator were small, requiring a large number of turns of copper wire, also conflicting with the weight and volume requirements. As a result of these limitations, construction of a miniaturized generator was not attempted.

Piezoelectric Crystal

Piezoelectric crystals of various materials have been used as electro-mechanical transducers for some time. These crystals can be obtained easily and usually are not expensive. The major advantage offered to this investigation by these crystals was their light weight and small volume.

The output voltage of a piezoelectric crystal can be very high. Values to 20,000 volts have been reported by the Clevite Corporation for a special ceramic type.[9] A typical phonocartridge will deliver 4 - 5 volts to an open circuit or very high impedance. However, the equivalent source impedance of a piezoelectric crystal is very high, normally 100,000 ohms or higher, presenting an impedance matching problem in circuits requiring a low impedance source. Measurement of the d-c power output of a Rochelle Salt crystal in a phonocartridge indicated that the crystal could deliver approximately 15 microwatts to a matched load at a voltage of approximately 0.5 volts.

A Test Model

To test the feasibility of using a piezoelectric crystal as a power source, a working model consisting of a phonocartridge, the associated impedance matching and rectifier circuits, and a simple negative resistance oscillator was built. This device was then tested to determine if the electrical properties were satisfactory. No attempt was made to miniaturize the phonocartridge or optimize its characteristics, it was used as it was purchased (for $1.27). This was probably the least desirable crystal source possible.

The Piezoelectric Power Source

Because semiconductor devices are more nearly current operated devices than voltage operated ones, an impedance matching transformer was used. It was a miniaturized transistor input transformer with a 100K to 1K impedance transformation, available from any electronics firm. It was possible to use a transformer, even for the relatively slow motions, if the mechanical coupling to the crystal was of such a nature as to allow a small amount of free mechanical vibration. These vibrations were found to vary in the neighborhood of 100 - 1,000 cps for the crystal used. The output of the transformer was then rectified with a semiconductor diode full wave rectifier with a 47 microfarad output capacitor. Although the diodes caused approximately 0.5 volts drop in the forward conducting direction, the power source was capable of delivering more than 10 microwatts at an impedance level of somewhat more than 1,000 ohms. This output was possible with a very light mechanical coupling to the Rochelle Salt crystal.

The Oscillator Circuit

The oscillator circuit connected to the power source employed a Hoffman Hu—10 Uni-tunnel diode in series with an inductor. Because the negative resistance of this connection was approximately 3,000 ohms, a series connection to the power source was possible. The resultant oscillation was a very fine sine wave. A schematic diagram of the complete test circuit is shown in Figure 2.

The power required to sustain oscillations was estimated from the d-c voltage and current requirements and was

$$P_{in} = V_{dc} I_{dc} = (0.120)(5 \text{ X } 10^{-6}) = 0.72 \text{ microwatts.}$$

The r-f circulating power was estimated by using the following formula taken from the General Electric Tunnel Diode Manual:[10]

$$P_o = \frac{I_p}{28} \text{ watts}$$

where I_p is the peak current at the beginning of the negative resistance region. Thus, for the tunnel diode used, with I_p measured on a curve tracer,

$$P_o = \frac{6 \text{ X } 10^{-6}}{28} = 0.22 \text{ microwatts.}$$

Problems to be Solved

The test model used to demonstrate the feasibility of using motion as a power source pointed up several problems which must be solved before such a system would become at least a partially satisfactory solution to the power supply problem. Some of these problems are: 1) a non-dissipative or very low dissipation voltage or current regulation method must be developed as the output of the system varied rather widely, depending upon the strength of the mechanical excitation; 2) miniaturization of the crystal support structure must be accomplished, and 3) a method of storing generated energy would be desirable for use during the periods when motion ceases. Solutions to some of these problems should be straightforward while others may require time to accomplish. The use of ceramic piezoelectric crystals should alleviate some of the difficulties because of the greater power output possible.

Conclusions

These brief investigations of direct biological energy conversion indicated that the existing potentials and gradients could deliver only small amounts of power and that the conversion of these energies to electrical current is a complex and presently unacceptable solution. The circulatory system could supply the required power but again suitable conversion methods are apparently not available at this time. Utilization of the energy of motion appears to be feasible and at least one method has considerable promise.

The study of the biological sources presented in this paper is by no means exhaustive and perhaps solutions to some of the problems described herein may be forthcoming in the near future. Certainly this study indicated that further investigations into this problem are warranted.

Acknowledgements

The author wishes to express his appreciation to Dr. Victor W. Bolie, Head of the Department of Bio-medical Electronics, Iowa State University, for suggesting the topic of this research and for his helpful guidance in many discussions during the pursuit of the investigation.

The author also wishes to thank Dean H. T. Person of the College of Engineering and the faculty and staff of the Electrical Engineering Department, University of Wyoming, for their understanding and for their assistance in the preparation of this paper.

Literature Cited

1. LeMunyan, C.D., White, W., Nyberg, E., and Christain, J. J. Design of a miniature radio transmitter for use in animal studies. Journ. Wildl. Magt. 24: 107-110. 1959.
2. Douglas, D. W. and Seal, H. R. Internalized animal telemetry system electronic considerations. Paper presented at the Aerospace Medical Association, 32nd Annual Meeting, Chicago, Illinois, April 24-27, 1961.

Multilithed. Van Nuys, Calif. Spacelabs, Inc. 1961.

3. Stacy, R. W., William, D. T., Worden, R. E., and McMorris, R. O. Essentials of biological and medical physics. New York, N. Y. McGraw-Hill Book Co., Inc. 1955.
4. Rehm, W. Stomach production of electrical energy. Amer. Journ. of Physiol. 154: 148-162. 1948.
5. Dukes, H. The physiology of domestic animals. 6th ed. Ithaca, N. Y. Comstock Publishing Co., Inc. 1947.
6. Sarnoff, S. J. and Berglund, E. The Potter electroturbinometer. Circulation Research. 1: 331-336. 1953.
7. Wiener, N. Generalized harmonic analysis. Acta Math. 55: 117. 1930.
8. James, H. M., Nichols, N. B., and Phillips, R. S. Theory of servomechanisms. New York, N. Y. McGraw-Hill Book Co., Inc. 1947.
9. Crankshaw, E. "PZT" ignition system. Clevite Corporation. Cleveland Graphite Bronze. Engineering Bulletin 13-60. 1960.
10. Tunnel Diode Manual. Liverpool, N. Y. General Electric Company, 1961.

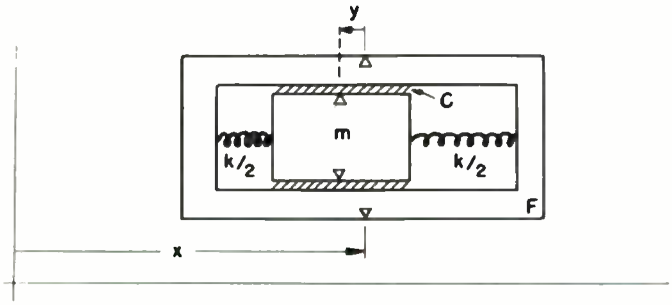
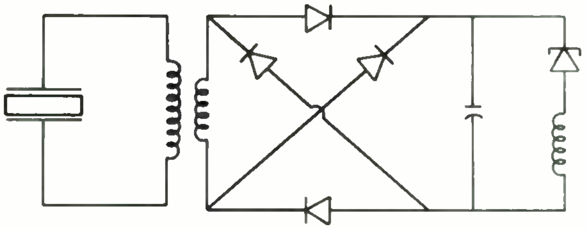


Fig. 1. Mechanical Oscillator.



Test Model Schematic
Fig. 2

ANALOGUE COMPUTATION OF ALVEOLAR GAS PARAMETER VIA DIRECT TECHNIQUE

T. W. Murphy
Associate, Sloan-Kettering Institute
New York, N.Y.

and

R. Crane
Electronic Gear Co., New York, N.Y.

Summary

The technique described uses the Bohr equation as the basis of the computation, i.e., Alveolar Ventilation

$$= \int_0^T \text{FLOW} \times \frac{(F_E CO_2 - F_I CO_2)}{(F_A CO_2 - F_I CO_2)} \, dt$$

(All values at the mouth)
(Peak $F_E CO_2 = F_A CO_2$)

The signal representing flow of gas at the mouth is obtained from a differential pressure strain gauge, and the carbon dioxide percentages from an infra-red analyzer. A system of relays (whose timing is obtained from the flow signal) used to obtain $F_I CO2$ and peak $F_E CO2$ values, and by appropriate subtraction, the peak and running difference between expired and inspired carbon dioxide concentrations. The flow signal is corrected and synchonized and the quantities multiplied, integrated (using modified standard analogue computing components) and read out in appropriate form.

Introduction

In first using analogue computer data processing in the study of respiratory physiology and pharmacology, Bellville and Seed (1) employed Gray's equation, an empirical relationship between tidal volume and alveolar ventilation (2) to obtain the latter from the former. A special purpose analogue computer was constructed by the present authors (3) using this equation and with it a comparison study of respiratory depressant effects of narcotics was performed (4). Tidal volume is easily determined by performing an integration of exhaled flow rate and applying suitable corrections for temperature, pressure, etc. Alveolar ventilation is generally considered to be of somewhat more fundamental importance than tidal volume as a parameter appropriate to the study of the function of the respiratory center. The following is a method of determining alveolar ventilation which adheres more closely to theoretical considerations, and has a higher order of accuracy.

Considering the volume of exhaled carbon dioxide (in the absence of inhaled carbon dioxide) the following series of equations can be developed:

$$V_E CO_2 = \int \text{FLOW} \times F_E CO_2 \quad dt \qquad (1)$$

But all the exhaled carbon dioxide came from the alveolar gas; (none from the dead space); and therefore represents an equivalent volume (V_A) at a concentration of gas equal to the alveolar concentration, namely $F_A CO_2$. Therefore we have:

$$V_A \times F_A CO_2 = \int \text{FLOW} \times F_E CO_2 \; dt \qquad (2)$$

From this is determined alveolar ventilation, V_A.

$$V_A = \int \text{FLOW} \times \frac{F_E CO2}{F_A CO2} \quad dt \qquad (3)$$

As $F_A CO_2$ is a constant, it can either be incorporated into the integrand or left outside. If carbon dioxide is inhaled the dead space air also contains carbon dioxide at a concentration, ($F_I CO_2$) which is greater than zero, and the equation assumes the following form:

Alveolar ventilation = V_A =

$$\int \text{FLOW} \times \frac{(F_E CO_2 - F_I CO_2)}{(F_A CO_2 - F_I CO_2)} \quad dt \qquad (4)$$

This is derived as follows:

$$V_E = V_D + V_A$$

i.e. the total exhaled volume is the sum of the gas from the dead space and gas from the alveoli, and similarly with the exhaled carbon dioxide,

$$\therefore V_ECO_2 = V_DCO_2 + V_ACO_2$$

But the gas in the dead space is the same as the inhaled gas,

$$\therefore V_DCO_2 = V_D \times F_ICO_2$$
$$= (V_E - V_A)\ F_ICO_2$$

and,

$$V_ACO_2 = (V_A)\ F_ACO_2$$

$$\therefore V_E\ CO_2 - V_E(F_ICO_2) =$$
$$V_A(F_ACO_2 - F_ICO_2)$$

and $V_E(F_ICO_2) = F_ICO_2 \int FLOW\ dt$

$$(V_ECO_2) = \int (FLOW)\ (F_ECO_2)\ dt$$

Since F_ICO_2 is constant and can be brought under the integrand;

$$\therefore (V_ECO_2) - V_E.F_ICO_2 =$$
$$\int (FLOW)\ (F_ECO_2 - F_ICO_2)\ dt$$
$$= V_A(F_ACO_2 - F_ICO_2)$$

$$\therefore V_A = \int FLOW \times \frac{(F_ECO_2 - F_ICO_2)}{(F_ACO_2 - F_ICO_2)}\ dt \qquad (5)$$

which obviously reduces to equation (3) when F_ICO_2 is zero.

The present paper is concerned with the instrumentation of this equation on a special purpose analogue computer. For a more detailed derivation of the respiratory equations, see references (5) and (6).

Transducers

A. Ventilation

The ventilation transducer consists of a Fleisch pneumotachograph, which is a heated honey-comb screen through which the subject inhales and exhales. The pressure differential across a fixed portion of the screen is then sensed by a Statham differential pressure strain gauge (Model PM 15) and amplified by a suitable amplifier, (Brush "Universal"). The push-pull output of this amplifier is then presented to the computer as the flow signal.

B. Carbon Dioxide Analysis

Carbon dioxide is analyzed with a Godart "Capnograph" infra-red sensor. The sample for analysis is obtained from a small tube inserted close to the entrance of the pneumotachograph instead of at the customary place, the lips. As the calculation depends on the formation of the product of flow and carbon dioxide fraction, it was thought more appropriate to obtain both values at the same point.

Computation

(See Figure 1)

The only portion of the computation presenting any difficulty is the formation of the product of flow and the exhaled fraction of carbon dioxide (F_ECO_2). This difficulty resides in the fact that there is a time discrepancy between the output signals of the flow and the carbon dioxide transducers. The flow-sensing device is extremely rapid and can be considered instantaneous. The carbon dioxide analyzer has both an electrical and a mechanical lag of about 80 to 100 milliseconds. This is the time necessary for the sensing membrane to change position and produce a new value of capacitance. This lapse, added to a sampling lag of similar order, yield is a delay of 180 to 200 milliseconds.

As both of these quantities (F_ACO_2 and flow) vary asynchronously, it is not possible to use an average with respect to time for either of them, because an average with respect to volume exhaled would be required. Consequently, a time delay was developed using a modification of the Padé approximation described by Dr. P.R. Hansen (7). The computer set-up performs the transfer function of the inverse Laplace transform of e^{-st} (t being the appropriate time lag). Instead of the formal expansion,

$$e^{-st} = 1 - \frac{st}{1!} + \frac{s^2t^2}{2!} - \frac{s^3t^3}{3!} +$$

a ratio of polynomials is used, which permits neglect of the higher order terms(above s^2t^2).

$$e^{-st} = \frac{1 - \frac{3\ II}{2\ ts} + \frac{II}{ts}^2}{1 + \frac{3\ II}{2\ ts} + \frac{II}{ts}^2}$$

The circuit to generate this transfer function is shown in Figure 2. Its advantage over the more usual expansion is that the feedback capacitors are identical for the two integrator stages, and it is thus easy to vary the time delay by altering the two capacitors. This type of approximation is quite adequate for the very low frequencies involved, namely those of less than three cycles per second. At these frequencies attempts to devise a tape delay are extremely difficult, as modulation and demodulation errors are quite significant. The transfer function technique seemed more appropriate and at least as accurate for these low frequencies. The response to step function, of course, is poor, but such functions are not encountered in respiratory physiology practice.

The output thus represents the delayed flow function, which is then fed into the Philbrick multiplier divider unit (GAP/R K5M) as part of the numerator. The output of this stage is also integrated to give a voltage representing tidal volume.

Computation

Carbon Dioxide:

Here it is required to find both the values for (F_ECO_2 - F_ICO_2) and (F_ACO_2 - F_ICO_2). The values for the alveolar carbon dioxide fraction and the inhaled carbon dioxide fraction (F_ACO_2 and F_ICO_2) are first determined and held on storage circuits. The CO_2 values are obtained by sampling the CO_2 signal at the appropriate times, namely at the end of exhalation and at the end of inhalation. By relay switching, the signal is transferred to the storage circuit in ten milliseconds (which is less than 1% of a typical cycle) and held there throughout the succeeding breath while computation is performed. The outputs of the stages representing exhaled, inhaled and alveolar carbon dioxide fractions (F_ECO_2, F_ICO_2, and F_ACO_2) are then fed to two subtraction circuits so that the appropriate differences are obtained. These are then inserted, one into the numerator and one into the denominator position on the multiplier-divider. Thus, the output of the multiplier-divider (which is of the form $\frac{AB}{C}$) is the integrand, namely

$$\text{FLOW} \times \frac{(F_ECO_2 - F_ICO_2)}{(F_ACO_2 - F_ICO_2)}.$$

This is then integrated and the output of this stage represents the alveolar ventilation.

Computation

Results:

The remainder of the computation consists simply in obtaining the values for tidal volume and alveolar ventilation both for each breath, and as ventilation rates on a per minute basis. The former values are obtained by feeding the outputs of the appropriate integrators to storage circuits at the proper time, namely at the end of exhalation. These stages therefore contain a voltage analogous to a volume (either tidal volume or alveolar ventilation) and are reset for the breath. The outputs of the tidal volume, alveolar ventilation and alveolar carbon dioxide stages are allowed to integrate for one minute and their outputs then transferred to storage circuits after which the integrators are re-set. The storage circuits then contain voltages representing ventilation in liters per minute, both tidal and alveolar, and also a one-minute average for the alveolar carbon dioxide fraction. These three values are read out on a Hewlett-Packard digital volt-meter and printer system. The per-breath values are displayed both on strip charts, and on an X-Y recorder, when appropriate.

Relay System

(See Figure 3)

This performs the functions referred to in the two preceding paragraphs.

Two separate but similar relay configurations are used, one operating at the end of each exhalation and the other operating at the end of each minute. The former consists of three six pole double throw relays operated by a voltage crossing detector which senses a fall in value from peak (about 20 volts) of the flow signal to 200 millivolts (which represents the end of exhalation). This is used as the signal for the operation of the six pole relays. Consider the three relays A, B and C: Relay A - When a relay A operates, its normally open contacts feed power to relay B, and when relay B operates, a pair of its normally open contacts feeds power to relay C through a 2 microfarad capacitor. Relay C's action is momentary owing to the differentiating effect of the series capacitor. A combination of a normally closed contact on A and a normally open contact on B will occur momentarily at the end of exhalation. A relay path of this form is used both to find exhaled fraction of carbon dioxide (F_ACO_2) and to readout tidal volume and alveolar ventilation at the end of each breath. As the opposing change of state occurs at the beginning of exhalation, a normally open contact on A and a normally closed contact on B will allow determination of the value for inhaled fraction of carbon dioxide (F_ICO_2). The momentary operation of relay C serves to short out the integrators so that a new computation may begin. The four-pole double throw relays used for the minute average values operate in exactly the same manner, except that the sequence of operations is initiated at the end of each minute by a timing clock.

Conclusion

An apparatus is described for analogue data processing of respiratory gas parameters which will allow determination of tidal volume and alveolar ventilation, both on a per breath basis as a volume, or on a per minute basis as a rate (liters/minute). Values for alveolar carbon dioxide fraction and inhaled carbon dioxide fraction (again on a per breath or a per minute average basis) are also derived. The minute averages are printed out, the other data recorded in strip-chart form.

Appendix

Symbols in Respiratory Physiology

The symbols used here conform to the standards published in Federation Proceedings 9:602, 1950.

General variables

V	Gas volume in general.
V	Gas volume per minute.
F	Fractional concentration of gas in dry phase.
P	Pressure, mm Hg.
I	Inspired gas
E	Expired gas
A	Alveolar gas
D	Deadspace
T	Tidal
E.G.V_A	Expired alveolar ventilation.
V_T	Tidal minute volume.

References

1. Bellville, J.W. and Seed, J.C., Science 130:1079 (1959).

2. Gray, J.S., Grodins, F.S., Carter,E. J. Appl. Physiol. 9:307 (1956).

3. Murphy, T.W. and Crane, R., R.S.I. To be published April 1962.

4. Murphy, T.W., Houde, R. and Wallenstein, S., To be published.

5. Comroe, J.S.: "The Lung".

6. Rahn, H. and Fenn, W.: "A Graphical Analysis of the Respiratory Gas Exchange", Am. Physiol. Soc.

7. Hansen, P.D., Personal communication

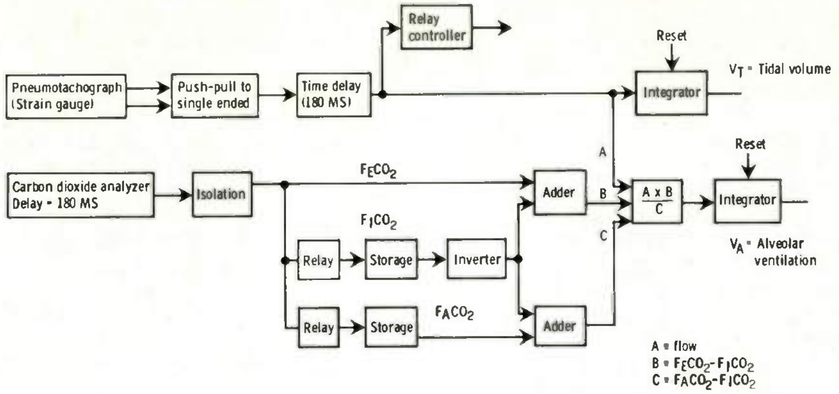
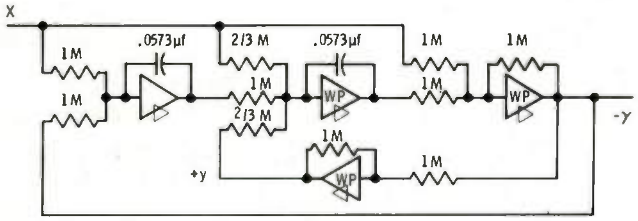


Fig. 1. Block diagram.



Delay = τ: Capacitor = $\frac{\tau}{\pi}$

180 MS time delay (suggested by P.D.Hansen)

Fig. 2. Time delay.

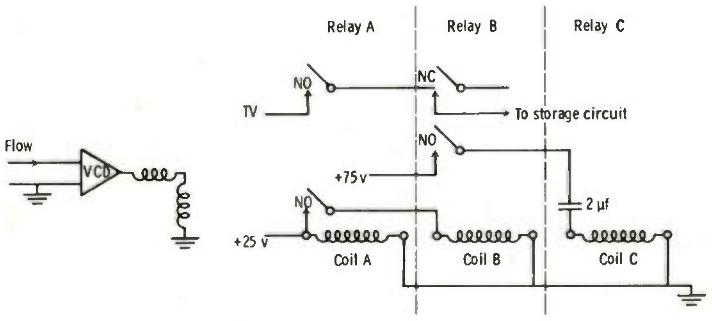


Fig. 3. Relay system (A sample path is shown).

DEVELOPMENT AND OPERATION OF CHRONICALLY IMPLANTED ELECTRONIC DEVICES WITH THEIR EFFECTS, IN EXPERIMENTAL ANIMALS*

Neal R. Cholvin,
Department of Surgery & Medicine, College of Veterinary Medicine
Michigan State University, East Lansing, Michigan

Summary

This study was an investigation of the feasibility of monitoring several physiologic parameters with chronically implanted transducers in experimental dogs. Sensors for detecting cardiac potentials, heart sounds, respiration rate and arterial pulse diametric fluctuations were developed, implanted and observed for periods of time ranging up to three months. A mathematical treatment was developed for a capacitive micrometer operating with a VHF transmission line. This type of transducer was fabricated and tested around large arteries of the dog. Problems associated with maintaining chronically implanted transducers and lead systems are enumerated. The tissue reactions and responses of the dogs to implanted electronic devices are discussed.

Implant Materials

There are two general types of materials used in surgically implanted electronic devices. Coating materials and electrical insulators make up one group. Ideally, these should be efficient electrical insulators, with low dielectric constants. They should be impermeable to body tissue fluids and to water vapor. In addition, they should be biologically inert (both chemically and with respect to foreign body effect), as well as noncarcinogenic and nonallergenic. They should be easily fabricated and under continual mechanical stress should remain stable. Materials of this group which have been implanted in mammalian tissue are polyethylene, formalinized polyvinyl alcohol, various medical grade silicone rubbers and halogenated carbons. The last two materials exhibit most of these desirable properties. Discussion of the uses and relative merits of some of these synthetics have been presented by Bing[1] and Brown, et al[2].

Electrical conductors comprise the second group. They can be used to deliver a signal from an internal source to a point on the body surface. Generally, these wires are coated with inert insulators. One reason for insulating lead wires is to minimize the dissipation of an electrical variant in the body tissues. Another reason is to prevent the electrolytic action of tissue fluids acting on metallic conductors, causing both lead deterioration and tissue injury. Venable and Stuck[3] have reported that most biologically inert metals, including the stainless alloys, show some reactivity when embedded in tissues. The third reason is to provide mechanical support for metals which fatigue and break when stressed by body movements. Several stainless metals are recommended for embedding in tissues. Of the surgical stainless steel alloys, SMO 18-8, types 316 and 304, are least affected by mechanical stress. A commercially available conductor**, consisting of stranded stainless steel wire which has been copper and then silver plated, was used for implanted leads in this investigation.

An extensible semiconductive silicone rubber*** is available. It has a volume resistivity of approximately 60 ohms per centimeter. This unique substance showed promise as an implantable strain gauge.

Implantable Transducers

Physiologic transducers which are to be chronically implanted must not only exhibit static and dynamic accuracy but must be physiologically nonreactive. They should also function faithfully in the presence of mechanical and chemical stresses imposed by body movements and tissue fluids. In order to obtain implantable transducers able to function under these stresses, some compromise of attributes for accuracy was required.

Electrodes for Cardiac Potentials

An electrode consisting of a 5 mm. terminal loop and lead wire of 0.010 inch stainless steel was fabricated. The lead wire was encased in teflon. The loop was affixed to a 2 mm. by 15 mm. by 15 mm. wafer of polyvinyl alcohol surgical grade sponge. On the surface of the wafer opposite the electrode loop was cemented a 15 mm. by 30 mm. piece of Silastic X-30146 sheet**** using Dow Type A Medical Adhesive. The spongy wafer was incorporated into the electrode design to

* This investigation was conducted in the Department of Veterinary Physiology and Pharmacology, Iowa State University, Ames, Iowa, during the tenure of a Postdoctoral Research Fellowship from the National Heart Institute, United States Public Health Service.

** Amphenol #414-004. R. F. Products, Danbury, Conn.

*** Silastic S-2086, Dow Corning Center for Aid to Medical Research, Midland, Michigan

**** Dow Corning Center for Aid to Medical Research, Midland, Michigan provide a synthetic matrix into which cells of the epicardium could penetrate, immobilizing the implant. The silicone rubber sheet, which strengthened the union between wafer and electrode, also provided a slick non-adhering surface that prevented adhesion between the implant and the pericardial sac.

Two of these devices were implanted in each of four dogs. One electrode was applied to the surface of an auricle, the other to the corresponding ventricle. Each plaque was affixed to the epicardium with a suture. This suture anchored the implant until tissue invasion immobilized it to the surface of the heart. A view of the union of the wafer with tissue cells is shown in Figure 1. Grossly, the tissue reaction provoked by the device was great. The magnitude of the direct cardiac QRS potentials detected by these electrodes did not exceed 10 millivolts for any of the four dogs. A relatively low output, plus the high physiologic "reactance" of the implant, negated the usefulness of the device.

EKG electrodes implanted in bones of the rib cage were investigated. Intraosseous placement of electrodes would tend to minimize electrical interference from skeletal muscle action potentials. In addition, the modulation of the electrocardiogram by changes in electrode position would be limited to that caused by the movements of the thoracic cage during respiration. This type of electrode consisted of a small stainless steel bone screw to which was attached the lead wire. One electrode was implanted in the sternum at the level of the fifth rib, and another in the dorsal process of the ninth thoracic vertebra in each of ten dogs. Implantation was achieved by a relatively simple surgical procedure.

Satisfactory tracings of EKG potentials were recorded for the trial for seven of the ten dogs. Figure 2 shows a tracing taken 14 days following implantation. The electrocardiograph tracings are modulated slightly by respiratory efforts. Typical recordings showed a steady baseline. In six of the ten implants, a lead wire was found to be broken and detached from the screw at the end of the experiment. Detachment was probably due to mechanical stress and deterioration from electrolytic action. In each case where detachment occurred, the wire end was securely embedded in connective tissue reaction adjacent to the rib. The recording in this situation did occasionally show baseline shifts not attributable to respiratory fluctuations.

Piezoelectric Heart Sound Transducer

This type of sensor was selected because of its relatively high output. It could function without an external power source. A small contact microphone coated with silicone rubber was implanted subcutaneously in the left thoracic area in ten dogs. The output from this heart sound transducer fell in the range of from one to five millivolts. Figure 2 shows a recording taken 14 days after implantation of the piezoelectric transducer in one of the dogs. The first and second heart sounds are easily distinguishable. Recorded sounds due to body movements were present in recordings taken during periods of vigorous activity.

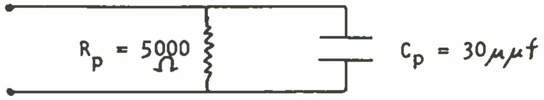
The disadvantage of this high impedance device was the progressive decline in output caused by the penetration of moisture into the crystal holder and/or into the lead system. The maximum interval over which crystal microphones functioned satisfactorily was 21 days.

In some cases, gross tissue fluid accumulation resulted when seals between components were faulty. In the remaining cases, water vapor penetrated the silicone rubber protective coating of the microphone. Sullivan, et al[4] encountered similar difficulties while investigating the operation of chest sound transducers in primates.

Capacitive Epivascular Micrometer

Rushmer[5] and Barnett, et al[6] noted the near-linear relation of blood pressure and aortic circumference in the dog. The concept and utilization of a capacitive device for detecting arterial pulse fluctuations has been reported by Adams and Corell[7]. This phase of the investigation deals with the detection of diameter changes in the large arteries of the dog utilizing a two plate capacitive transducer. The transducer plates were insulated by embedding in a thin shell of insulating serum-proof silicone rubber. Cyclic fluctuations in vessel diameter varied the spacing between the plates, correspondingly varying the capacitance of the transducer. Practical dimensions for an implantable transducer limit the capacitance of the device to less than 50 micromicrofarads. However, by energizing the transducer with VHF oscillations through a coaxial transmission line of proper length, the impedance at the input end of the transmission line can be converted to a convenient magnitude.

A capacitive transducer around the aorta of a dog would have the equivalent circuit shown below.



The impedance of this device at a frequency of 100 megacycles would be approximately 50 ohms. These values would give a quality factor, Q, of 100 for the transducer. The admittance of this device can be written as,

$$Y_1 = \frac{1}{Z} = \frac{1}{R_p} + \frac{j}{X_c} = \frac{j}{X_c}\left[1 - j \frac{X_c}{R_p} \right] \qquad (1)$$

If $\frac{1}{Q} = \frac{X_c}{R_p} = \delta$,

$$Y_1 = \frac{j}{X_c}\left[1 - j\delta\right] \qquad (2)$$

Since impedance Z_1 is the inverse of admittance Y_1,

$$Z_1 = \frac{-jX_c}{1 - j\delta}$$
$$= -jX_c\left[1 + j\delta - \delta^2 - j\delta^3 + \dots\right] \qquad (3)$$

However, terms of second order and higher can be ignored, since they will be much smaller than the term $[1 + j\delta]$, thus,

$$Z_1 = -j X_c\left[1 + j\delta\right] \qquad (4)$$

The capacitance of two parallel plates is inversely proportional to the spacing between them. McDonald[8] reported a 20 per cent fluctuation in the diameter of the thoracic aorta in the dog. It is, therefore, permissible to assume that the capacitance of this device may fluctuate approximately 20 per cent. The impedance of the transducer can then be said to vary ± 10 per cent from some value midway between systolic and diastolic values. If K represents the instantaneous fractional change from the average diameter, the value for impedance of the transducer can be represented by,

$$Z_1 = -j X_c\left[1 + j\delta\right]\left[1 + K\right] \qquad (5)$$

where $-0.1 \leq K \leq +0.1$

By expanding the expression and eliminating second order terms,

$$Z_1 = -j X_c\left[1 + K + j\delta\right]$$
$$= X_c\delta - j\left[X_c - X_c K\right] . \qquad (6)$$

This value for Z_1 may be substituted into the lossy line formula for the input impedance of a radio frequency transmission line. For ease in calculation, it can be assumed that line length is $\lambda/8$. For this investigation, a coaxial cable was used which had an impedance of $Z_o = 50$ ohms and an attenuation factor of 0.125 db per foot. Substituting these values into the formula gives the formula shown below for the input impedance of the system.

$$Z_{in} = 50\,\frac{(0.0234 + \delta) + jK}{(2 - K) + j\delta} \qquad (7)$$

A stepwise development of the mathematical treatment for this transducer is given by Cholvin[9].

Values of Z_{in} for a $\lambda/8$ tranmission line at 100 megacycles per second for selected values of K and δ are plotted in Figure 3. The characteristically low impedance of this capacitor-terminated transmission line resembles that of a series circuit around the resonance point. The addition of a $\lambda/4$ segment of transmission line to this system converts the impedance characteristic to that resembling a parallel circuit near the resonance point. This system would then have a high peak impedance value.

Changes of the impedance of the capacitor-transmission line system were detected by a locally built special VHF Wheatstone bridge. The input impedance of the transmission line could be changed by varying the length of the line. The impedance values, whether resistive, capacitive or inductive, can be calculated from the formula or can be estimated by using a Smith chart.

There are several requirements for a transducer which is to function over a period of time while implanted around an artery. First, it should be inert, so that the danger of vessel occlusion by chemical irritation is minimized. Second, the implant should fit snugly around the vessel to accurately follow cyclic dimensional changes. Third, it should not mechanically incite an occlusive tissue reaction. Finally, the characteristics (both mechanical and electrical) of the transducer must remain constant, or at least predictable, unless a means for repeated calibration can be devised.

Several models of capacitve transducers were fabricated and tested in dogs. The sites of implantation were the thoracic and abdominal aorta and the carotid artery. The best type of transducer system consisted of two 1 cm. by 2 cm. pieces of 0.005 inch brass shim stock, each soldered to short leads from the cable. These plates were then cemented to a rolled piece of Silastic X-30146 which had been calendered on straight weave dacron. This stage of fabrication is shown in Figure 4. This device was then fixed onto a glass rod and inserted into a cylindrical paper tube of slightly greater diameter than the transducer. Freshly catalized room temperature vulcanizing silicone rubber (RTV-502) was then poured into the form and allowed to solidify. When the vulcanization was complete, the form was removed and the excess rubber trimmed away. The finished product, which is shown in Figure 5, had the form of an overlapping split ring.

This type of transducer was implanted in four dogs for chronic studies. The primary drawback with this device appeared to be extreme sensitivity to movements of the transmission line between the artery and the external detector. Respiratory movements of the dog frequently caused an aberration of the pulse diametric waveform and level. This effect may have been due to changes in the longitudinal tension of the artery or to slippage of the transducer with respect to the vessel.

Figure 6 illustrates the potential value of the epivascular micrometer. This recording was taken one day after implantation of the device around the abdominal aorta distal to the origin of the renal arteries. A rather unique phenomenon occurred while this dog was under sodium pentobarbital anesthesia. Every fourth beat of the heart failed to eject blood, resulting in a pulse deficit. In the concurrent electrocardiogram, the QRS complex corresponding to the abnormal beat has a greater magnitude and a longer duration than the intervening normal complexes. Both phenomena commonly occur with idioventricular contractions. Ectopic beats are frequently too poorly organized to eject blood into the aorta. This recording of aortic diameter exhibits the pulse deficit. Each abnormal QRS complex is synchronized with a corresponding "decay" in aortic diameter.

This device was precalibrated by using a graded series of rods simulating a range of blood vessel diameters. A plot of diameter versus the d.c. detector output voltage showed a simple relationship.

Chronically implanted capacitive epivascular micrometers proved feasible from the biological standpoint. Tissue reaction did not inactivate the implants around the aorta or carotid artery.

Respiration Rate Strain Gauge Transducer

Respiration rate was successfully recorded by detecting the fluctuations of rib cage dimensions. The sensing device consisted of a piece of Silastic S-2086 rod, 0.05 inch in diameter and approximately 5 cm. long. This transducer is shown in Figure 7. Teflon insulated stranded stainless steel wire was attached to each end of the rod. A wafer of teflon served to immobilize each terminal of the transducer. The device shown here was then coated with RTV-502 silicone rubber to waterproof the terminals. Each wafer was then attached to a rib with a small stainless steel screw. When the transducer was placed under tension across several ribs, the expansion of the rib cage increased its length, and also its resistance. The change in resistance was detected by a low voltage d.c. Wheatstone Bridge circuit. The implanted transducer had a d.c. resistance varying around 50 kolohms. The response of the silicone material (change in resistance) was much too slow to record accurately the changes in length of the strain gauge. Nevertheless, the device did give an indication of respiratory rate. The resistance of the transducer appeared to remain constant following implantation. Figure 8 shows a recording taken seven days following implantation of the conductive silicone rubber strain gauge.

Implanted Lead Systems

The maintenance of electrical connections between the implanted transducers and a multiterminal receptacle on the external body surface proved to be a major problem. The two main difficulties that were encountered were corrosion of electrical connections by tissue fluids and mechanical breakage. Various materials were evaluated for use as a pass-through "grommet" in the skin. A device made of teflon functioned satisfactorily and was well tolerated in the skin of the dogs. Teflon skin prostheses remained in place for as long as 21 months. They did not appear to be uncomfortable to the animals. This grommet is shown in Figure 9. Also shown are the lead wires (Amphenol #414-004) and three transducers (EKG, sounds, respiration rate) which made up a multiple transducer unit. This was implanted in ten dogs. The terminals within the receptacle were waterproofed by etching the teflon grommet and adjacent insulated wire with Tetra-Etch* and then forcing freshly catalyzed RTV-502 into the base of the grommet, receptacle and adjacent wires. Adhesion of the silicone rubber to teflon is enhanced by the etching process. Etched teflon pretreated with Dow A-4094 primer produced a bond with RTV-502 sufficiently strong to withstand the stress of body movements. The etching procedure is necessary in order to obtain a fluid-proof seal.

Disruption of the leads between the transducers and the skin terminals occurred consistently with wires made of either copper or #302 monofilament stainless steel. When #316 wire was used, disruption was much less common. Better duration of continuity of #316 wire leads may have been due to the increased resistance to fatigue possessed by this alloy. Stranded stainless steel wire with snug teflon insulation gave the best service. This wire also deteriorated when exposed to tissue fluids, however.

Mechanical fatigue of lead wires appeared to be minimized by relatively large amounts of tissue reaction. This reaction consisted of fibrous connective tissue. It tended to support the implanted materials. Relatively great tissue reaction is not desirable, though, because the parameters which are being observed may be altered, and because the implants might cause discomfort.

Implantation Procedures

Implantation experiments were designed to evaluate materials, tissue reactions, transducers and surgical procedures. Forty-three dogs of mixed breeding were used in this study. Dogs of various weights and sexes, and indeterminate ages were used. Most of the dogs used for chronic studies were mature but less than two years of

* W.L. Gore Associates, 487 Papermill Road, Newark, Delaware

age. Only those judged to be in good health by a physical examination were selected for the trials.

All materials which were used in implantation procedures for chronic studies were sterilized. Steam under pressure was employed to sterilize surgical instruments, gloves, linens and implant materials which could withstand 121° C. Chemical disinfection by immersion in aqueous 0.02 per cent chlorhexidine provided the means for disinfecting materials with low heat tolerance, or instruments with sharp cutting edges that would become dull when repeatedly autoclaved. Ethylene oxide gas sterilization also would have been suitable for these materials.

The dogs were anesthetized by intravenous administration of sodium pentobarbital given to effect. Preparation of the animal for surgery consisted of removing the hair over the surgical areas with an electric clipper, and then repeatedly scrubbing the exposed skin area with a 25 per cent aqueous germicidal detergent solution. This solution contained 0.06 per cent benzethonium chloride.* If intrathoracic surgery was to be performed, an endotracheal catheter with cuff was inserted into the trachea. The cuff was inflated and an artificial respiration apparatus was attached to the catheter.

The hands and arms of the surgeon were repeatedly scrubbed with the germicidal detergent, rinsing after each scrub with tap water. The surgeon then dried his hands and arms with a sterile towel and donned a sterile full length operating gown. He then put on dry, sterile surgical gloves. The animal was then draped with sterile towels, leaving the surgical field exposed.

The success of implantation procedures especially depended upon careful surgical technique. Foreign materials were introduced into the tissues. Some might have been irritating. Others might have provided an ideal medium for infection. Principles of good surgical technique are: avoiding contamination and undue trauma to tissues, controlling hemorrhage and carefully closing divided tissues by meticulous suturing technique.

In a pilot study, strips of polyethylene, polyvinyl chloride, silicone rubber and teflon were inserted through a cannula into a bed of subcutaneous tissue in the lateral thoracic area. The gross tissue reaction to each material was observed following implantation. At appropriate intervals, a small specimen of tissue was removed for microscopic examination.

Each teflon skin grommet was implanted in the mid-dorsal thoracic skin posterior to the scapulae. A small circular plug of skin was removed with a sharp 3/16 inch tubular trephine. A radial incision was made joining the circular incision. The device was then inserted into the skin and subcutaneous tissues. The lead wires coming from the integral multiple transducer implant were then carefully sutured to the surrounding subcutaneous connective tissues. A purse-string suture was then placed in the skin around the grommet to support the tissues during the healing period. The transducers were next implanted by passing them through subcutaneous tunnels into the implant beds. Through small skin incisions each was then fixed in place by either sutures or screws. Each day for two weeks following implantation it was necessary to treat the skin around the grommet by cleaning away serum encrustations and by applying an antibiotic wound powder. After this interval, the skin usually healed firmly around the threaded shaft of the grommet. Exudation then stopped. Inflammation of the skin in this area rarely occurred after the initial healing period.

Lead wires for transducers on the heart or aorta were introduced into a tunnel in the mediastinum. This tunnel was produced by inserting a trocar and cannula from a point in the subcutaneous tissue approximately two inches lateral to the skin grommet. The cannula passed between fibers of the latissimus dorsi, serratus ventralis and dorsalis and iliocostalis dorsi muscles through the membranus septum between the longissimus dorsi and iliocostalis dorsi muscles. An entrance was then made into the sixth intercostal space subpleurally into the dorsal mediastinum. A small incision in the mediastinal pleura adjacent to each transducer implant provided an opening through which the lead wires could be inserted through the bore of the cannula. Removal of the cannula left the teflon insulated lead embedded within mediastinal and superficial tissues. The development of excessive tissue reactions and adhesions was minimized by this method of implantation because the pleural cavity was not traversed by foreign material.

Tissue Reactions to Implants

The tissue response to several synthetic materials was observed in the skin, subcutaneous tissue, mediastinum, epicardium and tissue surrounding the aorta and carotid arteries. Serial gross observations were made and specimens taken for microscopic examination. The synthetic materials which were evaluated were polyethylene, polyvinyl chloride, several silicone rubber products, teflon and polyvinyl alcohol sponge. The general type of tissue reaction to all but the last material, when implanted subcutaneously, was the same. The degree of reaction varied considerably, however, and was greater with polyvinyl chloride than with polyethylene implants. The silicone rubber materials, especially medical grade Silastic X-30146 and silicone rubber tubing, incited less reaction. RTV-502 provoked a slight degree of reaction. Teflon caused slightly less tissue

* Germicidal Detergent. Parke, Davis, & Co. Detroit, Michigan

change than did the medical grade silicone rubber. In general, the cellular reaction started with the development of a layer of primitive mesenchymal cells, one to several layers thick, around the implant. As the time interval lengthened, the peripheral layers of mesenchymal cells began to differentiate into immature fibroblasts. Frequently neutrophils and macrophages were intermixed with the fibroblasts. Maturation of the outer layers of fibroblasts followed and collagenous fiber accumulations appeared. Figure 10 shows the layers of tissue reaction to an uninsulated copper lead wire. The degree of cellular response is great. In contrast, Figure 11 shows minimal reaction to medical grade Silastic X-30146. All of the zones mentioned above, though very thin, can be identified.

The open cell sponge made of polyvinyl alcohol was infiltrated, rather than encapsulated, by living tissue. Plaques placed upon the epicardium were invaded by cellular components. The end result was the filling of the spaces by fibrous connective tissue.

Teflon skin grommets produced extensive changes in the skin. The prostheses were examined after implantation periods ranging from 8 to 656 days. Each device was surrounded by a ring of thickened, indurated skin. The film of tissue adjacent to the teflon appeared to be friable. The large amount of tissue reaction was not surprising. This developed because the grommet itself was large. Its mass, accelerated by body movements, exerted considerable trauma to the surrounding tissues. The inflammatory response to microbial contamination may also have contributed to a large tissue reaction. Figure 12 shows the tissue around a teflon skin grommet 656 days after implantation. The tissue here appears to consist of mature fibrous connective tissue with scattered accumulations of lymphocytes. Undifferentiated mesenchymal cells may have been present at the surface adjacent to the teflon, but were probably in the friable and mucilaginous membrane which was lost during sampling.

Minimum tissue reactions followed the subcutaneous implantation of teflon and medical grade silicone rubber. It would therefore seem advisable to choose from either group a material to coat irritating implantable devices. However, considerations other than minimal tissue reaction may govern the choice of coating materials. For example, where flexibility is desired for electrical leads, teflon may be too stiff, except for wires of small diameter. Both silicone rubber tubing or a coating of RTV-502 are quite flexible. Silicone rubbers are known to be pervious to water vapor, however. An extremely smooth, inert surface may be desirable to minimize the formation of a thrombus around an intravascularly implanted device. A silicone varnish may be acceptable if the implanted device is rigid since cured varnishes are brittle. If the object is flexible, however, a teflon coating may serve best, although fabrication of such a coating with a shiny unblemished surface of teflon might be difficult. Conversely, for some implants a rough textured inert surface may be desirable. Such an implant would become firmly embedced in the tissue and would not undergo migration.

In general, each implanted device and/or lead system presents a unique problem. Selection of an acceptable insulating and coating material depends upon preserving electronic specifications while still observing the desirability for a biologically tolerable substance.

References

1. Bing, J. Tissue Reactions to Implanted Plastics. Acta Path. et Microbiol. Scand. Suppl. 105: 16-26. 1955.

2. Brown, J.B., Fryer, M.P. and Ohlwiler, D.A. Study and Use of Synthetic Materials Such As Silicones and Teflon As Subcutaneous Prosthesis. Plast. Rec. Surg. 26: 264-279. 1960.

3. Venable, C.S. and Stuck, W.G. A General Consideration of Metals For Buried Appliances in Surgery. Internat. Abstr. Surgery 76: 297-304. 1943.

4. Sullivan, G.H., Schulkins, T.A. and Freedman, T. Internalized Animal Telemetry: Biomedical and Surgical Considerations. Paper presented at the Aerospace Medical Association 32nd Annual Meeting, Chicago, Illinois (Mimeo.) April 25, 1961. VanNuys, California, Spacelabs, Inc. 1961.

5. Rushmer, R.F. Pressure-Circumference Relations in The Aorta. Am. J. Physiol. 183: 545-549. 1955.

6. Barnett, G.O., Mallos, A.J. and Shapiro, A. Relationship of Aortic Pressure and Diameter in The Dog. J. Appl. Physiol. 16: 545-548. 1961.

7. Adams, R. and Corell, R.W. Cuffless, Noncannula, Continuous Recording of Blood Pressure. Statement of Concept and Description of Methods for Measurement Based on Principle of Capacitance. Surgery 47: 46-54. 1960.

8. McDonald, D.A. Lateral Pulsatile Expansion of Arteries. J. Physiol. 119: 28P. 1953.

9. Cholvin, N.R. Surgically Implanted Electronic Devices for Use in Experimental Physiology. Unpublished Ph. D. Thesis. Ames, Iowa, Iowa State University of Science and Technology Library. 1961.

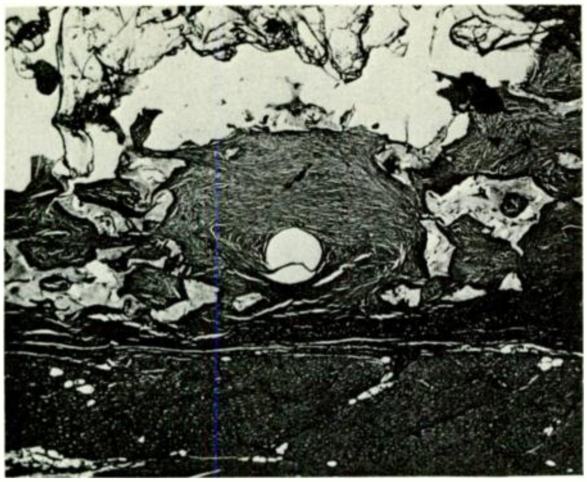


Fig. 1. Microscopic section of ivalon epicardial plaque and myocardium 64 days following implantation stained with hematoxylin and eosin. 36 X.

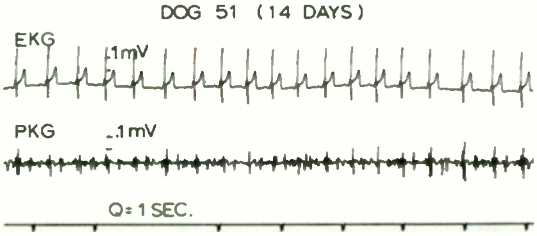


Fig. 2. Recordings of heart potentials and sounds taken 14 days following implantation of a multiple transducer system.

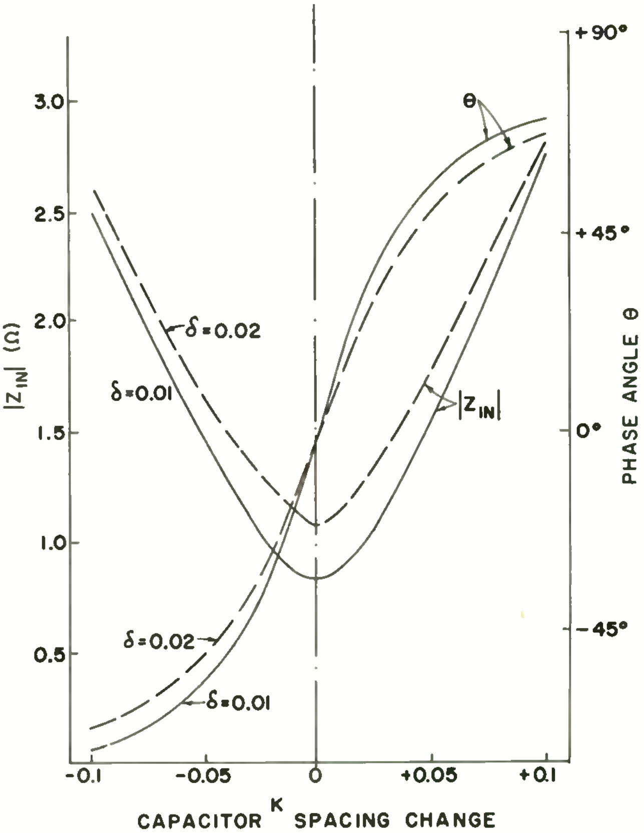


Fig. 3. Plot of calculated values for input impedance and phase angle versus change in plate spacing for a /8 transmission line terminated by a capacitive epivascular micrometer.

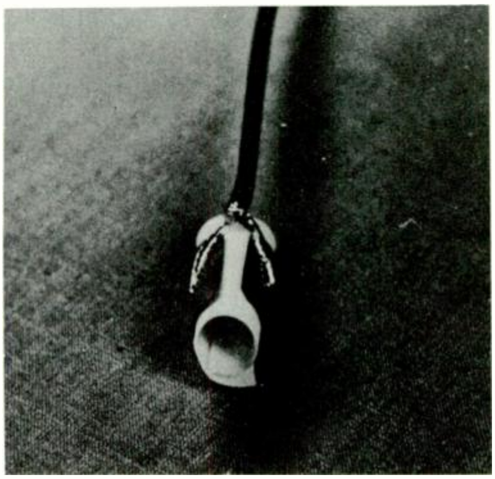


Fig. 4. Capacitive epivascular micrometer during construction.

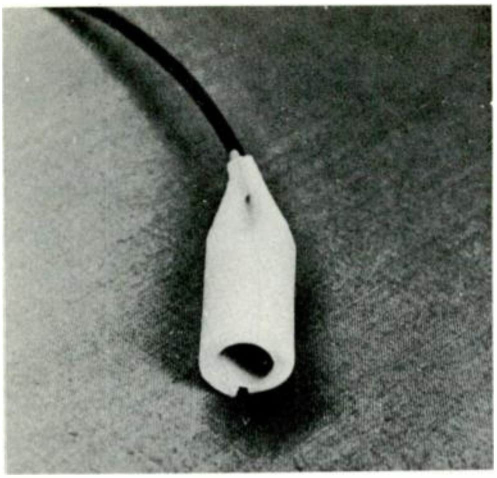


Fig. 5. Capacitive epivascular micrometer after coating with RTV-502.

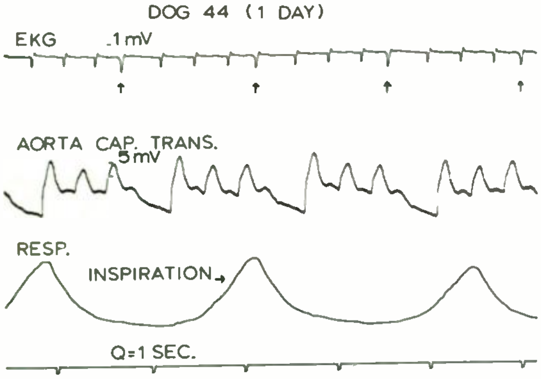


Fig. 6. Recording of heart potentials, aortic diameter and respiration of an anesthetized dog showing aortic pulse deficits as detected by an implanted capacitive epivascular micrometer.

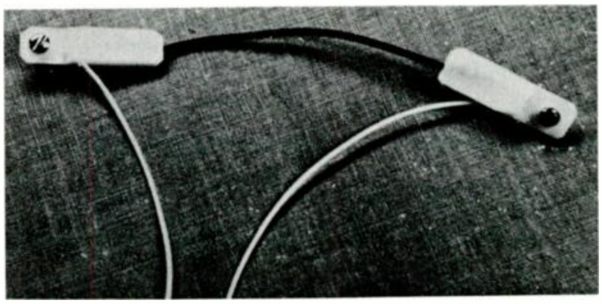


Fig. 7. Conductive silicone rubber respiration transducer.

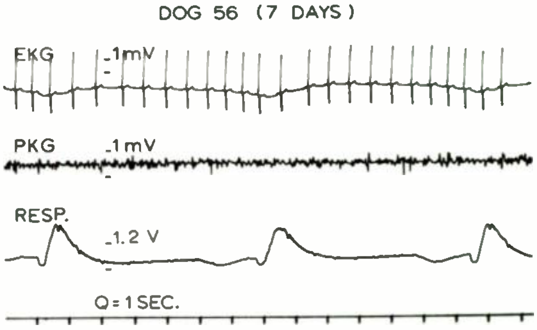


Fig. 8. Recording of heart potentials, sounds and respiration rate taken seven days following implantation of a multiple transducer system.

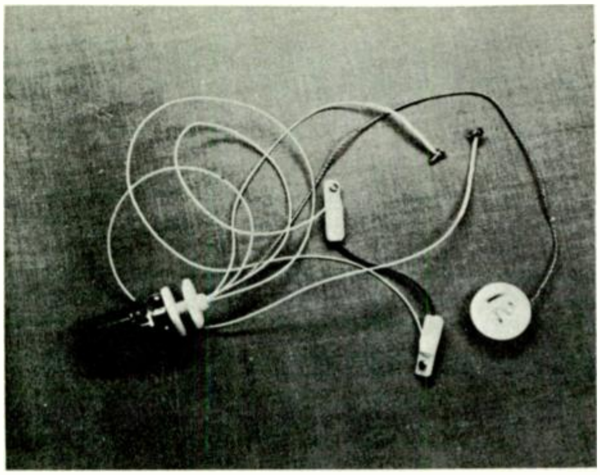


Fig. 9. Integral multiple transducer and lead system (assembled, before coating with RTV-502).

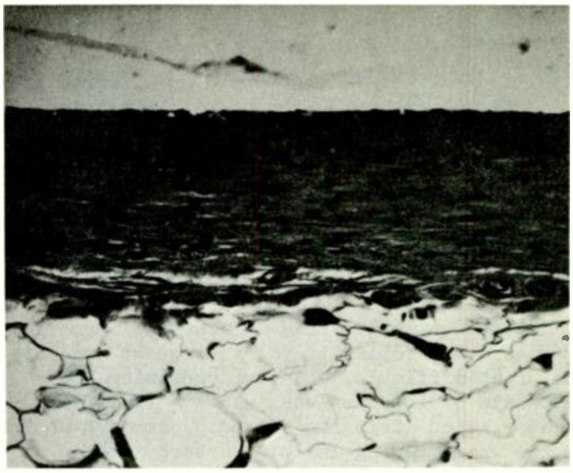


Fig. 11. Microscopic section of tissue reaction 50 days following implantation of medical grade Silastic in the lateral thoracic wall subcutaneous tissue, stained with hematoxylin and eosin. 240 X.

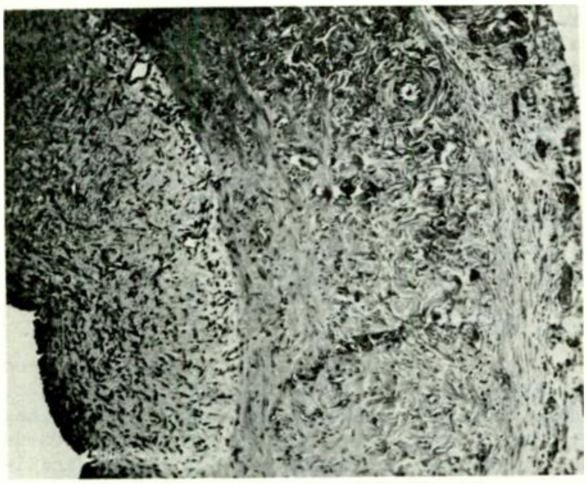


Fig. 10. Microscopic section of tissue reaction 127 days following implantation of a copper wire in the lateral thoracic wall subcutaneous tissue, stained with hematoxylin and eosin. 95 X.

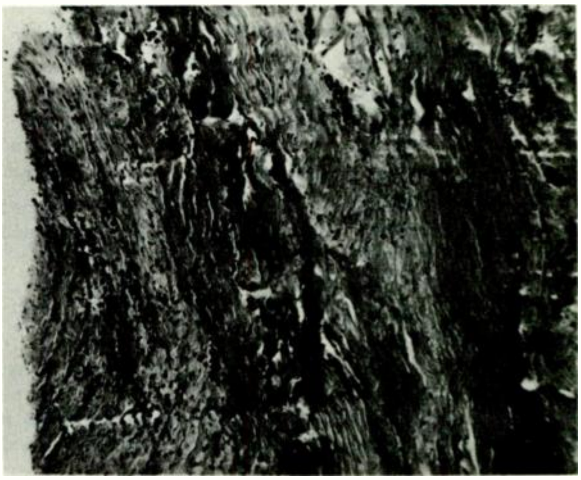


Fig. 12. Microscopic section of tissue reaction 656 days following implantation of a teflon grommet in the dorsal thoracic wall skin, stained with hematoxylin and eosin. 95 X.

MEDICAL MAGNETIC TAPE RECORDING
IDENTIFICATION AND SEARCHING SYSTEM

L. W. Paine and C. A. Steinberg

Airborne Instruments Laboratory
A Division of Cutler-Hammer, Inc.
Deer Park, Long Island, New York

In recent years there has been increased use of simultaneous recording of physiological data on both graphical recorders and magnetic tape recorders. In many instances, the researcher examines the graphical record for data of interest. He then desires to locate this data on the magnetic tape record and replay it for subsequent analysis. It is often difficult, and time consuming, to locate data of interest when using a tape footage indicator, when listening to audio comments on the tape, or when viewing the played-back data on an oscilloscope or graphical recorder.

To automatically locate a given section of data on magnetic tape, a tape coding and searching system has been designed and constructed. The coding device generates digital code numbers for recording on graphical and magnetic tape recorders. The searching device automatically searches the magnetic tape record for any selected code number.

The coding system records the desired number in a serial, binary-coded-decimal code on one track of the magnetic tape, at the same time that analog data is being recorded on other tracks. Analog data may also be recorded in the code number track, by means of time sharing.

The search system separates the code numbers from the data using unique characteristics of the code number as a criteria for acceptance, and can automatically search for a particular code number.

The manner in which the code number is recorded can take many forms. For example, binary ones and zeros can be recorded as positive and negative levels, positive and negative pulses, large and small pulses, or tone bursts of different frequencies.

The system to be described will use positive and negative voltage levels to represent the ones and zeros. In addition, this coding system uses a three digit code number, giving 1000 different code numbers.

The block diagram for the recording device is shown in Figure 1. A start pulse turns on a multivibrator used as a controlled clock. Each clock pulses causes a 4-bit ring counter to advance. Each time the 4-bit ring counter recycles, it causes a 3-bit ring counter to advance. The 3-bit ring counter is used to select the digit to be recorded, and the 4-bit ring counter is used to select the bit to be recorded. The number to be coded is set into a digital switch. This switch has ten positions, with appropriate binary coded decimal outputs for each position. The 3-bit ring counter selects the proper decade of the digital switch, and the 4-bit ring counter causes the appropriate bit to be gated into the output circuit. Each clock cycle further gates this bit to the output, where it becomes the code signal to be recorded on magnetic tape. The operation of the coder is shown in the timing diagram (Figure 2).

Operation is initiated by a start pulse which turns on the clock. At the end of each clock cycle the bit ring counter is advanced. This ring counter sequentially scans the 8, 4, 2, and 1 bit values. When the bit ring counter recycles, the digit ring counter advances. This ring counter scans through the hundreds, tens, and units digits. In the case shown, the code number 963 will be recorded. The gates in this case will select the 8 and 1 of the hundreds digit, the 4 and 2 of the tens digit, and the 2 and 1 of the units digit. These binary values are gated through to the coder output during each cycle. The coder output consists of a positive voltage to represent a binary one, and a negative voltage to represent a binary zero. After the last bit has been recorded, the cycle is complete, and the coder automatically stops.

The net result is that each time the coder receives a start pulse, an output is generated consisting of a series of positive and negative levels. This is the binary coded decimal representation of the desired code number. This code is in serial form, high bits first with a positive output to represent a binary one, and a negative output to represent a binary zero. Between bits, and when no code number is being generated, the output is zero volts.

The code number generated can be easily read and interpreted on a graphical chart recorder, providing that the clock rate and output voltage magnitudes are compatible with the graphical recorder. Ease in reading the number can be further facilitated by providing a space between each digit or set of 4 bits, as shown by the code number with gaps.

The search unit, shown in Figure 3, is an asynchronous device which automatically locates and identifies any preset code number. In operation, the input signal from the played-back magnetic tape recording enters the decoding circuit where a clock pulse is formed for each bit of the entering code number. This clock pulse is used to shift the bit value, whether a one or a zero, into a 12-bit shift register. The full register bit is used to tell when all 12 bits of the code number have entered the shift register. At this time, the register contents are compared with the digital switch contents by the coincidence check circuits. Coincidence can only be checked after the entire number is entered into the register, because the register contents passes through many different states while "reading in," and one of these states could coincide with the switch setting. The particular code number being searched for has been set into the digital switches by the operator. If coincidence is found, a pulse is generated which is used to stop the searching process and start the next process. For example, this might be an analog-to-digital conversion of the following data. A timing monitor circuit is used to assure synchronism between the incoming code number and the state of the shift register. The register must be set to zero before the first code bit enters, so that coincidence will be checked at the end of the code number and not at the middle. This timing monitor is, basically, a delay circuit which causes the register to reset some delay time after a clock pulse enters the monitor. By adjusting the delay time to be greater than the maximum time between bits, and by assuring that the delay time is restarted for each clock pulse, reset of the register will occur whenever there is a gap between bits greater than the delay time. This gap will occur between code numbers. In addition, the timing monitor will cause register reset if recorded noise or signals between code numbers are interpreted, by the decoder, to be bits.

A number of methods can be used to assure that noise or artifacts will not cause a false detection of coincidence. These become part of the decoding process. Much of the noise can be filtered, and the inherent redundancy of the bit width can be used to satisfy decoding criteria. For example, a voltage excursion of less than, or greater than, a specified amount can be rejected as not being a bit. Voltage excursions either shorter or longer in duration than a specified amount can also be rejected as non-bits. Another protection method is to record a "key" or identifying mark ahead of each code word. This mark (a tone burst at a particular frequency, for example) would unlock the register and allow it to read in the code. In general, as the conditions that a code must satisfy to be accepted as a bona-fide code increase, the likelihood of detection of false coincidence decreases.

This system, both coder and searcher, is entirely digital in nature and uses standard digital circuits such as ring counters, flip-flops, logic units, and one-shot multivibrators.

A number of features may be incorporated into this basic system to greatly increase its versatility and usefulness. One of these, mentioned before, is recording the code numbers in the same tape channel as the data. This is done by switching the recorder input from the data to the coder output by relay, sufficiently delaying the generation of the code number to allow the register to clear, and switching the recorder back to the data when the code number has been recorded.

By storing the code number on stepping switches, solenoid driven rotary switches, or the like, the code number can be automatically advanced each time it is used. Adding a cyclic timer to the system will allow periodic operation. Pulse starting can be used so that the coding process can be triggered automatically, from other equipment. A pulse can be generated at the end of the coding process for uses such as triggering a stimulus in neurophysiological experiments. Digital reading indicators can be used in both the coder and the searcher. The former are useful in the long term experiments where periodic coding is used and the experimenters have to refer to the code number. The latter is useful to indicate to the operator the various codes on the tape as they are played back. Another easily incorporated feature is high speed search. This is accomplished by setting the timing requirements of the decoder and timing monitor in the searcher to correspond with the code signals that are reproduced when the tape is played back at maximum speed.

Magnetic tape identification systems of the type described are currently being used for medical data processing in a number of areas.

One such application, at a large midwestern medical center, is in the field of research. Here, six channels of experimental data are simultaneously recorded on a magnetic tape recorder and a strip chart recorder, in analog form, while code numbers are recorded either periodically or by manual pushbutton in a seventh channel. The coding device being used is shown in Figure 4. This coder will generate a recordable code number each time the "operate" switch is depressed, or periodically as controlled by an interval timer. Various bit rates can be selected to correspond with appropriate chart speeds on graphical records. The number to be coded can be set manually on the code number switches. This number automatically advances each time it is used. A remote decimal display of the code number is also provided, but is not shown in this figure. In this application, an editing process is used where sections of recording acceptable for computer processing are visually selected from the strip chart and identified by their corresponding code numbers. Further processing then consists of automatically searching the analog tape for the selected sections of record on the basis of the code numbers. Once located, these sections are then converted from analog to digital form and recorded on digital magnetic tape in a format compatible with a general purpose digital computer.

In another application, a coder is used in a data collection system. In this application, recording charts are used to record ECG's on magnetic tape and on a graphical record. Figure 5 shows the magnetic tape recorder, the graphical chart recorder, and the control panel. The coder, as well as other electronic circuitry, is contained within the cart. In this application, the code number consists of a patient number, location number, and lead number, and is recorded before each electrocardiogram. Both the code number and the electrocardiogram occupy the same tape channel. Up to 400 ECG records and their identifying code numbers are recorded on a single 7-inch reel of 1/4-inch magnetic tape. Using the code numbers, desired cardiograms are automatically retrieved for visual inspection and analysis, or for entry into a computer for automatic analysis.

In summation, this coding and searching system provides a means for identifying and automatically locating specific sections of magnetic tape recordings, and has already proved itself useful in the medical field.

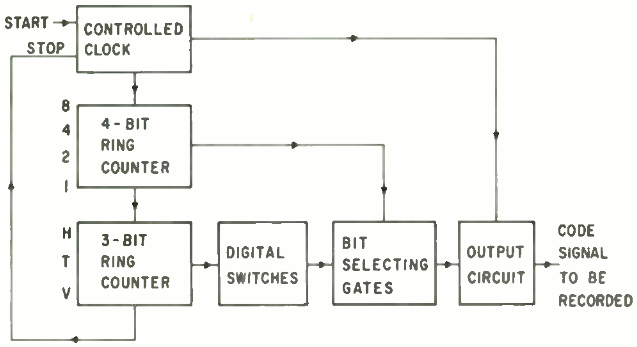


Fig. 1. Magnetic tape coder.

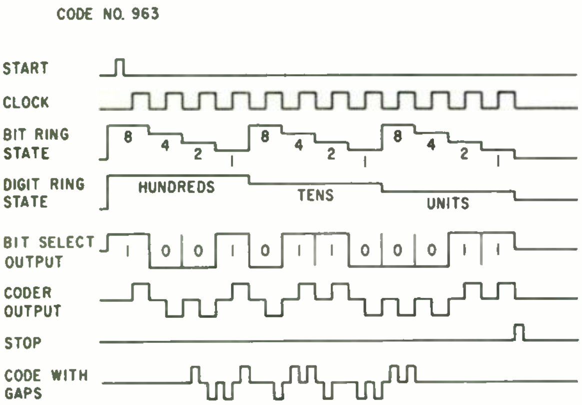


Fig. 2. Timing diagram magnetic tape coder.

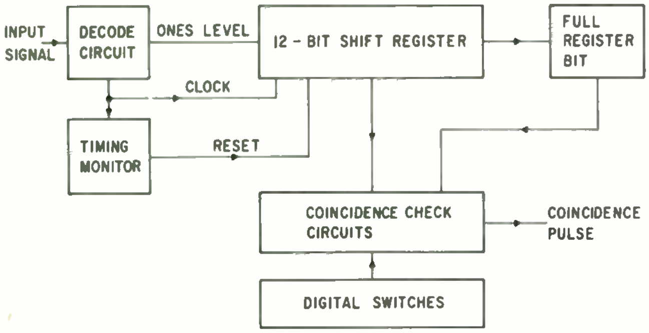


Fig. 3. Magnetic tape searcher.

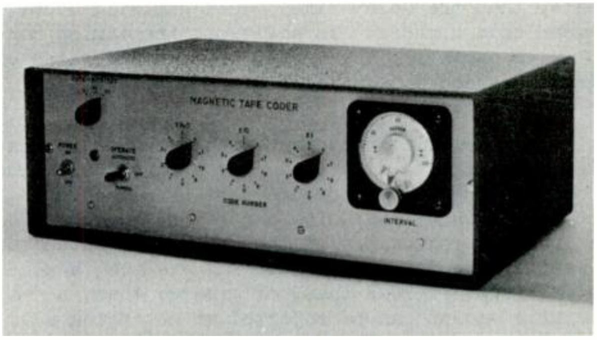


Fig. 4. Tape coder.

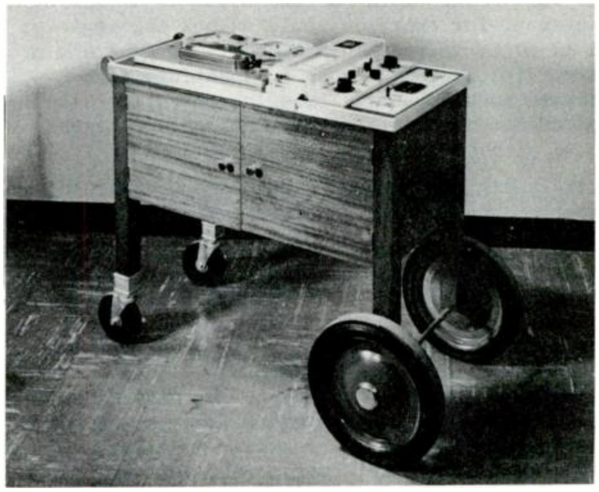


Fig. 5. Recording cart.

ENGINEERING IN THE LIFE SCIENCES

James B. Hartgering
Technical Assistant
Office of the Special Assistant for Science & Technology
The White House
Washington, D. C.

SUMMARY

The potential contribution of the engineering sciences to the life sciences in basic research, problem oriented studies involving man-machine complexes, and the development of instrumentation are recognized. The basis for development of any new field, particularly one involving integration of several disciplines, will depend on the education pattern devised. Experimental programs now offered by a few universities, leading to a degree in biomedical engineering, may result in graduates with limited potentials. An educational program is needed which will insure that first-rate life scientists and first-rate engineers have a working familiarity with the problems of and an understand ing of each others technologies.

The last decade has seen major developments in the biological sciences, and in particular those related to medicine. As a result of national interest, both within the government and the voluntary health agencies, about a billion dollars per year is provided for the support of research and training. Exciting contributions have been made to the field by those interested in the physical sciences. For example, the Nobel prize in medicine was awarded last year to Dr. Georg von Bekesy, a physicist. Recently there has been an increasing awareness of the potential contribution of the engineering sciences to the life sciences.

A good deal has been written about the needs and requirements of an interdisciplinary approach to research and development. As a matter of fact, this has become fashionable. However, when one examines closely the productive interdisciplinary groups, it usually turns out that the individual scientists, drawn together because of the scientific challenge, had very similar academic educations -- emphasis in the basic sciences -- although their degree might be in medicine, physical chemistry or mathematics. Discussions with medical and engineering educators make it clear that we are in the midst of an intellectual crisis which is having profound implications on our educational system. Pursual of curricula in catalogs of our universities shows an abrupt change during the past few years towards the emphasis of basic science irrespective of the final degree to be granted by the institution.[1]

The interests leading to the development of engineering in the life sciences may be somewhat arbitrarily divided into three areas: the study of fundamental questions in which a working knowledge of the theories of electricity, hydraulics, or thermodynamics provides new opportunities; problem oriented research involving man-machine complexes as typified in the Department of Defense and the National Aeronautics and Space Administration; and finally, the development of instrumentation to measure specific biological factors.

One can find examples to illustrate the problems and principles involved in each of the three areas, but perhaps, your familiarity with the computer makes it the easiest to discuss. Computers have been available for a good many years, but have, as yet, had little impact on the life sciences, except in a few universities such as the Massachusetts Institute of Technology with an interest in communication or related problems. When computers are available, they are utilized to perform primarily statistical functions. The complexities inherent in most biological systems have so far not been readily adaptable to the available input-output equipment. Most biologists educated even within the past decade have not had sufficient access to computers to develop either an appreciation of their potential as an intellectual tool, or to realize their inherent limitations, and only an occasional professor of engineering is working in a medical school or university department of biology. The possibilities are enormous and starts are being made, however shaky, in such traditional subjects within the art of medicine as diagnosis, medical records, and the coupling of physical examination data with clinical pathology. The hope is that the computers will not only permit the handling of large amounts of data, but more important will force us to ask ourselves critical questions.

The development of the behavioral sciences could be facilitated materially by intelligent applications of computer technology. The volume of data needed to understand communications among groups of individuals, the variables inherent in the study of personality, and our desires to understand the processes by which we think are dependent upon the flexibility of the computer. Many behavioral scientists feel that the computer may be the tool that will permit them to harden their data to the point that the relations between, for example, neurochemical and neuroelectrical events in the brain can be understood in terms of meaningful human behavior.

It is interesting that although we have gathered voluminous data on economic and other practical aspects of our everyday lives, we have little access to the demographic studies so essential to the establishment of cause and effect relationships in disease, and to an understanding of human genetics. The recent establishment of the National Health Survey within the Public Health Service is the first step to obtaining the relatively simple census-type data necessary for a population approach to heart disease, cancer, and mental illness.

The basis for the development of any new field such as biomedical engineering, lies in our educational system. Once challenging problems have been identified and areas of opportunity outlined, progress depends upon the availability of manpower. The shortages in biomedical engineering were identified at a conference held at the National Institutes of Health in December, 1961.[2] It is estimated that the number of biomedical engineers with good training and backgrounds in both biology and engineering may be as low as 25 for the entire country. What is more, applications for training grants have numbered only a handful per year. Present National Institutes of Health approved programs will provide six PhD graduates per year. This, of course, does not include the scientists who are currently pooling their several technologies for the solution of specific problems, but ways must be found for these individuals to provide the stimulus and education for graduate students in increasing numbers if the principles of the engineering sciences are to be applied in other than a superficial way to the life sciences.

At present, there is no agreement on how to plan an educational program. This is probably good. Ten universities are operating small programs on an experiment basis. It is interesting that most are located in schools of engineering and not in medical schools or university departments of biology. One principle seems essential to the development of the field. The engineer or the life scientist must be first-rate in his own field before he can hope to be of help in bridging the gap. In most universities, the educational requirements in engineering, medicine, or biology are such that the individual student is forced to devote all his efforts to a single field. Further, the advances have been so rapid and complex in the last decade that after graduation, the expert has difficulty in keeping up with his own speciality. I am concerned that the mixed program now offered by several universities will result in graduates with limited potentials, but as scientists, we should be willing to experiment and develop new approaches to education. Perhaps eventually we may decide on a thorough grounding in the basic sciences for all undergraduates, specialization in a particular discipline at the graduate level, and then additional work at the post-doctorate level in the principles of a new discipline. This sounds like too much time in school, and perhaps it is. The technologies of learning, developed by the behavioral scientists with the help of the engineers, may provide a solution, but not in the immediate future.

The physician is appalled at the complexity of an instrument that can perform a million additions per second, and has a memory of over a hundred thousand ten-digit numbers. The electrical engineer is equally appalled at working with a system such as the human, which is composed of 5×10^{27} atoms, and whose nervous system contains over 14 billion cells, each of which averages more than 200 connections. Each potentially has a good deal to offer the other, and what needs to be done is to find a mechanism, not necessarily to develop a new interdisciplinary science,to provide for each discipline an educational pattern which will insure a working familiarity with the problems, and a first hand feel for the usefulness of the other's technologies.

[1] Rosenblith, Walter A., On Some Social Consequences of Scientific and Technological Changes, Dædalus, ppg 498-513, Summer 1961.

[2] Brown, J. H. U., Resume of the Conference on Biomedical Engineering, National Institutes of Health, Bethesda, Md. December 7-8, 1961.

MAINTAINING THE THERMAL BALANCE IN MAN*

Loren D. Carlson
Department of Physiology
University of Kentucky
Lexington, Kentucky

Summary

Heat production in the human and heat exchange with the environment are characterized with respect to the extent of control and the systems involved in control.

This paper will briefly characterize the heat production of the human and the heat exchange with the environment. The sensors, integrator and output of the temperature control system operate within limits as a rather unique parallel system.

While all tissues of the body metabolize and produce heat, some do so at a higher rate than others and some have a capability of greater change in heat production. Certain systems, a small portion of the mass, are slaves to the entire system being necessary as pumps, ventilators, or serving other logistic functions. The heat produced depends on the tissue and its mass. Aschoff[1] has presented these characteristics schematically as shown in Figure 1. Some areas like the brain have a high maintenance energy cost and heat production but vary little, while muscle may increase to be the major source of heat. While liver and viscera may increase fivefold, the relative mass does not give this area the importance of muscle. We conventionally think of increases in heat production as due to increased muscular activity, but this is not the sole mechanism.

Just as heat production is not central or uniform, the heat exchange of the body is regional due to geometrical as well as control considerations. The body has a heat capacity which it may use as a source or sink, the peripheral tissues participating more actively than central tissues. Again the mass of these tissues determines the magnitude of the change. This has led to the use of the words core and shell. It is interesting that in the human one-half of the mass is within one inch of the surface. This characteristic of the human body is schematized in Figure 2 where isotherms in the body are drawn for two different temperature conditions. Inserted in this figure is a characterization of the system which brings about these changes - the circulatory system which by altering convection allows tissues to cool. The extent of cooling can be increased by using large vessels as heat exchangers.

The extent to which the system uses the mechanisms for exchange is summarized in Figure 3. In the lower portion of the figure the rectal temperature, an average skin temperature and temperatures representative of other areas are shown at various air temperatures. The bars diagrammatically indicate the change in heat content that occurs as the body temperature changes. Thus, a deficit of 200 kilocalories or more can accrue by tissue cooling. The upper graph gives some indication of the apparent change in conductivity due to circulatory changes, and specific measures of blood flow in three tissues are shown. An analogue can be constructed to simulate the main events (Fig. 4).

The overall response of the heat exchange system has recently been characterized by Benzinger.[2,3] His results indicate that the mechanisms operative against increase in body temperature are best correlated with brain (cranial) temperature, thus giving this area a primary importance in control (Fig. 5). Under negative heat load, however, both peripheral and central temperature are elements in the control system both with respect to heat production and circulation. Heat loss is dependent on internal temperature and skin temperature. The intensity of the metabolic response for any given internal temperature is dependent on the skin temperature (Fig. 6).

The control of temperature in the warm-blooded animal is a combined neural and humoral (endocrine) system, but all of the endocrine control is apparently subservient to a neural system. Due to the careful work of Hensel[7,8] the input of this neural control system can be characterized. The technique used in this type of investigation is known as a single fibre analysis and relates the number of impulses on a given fibre to temperature on temperature

* Work was supported in part by NIH Grant RG-8429 and U. S. Air Force Arctic Aeromedical Laboratory.

change. Figure 7 illustrates the technique and Figures 8, 9 and 10 summarize the results of this type of investigation. The impulse rate on any given fibre is related to the initial temperature, each fibre having a range in response (Fig. 9). Different fibres have a different spectrum of response (Fig. 10).

The output of the control system is similar. The calibre of blood vessels changes with the frequency of impulses on the innervating fibres and the amount of muscle activity is related to the frequency of impulses arriving at the neuromuscular junction within limits.

Blood vessels are characterized as resistance and capacitance vessels. In Figure 11 the change in these vessels is shown related to the frequency of impulses. In addition to changes in flow, changes in the volume of blood in the limb also occurs. The main vessels in a limb also serve as heat exchangers. The change in flow with temperature is greatest in the skin.

A different type of recording is shown as an illustration of the electrical activity related to muscle motion. In Figure 12 an area in the caudal pons has been stimulated to produce shivering. The electrical activity in muscle occurs after a delay and persists following the stimulation. Both the delay and the post-stimulatory effects are characteristics of the regulating center not clearly understood.

Our knowledge of the systems involved in maintaining thermal balance is characterized schematically in Figure 13. Sense organs are distributed differently in different areas of the body. The overall effect appears related to the number, the temperature and rate of change of temperature. Counting seems the simplest handling system for these. The stat itself is sensitive to temperature and to the input from the periphery. An activator is a necessary postulate for delays and for the effect of other inputs.

In more anatomical terms (Fig. 14) the temperature regulation center is in the hypothalamus. The functions of loss and conservation are to some extent separate. This area of the brain has many inputs other than temperature. Temperature fibres are traceable to the thalamus. The interconnection to the hypothalamus is certainly present but not clearly delineated. Two separate outputs are indicated, one related to blood vessel calibre and consequently heat transport and heat content; the other related to heat production by muscle action. Studies on acclimation to cold have indicated the presence of a non-shivering thermogenic system under the control of the sympathetic nervous system. In addition, rather marked changes take place in the mitochondrial systems involved in energy production and in the metabolic response to adrenaline and noradrenaline.[5,6]

References

1. Aschoff, J. and R. Wever. (1958) Kern und schale im warmehaushault des menschen. Die Naturwissenschaften 45:477.

2. Benzinger, T. H. (1961) The diminution of thermoregulatory sweating during cold reception at the skin. Proc. Nat. Acad. Sciences 47:1683-1688.

3. Benzinger, T. H., A. W. Pratt and C. Kitzinger. (1961) The thermostatic control of human metabolic heat production. Proc. Nat. Acad. Sciences 47:730-739.

4. Birzis, Lucy and A. Hemingway. (1957) Shivering as a result of brain stimulation. J. Neurophysiol. 20:91-99.

5. Carlson, L. D. (1962) Criteria of physiological responses to cold. Temperature, its measurement and control in science and industry. Vol. III. New York, Reinhold. J. D. Hardy, ed.

6. Carlson, L. D. (1962) Human bioclimatology, cold. Climate, health and disease. Vol. VII. Physical Medicine Library.

7. Hensel, H. (1952) Physiologie der thermoreception. Ergebn Physiol. 47:166-368.

8. Hensel, H. (1955) Uber die function der lorenzinischen ampullen der selachier. Experientia 11:325-332.

9. Melander, S. (1960) Comparative studies on the adrenergic neurohumoral control of resistance and capacitance blood vessels in the cat. Acta Physiol. Scand. 50, Suppl. 176.

10. Witt, I. and H. Hensel. (1959) Afferente impulse ause der extremitaten der katze ber thermischer und mechanischer reezung. Pflug Arch. 288:582-596.

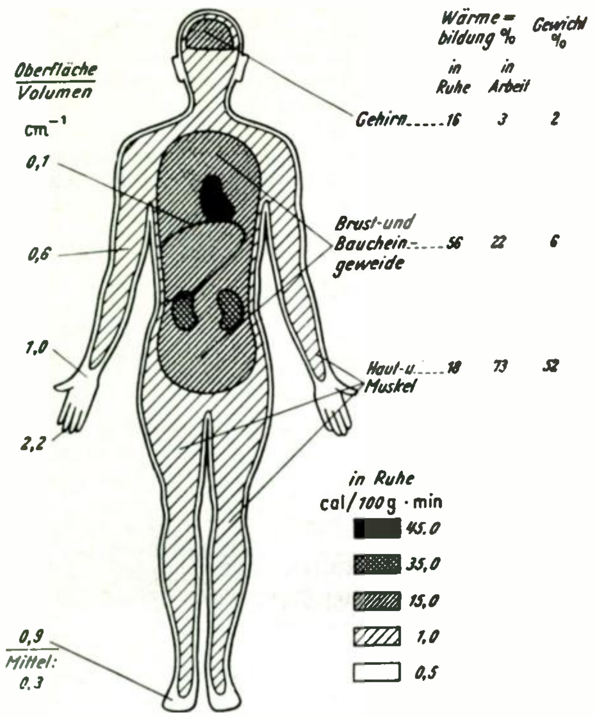


Fig. 1. Topography of heat production in the core and shell in man. From Aschoff.[1]

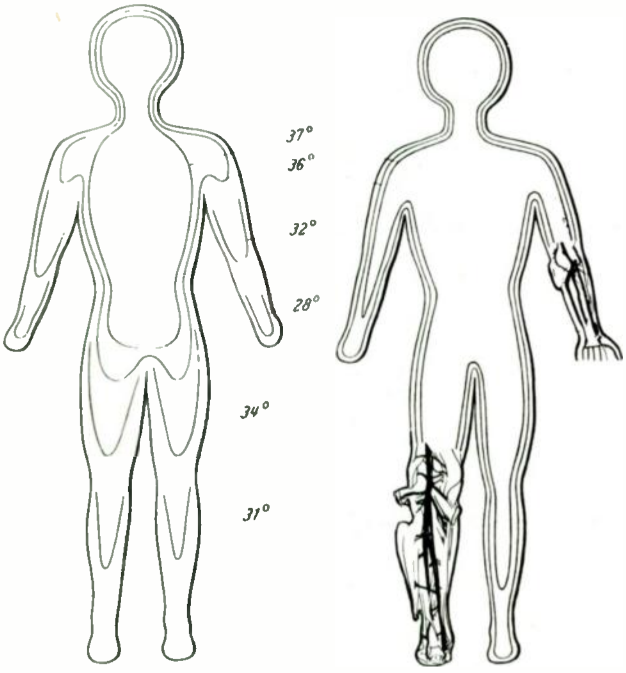


Fig. 2. Schematic isotherms in body cooling. Warm body on the right, cooler on the left. The insert shows the circulatory system in the leg illustrating the remarkable circulatory arrangement for heat exchanging. After Aschoff.[1]

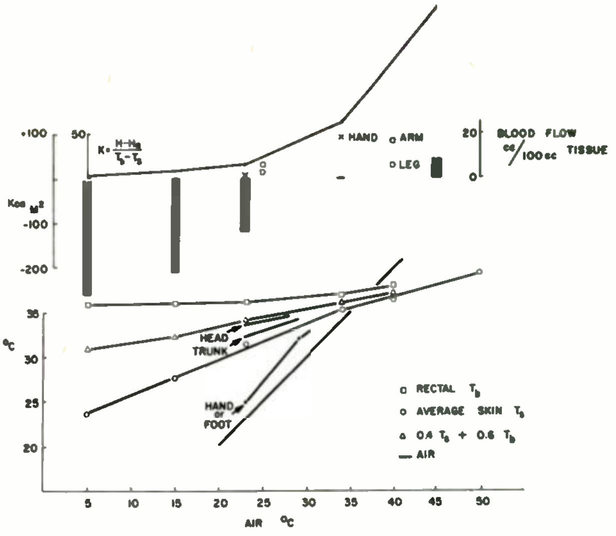


Fig. 3. Representations from various authors of temperature regulation. In the lower figure the change in rectal temperature, average skin temperature and hand or foot temperature are plotted for various air temperatures. In the bar diagram these data are used to calculate the conductivity and the loss or gain of heat from body tissues. Blood flow in the hand, arm and leg are shown in the insert.

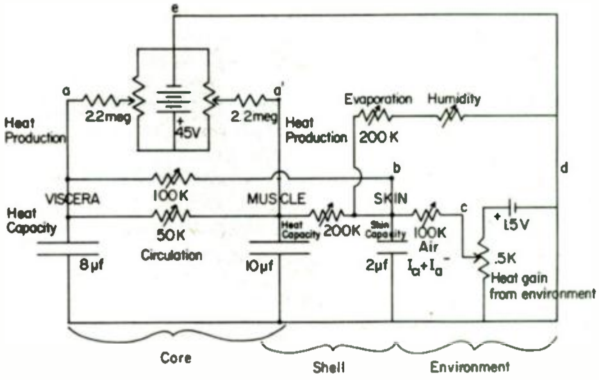


Fig. 4. Electrical analogue of heat production and heat loss to illustrate the major components of the system.

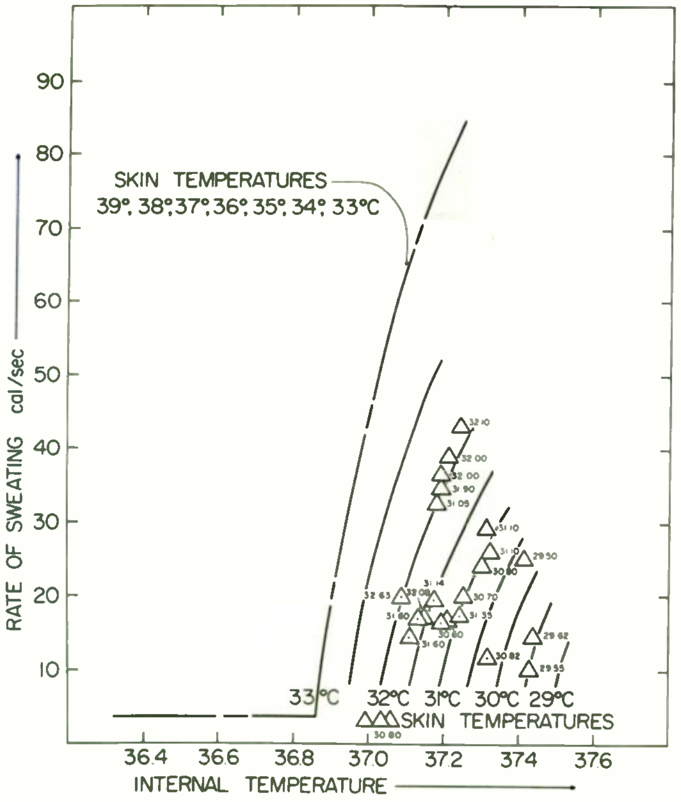


Fig. 5. Intensity of thermoregulatory sweating during cold reception at the skin. Sweating rates were plotted against internal cranial temperatures. Measurements obtained at similar skin temperatures were connected with "best lines." At any given cranial internal temperature, sweating rates are seen to be diminished by approximately 40 cal/sec for every degree C. decrease in level of skin temperature. Figure 5 contains no resting observations as these paradoxical conditions cannot be produced in steady states at rest. Work rates were mechanically equivalent to 6 cal/sec (▲) or 11 cal/sec (△), respectively. Increase in work rate enlarges the range of observations to the right (low skin, with high internal temperature). (Experiments were carried out between April 4 and June 5, 1961 with one subject, D.D., nude, age 26, weight 88.6 kg., height 176 cm.) From Benzinger.[3]

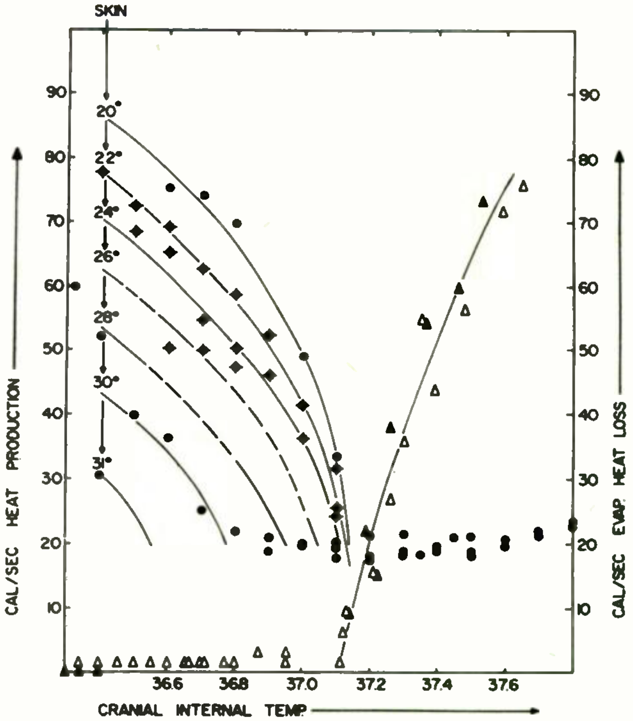


Fig. 6. Experimental Resolution of "Chemical" and "Physical" Temperature Regulation. Thermoregulatory heat production (abscissae, circles, cal/sec) at low and constant cranial internal temperature (ordinates) is determined by steady cold-stimulation of skin (temperatures 31° to 20° C.). Central warm-stimulation (ordinates) leads to depressing counteraction, which becomes complete at individual "setpoint of thermostat," 37.1° C. Thermoregulatory sweating (triangles, abscissae) is uniquely determined by internal sensory warm-reception It begins at setpoint, rises to comparable evaporative loss (cal/sec). Result is tenacious maintenance of setpoint (homeostasis) over fourfold range of production or losses. Note: For narrow temperature range on ordinate, resolving power of classical methods was insufficient. From Benzinger.[2]

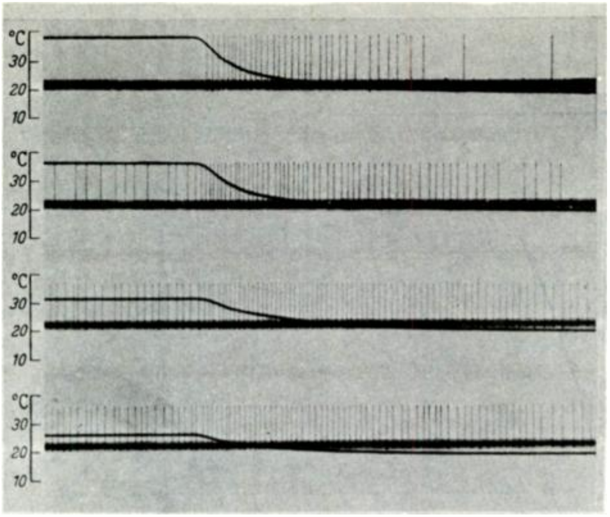


Fig. 7. Change in impulse frequency with cooling of the skin. (Witt and Hensel[10]).

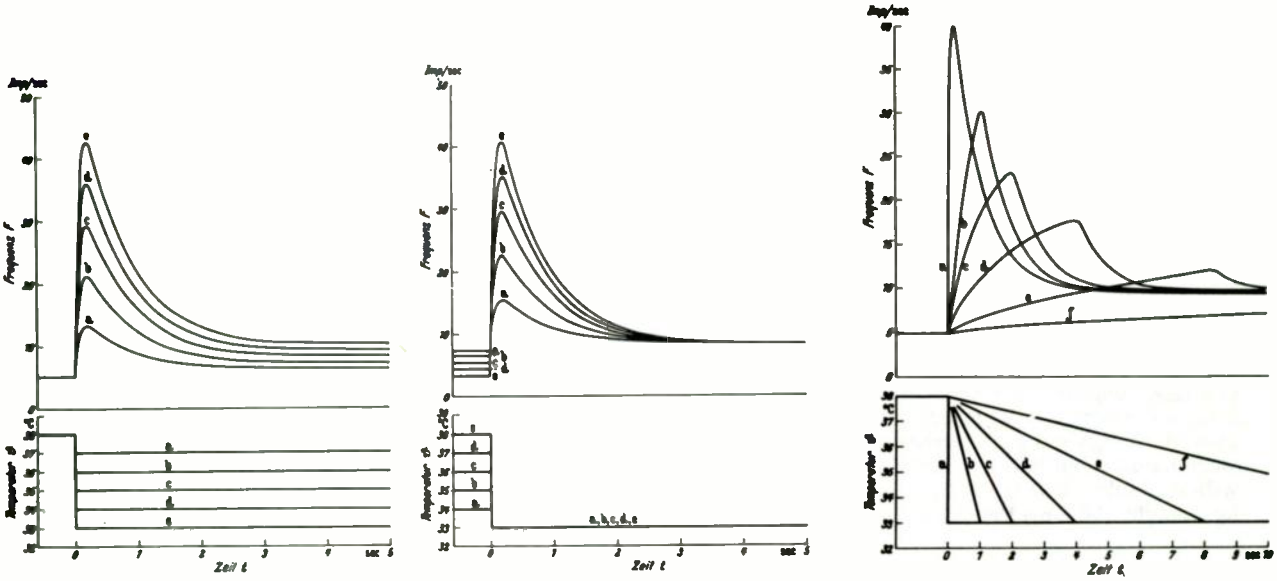


Fig. 8. Relationship of impulse frequency to temperature and rate of change of temperature. (Hensel[7]).

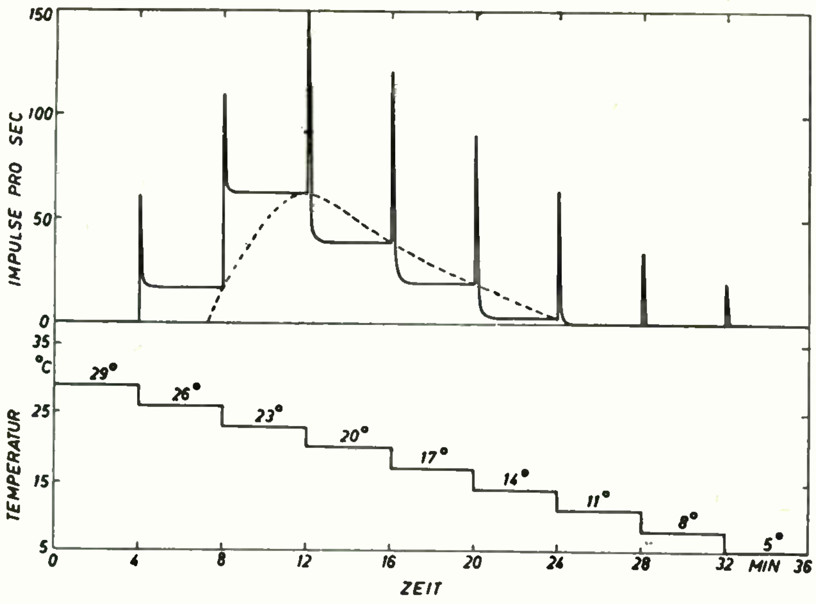


Fig. 9. Impulse frequency of a single fibre with changes in temperature. (Hensel[8]).

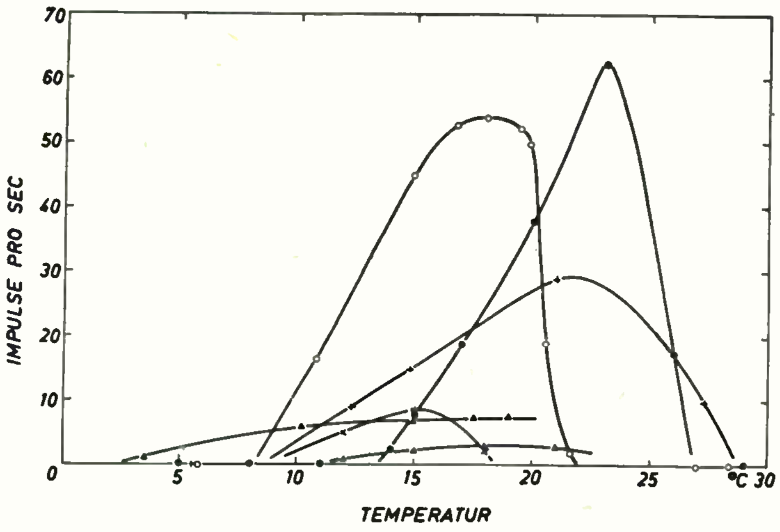


Fig. 10. Stationary impulse frequency of different fibres as a function of temperature. (Hensel[8]).

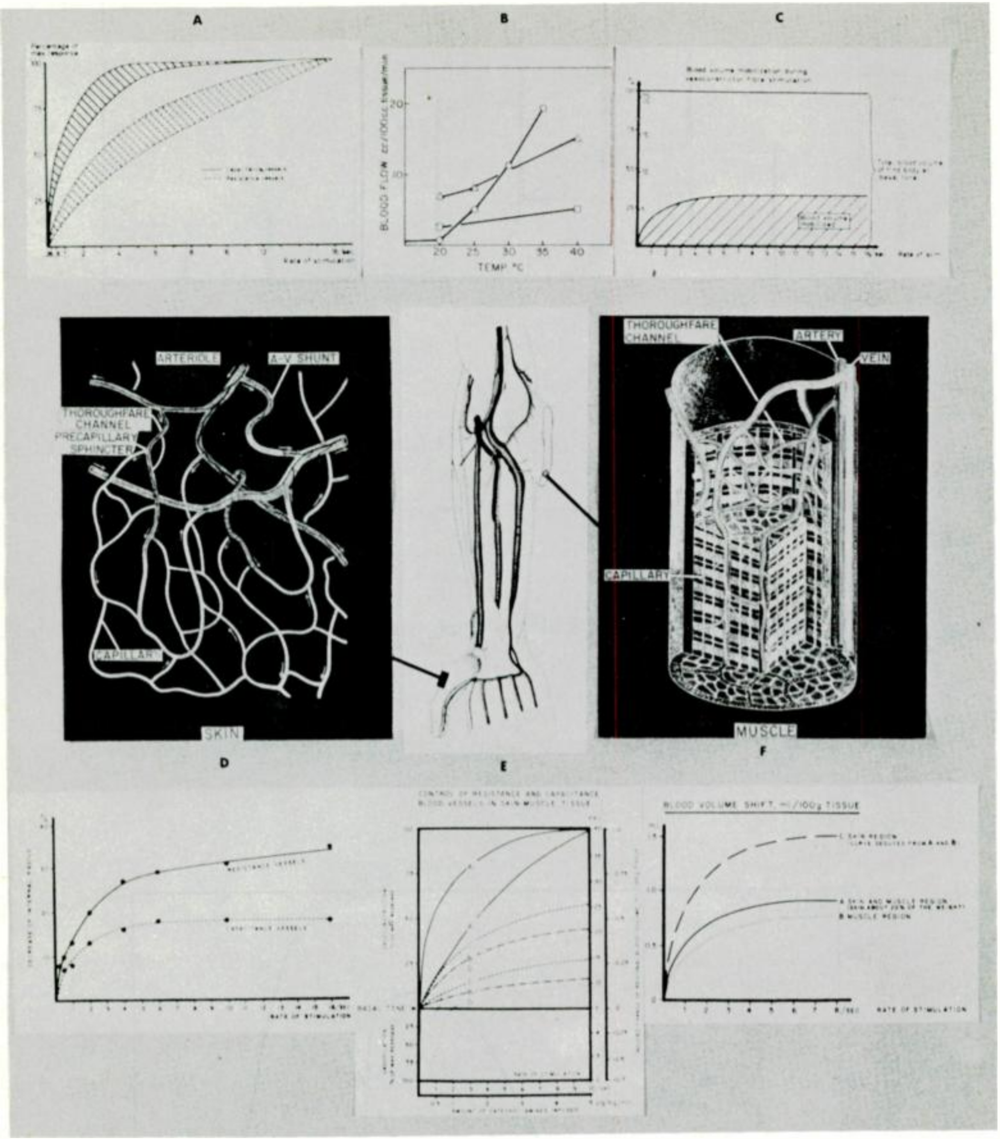


Fig. 11. Response of resistance and capacitance vessels of skin and muscle as a function of rate of stimulation and related to blood volume of the limb. Upper left figure indicates percent of response related to impulse frequency, lower left the decrease in radius. Right-hand figures indicate the blood volume shift. In the lower center figure the neural effect is compared to a humoral effect. Actual blood flows at different temperatures are given in the upper center figure for hand, O, leg, □ and arm, Δ . Capacitance and resistance curves from Melander.[9]

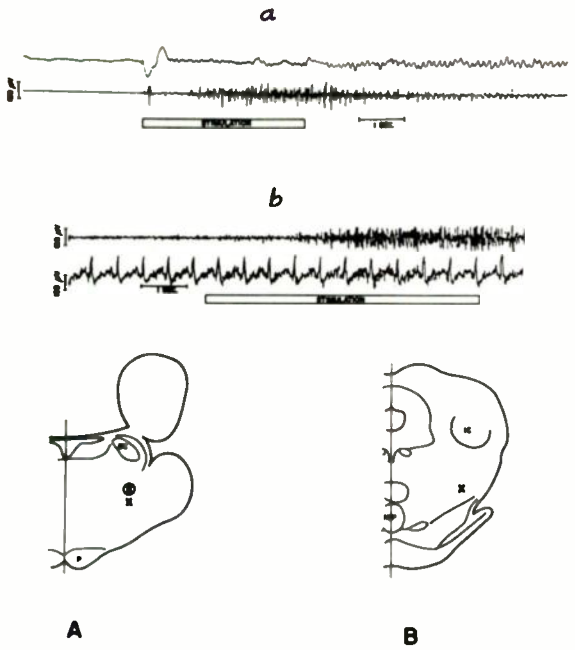


Fig. 12. (A) Cross-section through caudal pons of cat 31, showing stimulation points (X). (B) Cross-section through pons of cat 30, showing stimulation point (X). (a) Top tracing - mechanogram; bottom - electromyogram of shivering obtained during stimulation of encircled X in A. (b) Top tracing - electromyogram from hind leg; bottom - electrical record from intercostals during shivering produced by stimulation of point X in B. (Birzis and Hemingway[4]).

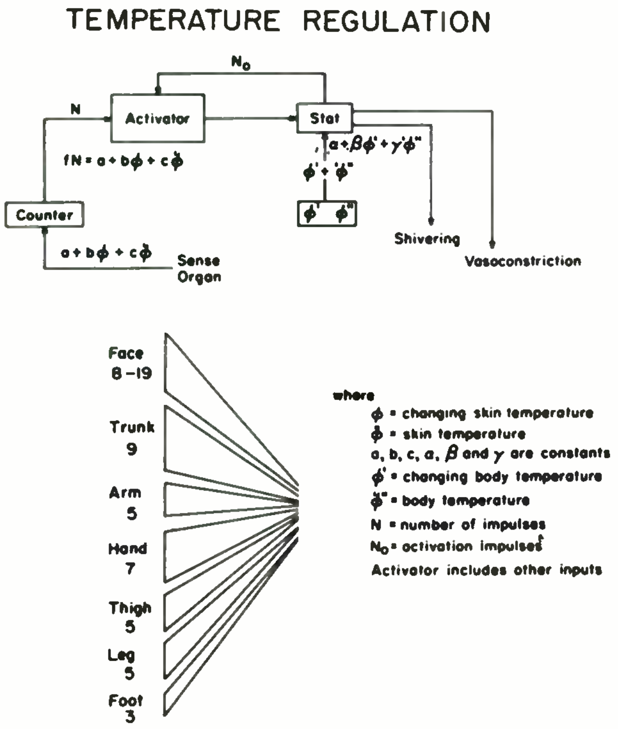
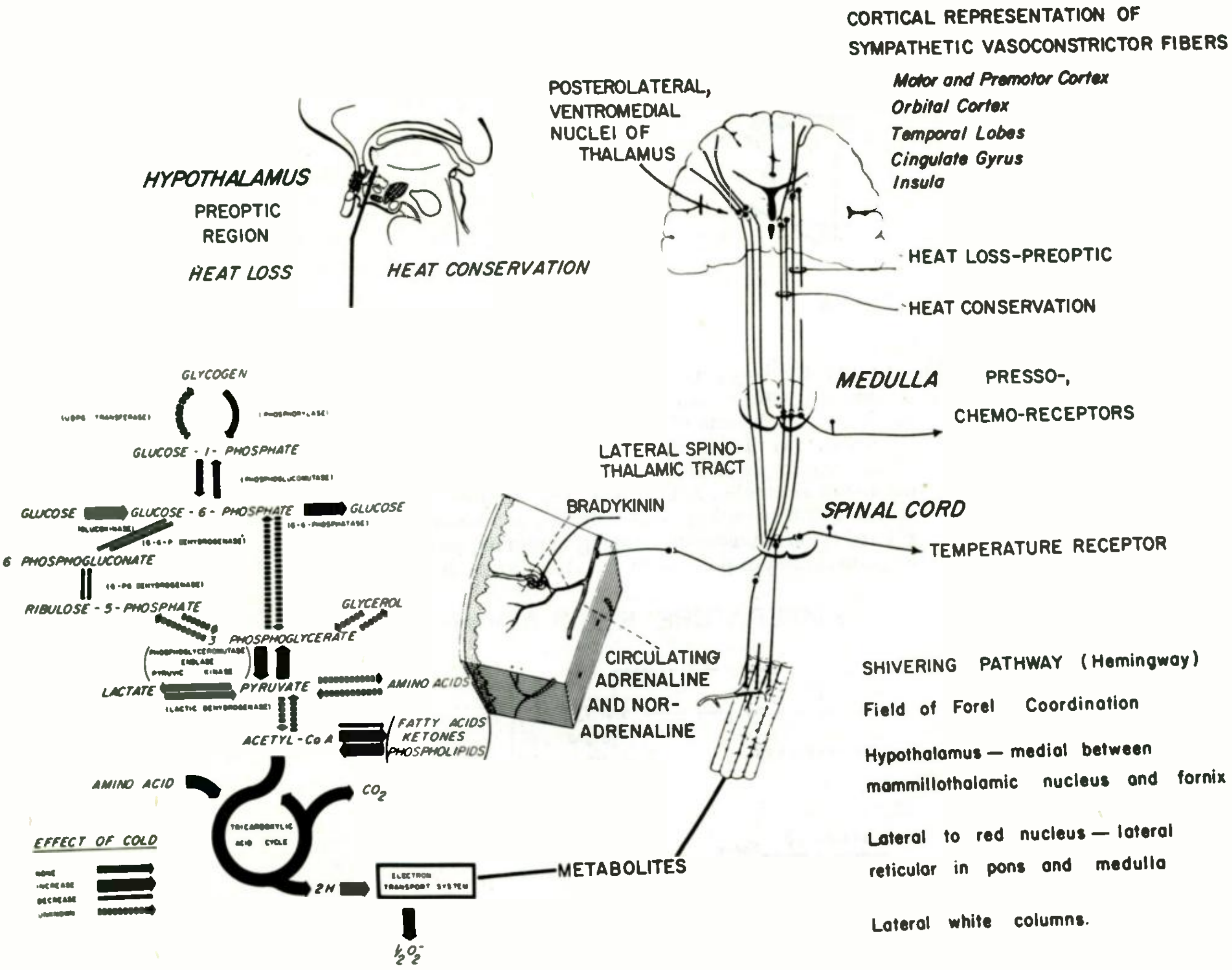

TEMPERATURE REGULATION
N_o
N
Activator
Stat
$fN = a + b\phi + c\mathring{\phi}$
$\alpha + \beta\phi' + \gamma\mathring{\phi}''$
$\phi' + \mathring{\phi}''$
Counter
ϕ' ϕ''
$a + b\phi + c\mathring{\phi}$
Sense Organ
Shivering
Vasoconstriction
Face 8-19
Trunk 9
Arm 5
Hand 7
Thigh 5
Leg 5
Foot 3
where
ϕ = changing skin temperature
$\mathring{\phi}$ = skin temperature
a, b, c, α, β and γ are constants
ϕ' = changing body temperature
$\mathring{\phi}''$ = body temperature
N = number of impulses
N_o = activation impulses
Activator includes other inputs

Fig. 13. Schematic of temperature regulation.



SHIVERING PATHWAY (Hemingway)

Field of Forel Coordination

Hypothalamus — medial between mammillothalamic nucleus and fornix

Lateral to red nucleus — lateral reticular in pons and medulla

Lateral white columns.

Fig. 14. Main neural factors in temperature regulation and the metabolic systems involved in the response.

ENGINEERING PERSPECTIVES IN THE ANALYSIS OF BEHAVIOR*

Bernard Weiss, Ph.D.
Department of Pharmacology and Experimental Therapeutics
and Division of Clinical Pharmacology
The Johns Hopkins University School of Medicine
Baltimore, Maryland

Summary. The behavior of organisms is largely under the control of its consequences. By programming consequences according to a certain plan, which is termed a reinforcement schedule, behavior can be obtained from the organism that is stable and that follows a certain form which arises from the nature of the reinforcement schedule. The problems that are of interest to engineering science have to do with the properties of the schedules themselves, such as inherent periodicities and sequential dependencies, and the description of the resultant behavior as time series with stochastic features or finite Markov processes.

Introduction

Biological scientists have usually looked to the engineering sciences as a source of new instruments. Now they also look to them for new ways of analyzing and describing the systems with which they deal. Behavior scientists may also begin to shift their point of view. During the past decade, the science of behavior has benefited enormously from the development of instrumentation specifically designed to attack problems in the analysis of behavior. It is now getting to a stage at which, as it adopts more advanced instrumentation, it will have to adopt some of the techniques employed by engineers in studying physical systems.

In my talk today, I shall outline a class of phenomena that seem especially well suited to these techniques. This class consists of behavior under the control of what are called reinforcement schedules. Let me begin by first describing the fundamental situation.

Most behavior is strengthened or weakened by its consequences, that is, by what happens as a result of the behavior. The laboratory paradigm of this relation is constructed by placing an organism in an experimental space and then arranging certain states of the environment to follow certain actions by the organism.[1]

With a species such as the rat, the experimental space is arranged to contain a lever and a device for delivering stimuli such as food, heat, water, electrical stimulation to the brain, etc. Figure 1 shows an arrangement used in our laboratory with which we can deliver either bursts of heat or food.

By appropriate circuitry, depression of the lever can be made to deliver one of the class of stimuli enumerated above. If, for example, depression of the lever is arranged to eject a pellet of food into a food tray placed in the chamber, the frequency with which the rat presses the lever is greatly increased over the base level. Depressions of the lever are called responses. The consequent stimuli are called reinforcements. The basic paradigm for manufacturing behavior is to deliver reinforcements after the emission of the behavior that is arbitrarily defined as a response by the experimenter.

The reinforcement process, then, may be looked upon as one which enhances a particular state of the system. One of the most interesting features of behavior developed in this way is the fact that it is not necessary to reinforce the behavior each time that it appears. Behavior can be reinforced intermittently and still remain in strength. The plan that designates when these reinforcing events appear is called a reinforcement schedule.[2]

The property of reinforcement schedules that makes them particularly suitable for treatment by engineering methods is that they generate stationary processes. Behavior scientists have paid much less attention to this property than they have to learning phenomena, which are not stationary. I do not think, however, that anyone can make an effective argument for the view that the factors that maintain behavior are basically different from those that alter it; both sets of factors consist fundamentally of the processes associated with reinforcement. Our understanding of behavior might be advanced more by its study as a relatively orderly process than as a relatively disorderly one. I would like to consider, then, how the study of behavior generated by reinforcement schedules might benefit from an approach that an engineering scientist might take.

Schedules with Periodic Components

On certain schedules, the reinforcement is programmed to appear a constant length of time following the previous reinforcement. This is termed a fixed-interval schedule. Suppose that at time $\underline{t}$ a response is followed by reinforcement. The probability of a response producing a reinforcement between time $\underline{t}$ and time $\underline{t}+n$ (where n = seconds) is zero. At time $\underline{t}+n$ the

*The preparation of this paper was supported by grants MY-3229 from the National Institute of Mental Health and B-865 from the National Institute of Neurological Diseases and Blindness, National Institutes of Health.

probability that a response will produce reinforcement suddenly jumps to 1. Immediately after a response has produced the reinforcement, the probability falls back to zero. The same sequence of events occurs at $\underline{t}$+2n, $\underline{t}$+3n, etc. The pattern of behavior that finally emerges from prolonged exposure to such a set of conditions possesses the following properties. Directly after a reinforcement has been delivered, the likelihood that a response will be emitted is very low. As the end of the interval approaches, the rate at which responses are emitted typically grows higher and higher. This pattern of behavior is often equated with a time discrimination.

Although such behavior is relatively constant from interval to interval and has been produced in numerous laboratories, a precise understanding of it is still lacking. One reason is that our analytical methods may be somewhat primitive. Various attempts have been made to describe the properties of the temporal distribution of responses on such a schedule. One method is to measure the quarter life of the interval, that is, how long it takes the animal to emit one quarter of the total responses made during the interval.[3] Another method has been to divide the interval into segments and then plot a straight line according to the number of responses in each segment.[4] The slope of the line reflects the distribution of responses within the interval. A third technique expresses curvature as a deviation from a constant rate of responding through the interval.[5]

None of these adequately describe the serial properties of the data. For example, it could be proposed that as the interval proceeds, succeeding interresponse durations grow shorter and shorter. One way of analyzing fixed-interval behavior might then be to compare the actual data with this hypothesis. Thus, given a set of interresponse times, IRT_1, IRT_2,.....,IRT_n, where IRT_i gives the ordinal position during the interval, we would expect the size of the interresponse time to show an inverse relation with ordinal position. Deviations from this relation would indicate lack of control by the temporal properties of the schedule. Another test of the control exerted by the schedule would be to see whether, given a sequence of $\underline{n}$ interresponse times, it is possible to predict the sequence of the next $\underline{n}$ interresponse times within a specified error. Such an analysis might enable us to determine whether the animal was changing its behavior in accordance with a derivative or integral function; that is, whether the controlling variable is the rate of change in interresponse time or the number of responses emitted since a particular time.

A fixed-interval reinforcement schedule is not the only way of trying to observe the sensitivity of the organism to periodic fluctuations in his environment; this probability could also be made to vary in a continuous fashion. For instance, one could arrange the programming circuitry to impose a sine function on the probability or density of reinforcement, letting the baseline of the function be displaced to eliminate negative values.

There are some interesting psychological problems that such an approach would bear on. Intermittent or partial reinforcement is still a somewhat unclear phenomenon. Why, for instance, should an organism continue to respond between reinforcements on a fixed-interval schedule? One answer to this question is that the organism is really not emitting single isolated responses, but, instead, is emitting behavior in a certain serial pattern. The degree to which patterning can be observed on a periodic reinforcement schedule of the kind just described should be able to tell us over how great a period of time such patterning will be able to affect the organism's behavior. We might imagine that if the periodic function that shifts the density of reinforcement has a total period of, say, 5 minutes, it will be easier to see a temporal patterning of behavior than if this function has a period of 7 or 8 hours. The question of how long we can extend this period and still perceive a corresponding periodicity in the behavior is relevant from the standpoint of an organism's capacity to discriminate the temporal properties of its environment.

Such an approach lends itself to analysis by well known mathematical and engineering methods such as autocorrelation, cross-correlation, and power spectral density functions. These would allow us to describe the fidelity with which the animal's output of behavior followed the schedule. Similar analyses have been applied to other aspects of the serial properties of behavior such as the description of periodicity in the spontaneous activity of animals, which is an example of what physiologists call circadian rhythms.[6] These techniques have also been applied to the description of the human operator as a control system. For example, in making judgments of angular excursion, human subjects behave as a first-order linear autoregressive system; that is, for response X_{t+1},

$$X_{t+1} = aX_t + \epsilon_{t+1} \qquad (1)$$

where $\underline{a}$ is the first order coefficient and ϵ is random error. Such behavior fits the Markov process model, meaning that the past history of the judgments, except for the just preceding judgment, does not affect the judgment made on a particular trial.[7] This technique, it should be noted, allows one to make a significance test of the order (i.e., serial dependency) of the system.

If reinforcements appear at irregular intervals of time, the schedule is termed a variable-interval schedule. In constructing such a schedule, most experimenters use a distribution of long, short, and medium intervals in which the relative frequencies are adjusted to give a steady rate of responding.

Although a combination of interreinforcement intervals may be formed that gives excellent control over behavior, it tells us little about the system emitting the behavior because it does not allow more than a gross description of the resultant behavior. A more fruitful analysis might result from specifying such a schedule as a complex waveform with known component frequencies in which the density of reinforcement varies with the amplitude of the wave, as with the periodic schedules described earlier. Such a waveform could be constructed by combining a base frequency with several harmonics. Or, it could be specified as a periodic function with a stochastic component such as Gaussian noise. By describing a schedule in this way, it should be possible to analyze the component periodicities in the behavior and to determine the extent to which these reflect the properties of the schedule.

To summarize, the distribution of reinforcements on time-parameter schedules should be described as time series with clearly stated properties that make possible a time series analysis of the resultant behavior.

Rate-dependent Schedules

On the reinforcement schedules described so far, reinforcements are programmed according to temporal parameters. There are other schedules on which the frequency of reinforcement depends only on the frequency of responding. These are called ratio schedules. They impose the requirement that a certain number of responses be made to produce the reinforcement. If the requirement is invariant, it is called a fixed-ratio schedule. If the requirement varies, it is called a variable-ratio schedule.

Certain features of fixed-ratio schedules are particularly interesting. One is that if the ratio is large, or the reinforcement small, regular periods of no responding appear. When it does respond, the animal tends to do so at a high and steady rate; under the above conditions, extended pauses appear, usually early in the ratio. The reason for this is that early in the ratio the probability of a reinforcement is zero, and the discrimination by the animal of this condition is aided by recency of the reinforcing stimulus.

Another way of constructing a rate-dependent schedule is to vary the probability of each response delivering a reinforcement. Thus, the equivalent of a 100:1 ratio schedule is a reinforcement probability of .01 associated with each response. One might expect, if successive probabilities were independent, that such a schedule would result in a rather steady rate of responding with the characteristics that successive interresponse times were also largely independent. What would happen if we progressively increased the degree of serial dependence in the program? Would we then increase the degree of serial dependence in the behavior? Such a question might be thought of as analogous to the question that Shannon and others have put to language; namely, how does the next word used in a sequence of words vary as a function of the serial dependencies of the previous words in the sequence?

One approach in the present instance might be the following. Arrange the schedule to specify a particular probability of reinforcement given a particular rate of responding over the last _n_ responses. That is, for any single response, n+1, the probability of reinforcement,

$$P_{n+1} = k\left[\frac{1}{n}\sum_{i=1}^{n}\tau_i\right] \qquad (2)$$

where τ_i represents a single interresponse time. Such a schedule, which specifically reinforces a rate of responding, is likely to demonstrate serial dependencies of greater magnitude than the schedule which designates independence in the reinforcement properties of successive responses. One could ask, moreover, because the question can be made explicit, to what extent the properties of ordinary fixed ratio schedules depend on the extent to which they reinforce a rate of responding rather than a fixed number of responses.

The Description of Changes in State

Certain properties of reinforcement schedules may be devised in which the resulting behavior is described not by interresponse times alone but also by the sequential character of the behavior. An experiment specifically calling for such an approach was performed by Frick and Miller in 1951.[8] They devised an apparatus with a bar at one end and a feeder at the other, the path to which crossed a photocell beam. The object of this design was to secure data that would allow an analysis of the sequential properties of the behavior. The behavior was so encoded that one symbol, A, represented approaches to the food tray, and another, B, represented a bar press. Sequences up to four units in length were analyzed to obtain the average "uncertainty" for different orders of serial dependency. (Their measure of uncertainty is equivalent to the information measure, H, which is called entropy in communication theory.) By plotting the uncertainty for different sequence lengths, it was possible to determine that a sequence longer than two responses did not add a significant increment in the predictability of the next response. This was confirmed by the autocorrelation functions, which also reflected the sequential properties of the shifts. Before the introduction of reinforcement for lever pressing, A's tended to follow A's and B's tended to follow B's. When the bar was connected to the feeder, the pattern shifted to one in which the A's and B's tended to alternate, and the autocorrelation plot displayed clearly defined oscillations.

An analogous situation is one used in our laboratory. A rat living in the cold is permitted to obtain both heat and food by pressing appropriate levers. On one program that we have arranged, the number of presses required for a pellet of food increases by a fixed number after each reinforcement. (In Figure 2, the number is 11.) The total number of steps is 22. Each time that the rat allows more than a certain interval to elapse without a response, the ratio decreases by one step. (In Figure 2, this interval is 30 seconds). The records that we have obtained so far suggest that as the ratio becomes greater and greater, the rat spends more and more time between reinforcements working for heat. Again, a description of the transition probabilities is necessary for an adequate account of the behavior.

A sequential analysis could also be applied to a situation used by Mechner[9] in which the experimental chamber contains two bars. In one variation of this arrangement, _n_ presses on bar A close a circuit that permits a press on bar B to actuate a feeder. This situation also lends itself to an analysis compatible with the techniques of communication theory. One assigns, for a fixed ratio of _n_ responses on bar A, the designations $a_1, a_2, \ldots, a_n, a_{n+1}, \ldots, a_{n+m}$, to sequences on bar A completed before a response on bar B. If, for example, the behavior is a simple Markov chain, then the length of the sequence that represents a particular state depends only on the length of the previous sequence. Mechner's graphs of the lengths of successive runs suggest that the dependence is higher than first-order; definite periodicities are apparent and lumped chains would probably be necessary in order to treat the data as a finite Markov process.

Findley[10] has recently reported a series of studies that are also amenable to such an analysis. After completing a state represented by a segment of a reinforcing schedule such as a single fixed interval, the animal is given a choice of completing, say, either a fixed ratio segment of a certain size or perhaps another fixed interval segment. The animal has previously been exposed to the consequences of completing these states, which are designated by stimuli such as lights above the appropriate lever.

The entire system could be described as a finite state device with one absorbing state (reinforcement). The transition probabilities appear to depend on previous states, according to Findley. An analysis of this situation as a finite Markov process may prove rewarding.

Adaptive Schedules

Up to now I have been talking mainly about reinforcement schedules in which the contingencies are pre-set. That is, nothing that the animal does affects the schedule. There are other types of schedules in which this is not the case. These have been called adjusting, titration, and conjugate schedules. Their basic principle is that a function or value is varied, in accordance with some rule, by what the subject does.[11,12,13]

In our laboratory such a schedule has been adapted to the study of pain tolerance in rats, monkeys, and humans.[14] The subject receives an electrical signal whose intensity rises in steps. Figure 3 shows a monkey, in a restraining chair, whose feet are strapped into an electrode "shoe" through which we deliver the stimulus. By pressing the lever, the subject brings the current down. We have been interested in the level of tolerance to this electrical stimulus as a function of various parameters of the situation. We can vary the following: the interval between increments, the size of the incremental step, the size of the decremental step, and the number of responses required to produce a single decrement in the intensity of the stimulus. All of these factors are important determinants of the level at which the subject will maintain the stimulus. Moreover, they also help to determine the kind of variation we observe on the record.

The records in Figure 4 show what appear to be periodicities or fluctuations in the subject's behavior. These are due to the following. The subject allows the stimulus to rise in intensity by a certain amount before reacting. Then he emits a series of responses which brings the stimulus down below the point at which it ordinarily fails to evoke a response. The shock is then allowed to rise again until it calls forth another series of responses. The response record shows clearly that this behavior tends to occur in bursts. One of the questions that it seems of interest to examine here would be the periodicities inherent in the animal's behavior as a function of the parameters of the experimental situation. It seems fairly certain that one reason the subject engages in such behavior is that a single response which brings the step down by a very small amount is not reinforcing simply because it cannot be discriminated. Only long series of responses, as in a fixed-ratio schedule, are reinforcing.

The titration type of experiment offers several advantages to the experimenter. For one thing, it is more efficient. Adaptive experiments of this sort have a non-zero probability of being shorter than preset experiments designed for the same problem. However, they also offer some new problems in the analysis of data. In the typical operant conditioning experiment, we can get a great deal of information simply by counting responses made by the subject under various conditions. In the titration type of experiment, this is not enough. For example, in tracing out tolerance to pain, the subject with an elevated threshold makes the same number of responses, if he keeps his threshold at a steady level, as the subject whose threshold is low. One way in which our laboratory has handled this problem is to convert the amplitude of the shock signal to a pulse rate which then allows us to integrate the shock. But this is not sufficient to characterize some of the most interesting aspects of the record. Here is one problem on which the help of the engineer might be particularly valuable because the titration schedule, at least in certain respects, may be viewed as a process control system with an input that consists of the shock, a network which consists of the subject, an output which consists of pulses, and a controlled signal, the intensity of the stimulus, which is fed back to the network.

Some of the following considerations seem pertinent from this viewpoint.

1. At present, the system is a proportional controller; the magnitude of control action, which is the output rate of the subject, is a function of the shock intensity. The higher the shock, the greater the rate.

2. One apparent feature of the system is the tendency for the size of the displacement from zero to vary as a function of the rate of rise of the shock. The greater the rate, the greater the displacement.

3. Given these considerations, could one insert parameters into the system so that displacement does not vary with the interval between increments? One possibility is the addition of derivative control. This would be accomplished by making the size of the step that decreased the shock a function of the rate of response; the greater the rate, the greater the size. One might achieve an even more stable result by varying step size in accordance with the time integral of shock.

4. The frequency characteristics of the output could be analyzed in two ways. One is by the interresponse times of the decrements. The other is by an analysis of the amplitude variation of the record by autocorrelation, cross-correlation, and power spectral density functions.

Analytical Methods and Computers

As I said before, the aspect of reinforcement schedules that makes them particularly attractive for analysis by engineering methods is that the resulting behavior is characterized by stationarity. That is, given a suitably long segment, the statistical properties of the behavior do not vary as a function of time. This means that the analytical apparatus of communication theory and related techniques can be brought to bear on the system and afford a much more precise description of it than is possible under other circumstances.

These techniques require access to digital computers and the collection of data in a form compatible with computer analysis. At this stage of technology, it might also be asked what contributions to the study of behavior could be made by computers apart from their use in the analysis of data. I believe that an on-line digital computer of moderate size could make important contributions. For example, it could vary the contingencies of reinforcement as a function of its input history. In this way, it could simulate social reinforcement, since the computer could store up the history of the environment just as a person does. In the laboratory, our reinforcing circuits have no memory. In many of the situations that we encounter in everyday life, we are not only faced with an environment that has a memory, usually in the form of another person, but with an environment in which the reinforcement contingencies are basically stochastic processes. A digital computer can simulate this kind of situation.

An on-line computer would also be useful in adaptive schedules. Again the computer has the advantage that it is easily arranged to act as a sort of proportioning controller which takes account of the general trend of the behavior and not simply the behavior at a particular time, unrelated to the history of that state. Perhaps certain local variations in rate of responding on various schedules that cannot be explained are due to the fact that not enough of the behavioral history is taken account of in arranging the consequences. By incorporating a segment of the recent behavior into the selection of contingencies, we may be able to erase these local variations in rate. If this is the case, then on-line computers could reveal some facts about behavior that at the present time remain inaccessible to us.

Concluding Remarks

I have tried, drawing on my limited knowledge of methods used by the engineering sciences, to present some problems to which these methods might profitably be applied. By no means have I exhausted the possibilities for such applications in the study of behavior, and I am certain that I have listed only a fraction of the possible contributions that engineers could make.

The behavioral paradigm that I believe is most amenable to such approaches is behavior under the control of a reinforcement schedule that generates stationary processes. Most of the mathematical techniques that psychologists have been concerned with apply to the acquisition of behavior.[15,16] The fact that acquisition (or learning) is a transient process often makes the mathematical apparatus unwieldy and complex. Perhaps a more intensive study of stationary behavior, with simpler mathematical tools, could lead to the development of new techniques for describing learning.

Moreover, the study of stationary behavior by the engineer might make it simpler for him to match the behavior of living organisms to models generated by theories such as those pertaining to finite automata.[17] It seems to me that it is much easier to represent, with a logical net, a system whose behavior is relatively constant than one in which multiple parameters are varying

References

[1] B. F. Skinner, "Science and Human Behavior," Macmillan, New York, N. Y.; 1953.

[2] C. B. Ferster and B. F. Skinner, "Schedules of Reinforcement," Appleton-Century-Crofts, New York, N. Y.; 1957.

[3] R. J. Herrnstein and W. H. Morse, "Effects of pentobarbital on intermittently reinforced behavior," *Science*, vol. 125, pp. 929-931; 1957.

[4] B. Weiss and E. W. Moore, "Drive level as a factor in distribution of responses in fixed-interval reinforcement," *J. Exper. Psychol.*, vol. 52, pp. 82-84; 1956.

[5] W. Fry, R. T. Kelleher and L. Cook, "A mathematical index of performance on fixed-interval schedules of reinforcement," *J. Exper. Anal. Behav.*, vol. 3, pp. 193-199; 1960.

[6] Cold Spring Symposia on Quantitative Biology, Vol. 25, "Biological Clocks," Biological Laboratory, Cold Spring Harbor, L. I., N. Y.; 1960.

[7] B. Weiss, P. D. Coleman, and R. F. Green, "A stochastic model for time-ordered dependencies in continuous scale repetitive judgments," *J. Exper. Psychol.*, vol. 50, pp. 237-243; 1955.

[8] F. C. Frick and G. A. Miller, "A statistical description of operant conditioning," *Amer. J. Psychol.*, vol. 64, pp. 20-36; 1951.

[9] F. Mechner, "Sequential dependencies of the lengths of consecutive response runs," *J. Exper. Anal. Behav.*, vol. 1, pp. 229-233; 1958.

[10] J. Findley, "An experimental outline for building and exploring multi-operant behavior repertoires," Laboratory of Psychopharmacology, Univ. of Maryland, Technical Report No. 61-66; 1961.

[11] G. v. Békésy, "A new audiometer," *Acta Oto-Laryngol.*, vol. 35, pp. 411-422; 1947.

[12] D. S. Blough, "A method for obtaining psychophysical thesholds from the pigeon," *J. Exper. Anal. Behav.*, vol. 1, pp. 31-43; 1958.

[13] O. R. Lindsley, J. H. Hobika, and B. E. Etsten, "Operant behavior during anesthesia recovery; A continuous and objective method," *Anesthesiology*, vol. 22, pp. 937-946; 1961.

[14] B. Weiss and V. G. Laties, "Titration behavior on various fractional escape schedules," *J. Exper. Anal. Behav.*, vol. 2, pp. 227-248; 1959.

[15] R. R. Bush and F. Mosteller, "Stochastic Models for Learning," John Wiley & Sons, Inc., New York, N. Y.; 1955.

[16] W. K. Estes, "The statistical approach to learning theory," in "Psychology: A Study of a Science," Vol. 2, McGraw-Hill Book Co., New York, N. Y., pp. 380-491; 1959.

[17] R. McNaughton, "The theory of automata, a survey," in "Advances in Computers," Vol. 2, Academic Press, New York, N. Y.; 1961.

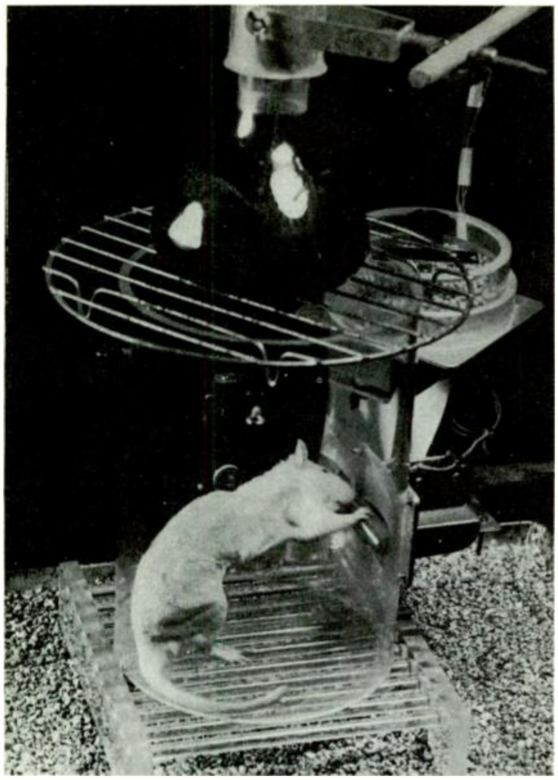


Fig. 1. An experimental chamber constructed so that depression of one lever delivers a pellet of food and depression of the other lever delivers a burst of heat from the heat lamp above the chamber.

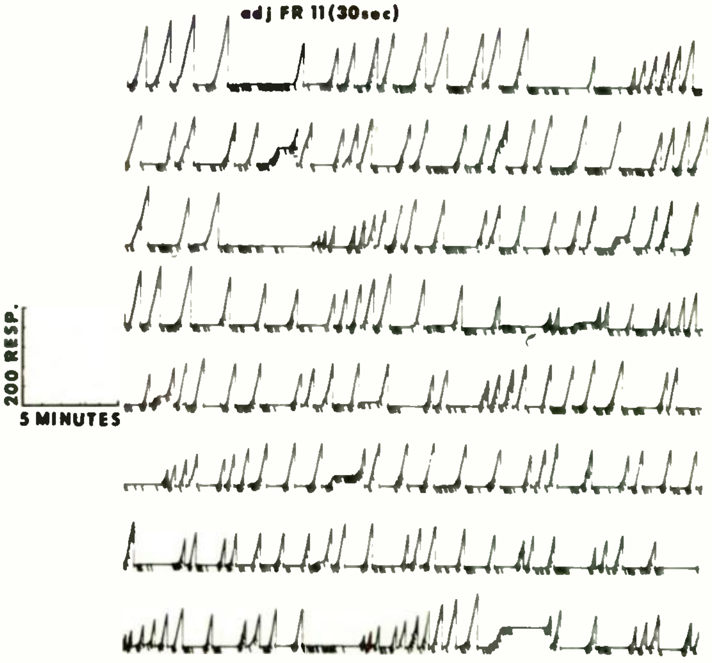


Fig. 2. Behavior of a rat in a chamber in which presses on one lever deliver heat and presses on another lever deliver food. The responses on the food lever were recorded cumulatively, the pen resetting to the baseline after each food reinforcement. The number of lever presses required for food rose by 11 after each reinforcement and decreased by the same number after each 30 second period without a response. Each heat reinforcement (a duration of two seconds with the lamp output set to 375 watts) made an oblique slash mark on the record. Each segment represents 30 minutes and the progression of the session in the figure is downwards.

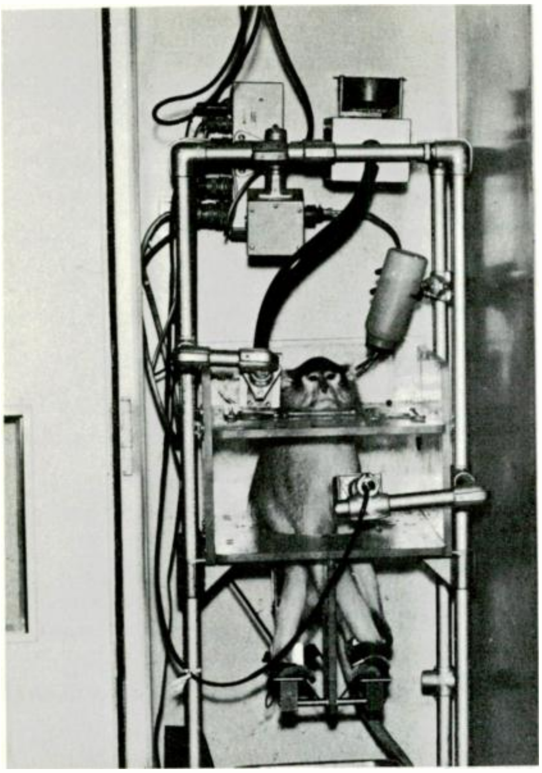


Fig. 3. A monkey (Erythrocebus patas) in a restraining chair arranged for pain titration schedules. The monkey's feet are secured to bent aluminum strips which serve as the electrodes. The lever is enclosed in a box in front of the monkey.

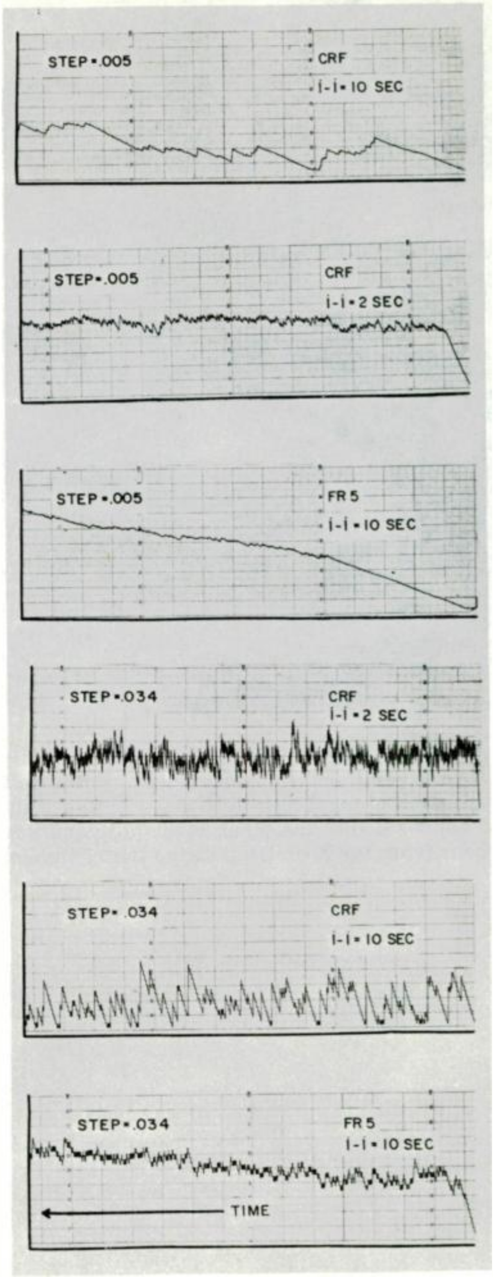


Fig. 4. Records of a monkey on a titration schedule. Each segment represents a 1-hour session, time reading from right to left. The i-i interval refers to the interval between increments. The step sizes are given as the fraction of the total range, from zero to maximum shock, represented by each step. CRF (continuous reinforcement) means that each response produced a decrement. FR-5 (fixed-ratio of 5) means that 5 responses were required to produce a decrement. The size of the decremental step was equal to the size of the incremental step.

BIOLOGICAL OCEANOGRAPHY

Carl N. Shuster, Jr.
Department of Biological Sciences
University of Delaware
Newark, Delaware

"Knowledge of the oceans is more than a matter of curiosity. Our very survival may hinge on it."
President John F. Kennedy

Quotation from a special message to Congress in March, 1961.

Summary

This paper is a biologist's synopsis of oceanography. It also points out some uses of instruments, since instrumentation is playing an important role in oceanographic research; future utilization of instruments will be intensified. For example, instruments can be used to increase our ability to obtain, process, and interpret synoptic data. These data, fed back into computers, can be used to predict oceanic events.

Biologists are allied with other specialists in the scientific exploration of the world ocean, documenting, among many other endeavors, information about the multitude of natural processes occurring in the ocean; processes which now and in the past have been associated with the evolution of the Earth and life upon it. This alliance of researchers is largely due to the complexity of oceanic conditions, the world-wide distribution of organisms, and the magnitude of natural processes in the ocean. Some of these processes include: energy flow and transformation, as in the one-way flow and step-wise decrease of energy through a biological community; the recirculation of inorganics; and, the dynamics of other environmental conditions such as temperature changes and sound propagation

Research in oceanography is providing a fund of knowledge that will become increasingly important as it is applied to some of man's problems, such as: reducing pollution of coastal waters; increasing the sea food harvest; prospecting and mining underwater mineral deposits; improvement of weather forecasting; prevention of shoreline erosion; conversion of salt water to fresh, and national defense. Perhaps the most challenging of the problems concern radioactive waste disposal and radioactive fallout in the ocean.

The Scope of Oceanography

Oceanography deals most directly with the distribution of matter and energy, in both time and space, within the world ocean. This statement, however, does not entirely define the scope of oceanographic research. Understanding more about the oceans and oceanic life demands that conditions existing in the three dimensions of space and the fourth dimension, time, be studied and analyzed further than just the physical boundaries of bodies of water. This demand has imbued oceanography with a uniqueness shared by a few other fields of research; a uniqueness that even the non-oceanographer can sense: it is the oneness or interrelatedness of natural processes which are now occurring or have occurred, not only in the ocean, but in and between other major components of the Earth. This interrelatedness makes it necessary for oceanographers to consider the processes involved in two or more of the Earth's components in order to gain a fuller understanding of energy transfer and transformation of matter in the ocean. The major components or subdivisions of the total Earth system, which are schematically shown in Figure 1, ultimately must be considered in context with the evolution of the universe, since even such common energy sources as sunlight and the gravitational pull of the sun and moon upon the air, land, and ocean emanate from outside our Earth.

The Cubic Approach. One way in which the interactions among the processes and phenomena related to the major components of the Earth can be studied is in a cubic section of the Earth's surface. Such a section is visualized in Figure 2. Two sets of processes can be analyzed: 1) the flow of energy into and out of the cube, through each of its six faces, and 2) the transformation of energy within the cube. Some aspects of this analysis, as in the case of community metabolism, are treated in subsequent sections of this paper.

Contribution No. 21, University of Delaware Marine Laboratories.

Biological Oceanography

Biological oceanographers study the biology, principally the ecology of marine organisms. To do this they spend considerable time at sea and in the laboratory observing, recording, and comparing, respectively, natural and experimental conditions existing among organisms and the environment in which they live.

My remarks treat broadly of what biological oceanographers are interested in and what they do research on, rather than a review of how they do their research. Instrumentation and problems relating to instrumentation -- which are mainly due to corrosion and fouling of instruments; handling of long cable lengths; problems in interpreting the data collected when a catenary develops in a cable carrying instruments, due to variable velocities of vessel or currents; problems created in transmission of data from depths to the surface; high ambient pressures; selection of radio frequencies for transmitting data from buoys and satellites to ships and shore stations; etc. -- will be mentioned briefly.

The Hydrospheric Climate. Ruttner[28] has succinctly outlined the two-fold effect of the water environment upon life within it: 1) through its physical properties, as a medium in which plants and animals extend their organs and move or swim; 2) through its chemical properties, as a bearer of the nutrients which produce the organic from the inorganic through the primary production of the plant kingdom. Thus, because of their importance to life, physical and chemical attributes of the water environment are often as fully studied as the organism by the marine ecologist. The scope of these studies is suggested by the magnitude of the celestial, planetary, and hydrospheric conditions and processes depicted in Figure 3. For example, the sequence from light to fresh water runoff in the outer circle is the hydrologic cycle. The effect of this cycle upon marine organisms is particularly marked in estuaries and coastal waters through the variable modification of components of the local hydrospheric climate, as sediment, salts, nutrients, and dissolved gases in the water. For simplicity, not all of the components of the geochemical cycle are shown in Figure 3, but I wish to call attention to this cycle as it is also important in the ecology of marine organisms. It is linked to the hydrologic cycle through sedimentation, i.e., the erosion, transport, and deposition of sediments from the continents. An excellent discussion of the biogeochemical cycle can be read in Odum and Odum.[22]

Moore[20] believes that the latitudinal and vertical gradients are the most useful, of several gradients that might be chosen, in classifying marine organisms and their environments. The intensity and duration of light is the principal factor in both gradients. Incident light at the water surface is dependent upon the latitude; temperature is related to illumination, hence also shows a latitudinal gradient and decreases poleward. In the vertical gradient there is an extinction of sunlight illumination, usually by 80 to 200 or more meters of depth; temperature changes (see Figure 5C) are generally limited to a few hundred meters in depth at the surface of the ocean; and, pressure increases with depth at the rate of 0.442 pounds per square inch per foot of depth (about one decibar for every meter of depth).

I have emphasized the large scale phenomena associated with the hydrospheric climate to call attention to the fact that the biological oceanographer must bear these in mind and accordingly conduct his research within this context. A set of local environmental influences relating the hydrologic and geochemical cycles is shown by a hydrospheric climatograph in Figure 4. These are of interest because they help explain seasonal fish mortalities. The climatograph indicates that the amount of dissolved oxygen is apparently more nearly associated with the seasonal pattern of river flow rate than to water temperature. Since the amount of dissolved gases is normally equated to water temperature another factor is obviously hidden in the data. That this is so has been suggested by Dr. A. Joel Kaplovsky, Director of the Delaware Water Pollution Commission (personal communication). Dr. Kaplovsky believes that the reduction of dissolved oxygen may be due to the stirring up of the bottom sediments by the increased river flow. Since sediments tend to absorb chemicals, it is entirely possible that oxygenophilic substances are mixed into the water column during increased river flow. The seasonal occurrence of fish mortalities in the Delaware River, particularly during the time of the spring flows, suggests that the hypothesis deserves careful study.

Since sound propagation in water is likely to be of interest to radio engineers, I have included some graphs of the velocity of sound in the sea which show that the velocity varies according to the water temperature, salinity, and pressure (Figure 4). This variability in the transmission of sound in the sea has an obvious bearing upon the reliability of data collected by sonar and echo sounders. Further, the kind of sound must be identified since sound originates from three principal sources: geological, biological, and mechanical (as from submarines).

Community Metabolism. Several interacting environmental factors were shown in Figure 3, with an indication that many of these factors ultimately affect the metabolism of marine species. Metabolism was selected to represent life processes, since it is one of the unique characteristics of living organisms. Further, a concept of community metabolism is useful when an association of organisms is studied as a unit, particularly in an estimation of biological productivity.

The concept of community metabolism has been advanced most strenuously by Odum[23]. He has designed an electrical analogue circuit which furnishes an interesting model for this concept[24].

In the community metabolism diagram (Figure 6), we see the relationship between photosynthesis, respiration, and the flow of chemical energy in organic form through the community. Plants, through an elaborate mechanism involving absorption of sunlight and photosynthesis, utilize sunlight energy as an energy source for metabolism and the growth of plant tissues. Since animals are dependent upon organic food, plants are the basic source, directly or indirectly, of all organic food for animals. Utilization of the organic chemical energy dissipates this energy; this energy loss can be measured in the respiration of plants and animals. Thus, the energy stored in plant tissues can be visualized as flowing in one direction through the community, with a marked dissipation of energy at each consumer level from plants to herbivores to carnivores to decomposers.

Biological Productivity. The rate of biological productivity, usually measured in such units as grams or kilocalories per square meter per year, is often considered in connection with community metabolism as well as in the study of individual species such as fishes. Primary productivity, the rate at which plants and some bacteria (collectively called "producer organisms") store energy, is often studied to provide a reference value even when the principal study is upon other components of the community. An excellent review of biological productivity is given by Odum and Odum[22].

Bioengineering. A very limited presentation of this topic is given here, primarily to indicate the possibilities in the bioengineering approach to the study of marine organisms. As already stated, biological oceanographers study life and environmental conditions in the oceans, although many studies on marine organisms may be conducted in laboratories and museums far removed from the ocean. These studies, however, usually are based upon field observations. Among excellent laboratory-type studies, mathematical and engineering approaches to the analysis of the functional aspects of the structure of marine species are largely unexplored. In some cases it is advantageous to work upon preserved specimens, in others, to experiment upon living organisms. Examples of these studies will be cited from one area of my research interests.

Development of the mollusk shell often reflects changes in environmental conditions, particularly of temperature and siltation. Researchers, using the oxygen-isotope thermometer technique devised by Dr. Harold Urey, are able to determine the temperature, even in fossil specimens, at which the shell laminae were deposited[14]. Thus, analysis of the shell architecture[30] provides information on the effect of the environment upon shell deposition as well as upon the past history of the animals. Since the principal architectural pattern consists of laminated structure with buttress-type strengthening in regions of greatest curvature, much could be learned about "engineering" principles involved in shell architecture.

D'Arcy Thompson[11] provides a notable introduction to the study of growth and form among organisms; one attempt to further analyze shell form and growth, in mollusks, has been made by Owen[25]. Certain future lines of research are indicated, as for example, application of the kind of mathematical analysis used by Dean[12] in the study of plywood and the application of common engineering tests to a quantitative determination and statistical study of the mechanical and structural properties of mollusk shells. This would permit a sophisticated level of study of the structural form and strength of the shells and their component parts not yet attained by biologists in their analysis of shell structure.

Several pioneering researches indicate other lines of intriguing research as in studying bioengineering aspects of the development of hydrostatic pressure. This pressure is effectively used in swimming and jet-propulsion by scallops[37] and squid, and in the digging by clams and marine annelids[6]. The significance of this pressure in the activities of these marine organisms is indicated by the term "hydrostatic skeleton" used to describe the fluid mechanism which provides a functional relationship between contractile elements of the animal and its body fluids[6]. As to engineering applications, I wonder for example, whether a suitably and simply propelled, water-inflatable, double-walled, pointed tube of plastic, resembling the shape of a giant squid, could be designed as an effective method of long distance transport for a "frogman."

Collection of Specimens and Data

Three essentials usually are required in the collection of oceanographic information: 1) oceanographers and/or 2)instruments, and, 3) a vehicle or platform to carry or support the observers or instruments. Combinations of these essentials may be as simple as that involved in strewing plastic-covered cards or driftbottles from a small boat to study surface currents, or in the utilization of precision instrumentation such as loran and GEK (Geoelectrokinetograph) on a modern oceanographic research vessel to measure subsurface currents.

In most cases, there are but two general approaches to either observing or collecting marine specimens, sediments, hydrographic data, etc.: either go down yourself or send down collecting or "sensing" devices to do the job for you. Since these two approaches are often combined, a whole spectrum of possibilities has evolved. For example, the evolution of manned sensing devices has proceeded along several lines from surface swimming to unprotected diving to diving ships. A big step was made when primitive underwater craft, capable only of shallow water observations, were superseded by the cable-suspended, bathysphere which made deepsea explorations in the 1930's. Then, in the 1950's came another significant advance: the self-propelled bathyscaph, which recently probed the deepest oceanic trench, diving to depths up to 10,700 meters; in shallower water a two-man diving saucer is being used. Now an aluminum submarine, the Aluminaut[2] is being built, with launching scheduled for 1963. The Aluminaut will be capable of undertaking a wide variety of geological, biological, and physical research; operating to depths of about three miles. This submarine will allow the crew to explore about 60 percent of the world ocean floor, most of it for the first time.

Many more kinds of vehicles or platforms are in use or are being designed. These include: special oceanographic research vessels; inflatable, metal-reenforced, rubberized sea-going craft that can be transported by air and then assembled at a research location within a few days; aircraft, including helicopters and air cars; Texas towers; the Cuss I, used in the first attempt to probe the thin crust of the earth at the bottom of the ocean in the Mohole project; and, two variations of tubular research platforms. These latter, some 16 feet in diameter and 300 feet long, have gimbaled compartments and instrument panels, that will enable researchers to tow the platforms in a horizontal position to the study area and then position them upright for work: the FLIP or Floating Instrument Platform[21] and the SPAR or Seagoing Platform for Acoustic Research[3]. In time there no doubt will be oceanographic-meterologic space ships orbiting the earth and reporting "bird's eye views" of the ocean surface and atmospheric conditions.

Scattered sampling of oceanic depths yields incomplete or biased oceanographic records. Repeated, adequate sampling is an obvious necessity if either space man or oceanographers are to achieve a better impression of an environment and its occupants. Statistically valid sampling of specific environments at certain geographical locations, or of transient or moving conditions, requires precision navigation and/or the use of sensing devices which can track the conditions being followed, as: radar, sonar, loran, RDF, and the Swallow float, to name a few. The last named was used effectively in studying deepsea currents during the International Geophysical Year and in subsequent studies[26, 34].

If it is not already clear from the discussion thus far, it should be emphasized that collection of marine organisms and oceanographic data is not only an expensive undertaking, but it is also very time-consuming.

Utilization of Knowledge

As in other sciences, when we review the advances that have been made in oceanology, we are impressed by the quality and volume of information obtained by the pioneers of the science, despite the crude methods and instruments available to them. Much of this information is contained in publications of limited issue, and these are not available in every oceanographic library. Not only should this past record be more widely known and consulted, but there are "growing pains" also. Even today, with our relatively high level of technology, there are several bottlenecks preventing rapid utilization of accumulating knowledge. There is a need for more rapid processing of the material collected on oceanographic cruises, through laboratory studies, etc., and translating these into useful information.

Major Bottlenecks

The most critical bottlenecks stem from lack of sufficiently trained, qualified personnel and from lack of accurate, labor-saving methods and devices in the collection, processing, and interpretation of data.

Of course, greatly increased financial support of oceanographic research is sorely needed, but this in itself will not break the present bottlenecks. To avoid perpetuating and intensifying the communication-of-knowledge bottleneck, vastly improved feed-back of information is needed. Instrumentation may be the key to speeding up this feed-back.

Improved Library Facilities. The rate of library acquisitions is seemingly progressing in geometric proportions, such that even foremost researchers have difficulty staying abreast of new literature. Coupled to this publication rate is the lack of adequate financing and facilities at most marine science laboratories to either accommodate acquisitions or to purchase all of the references required for their work. Establishment of large, centralized, well-financed libraries seems to be an answer to this bottleneck. Techniques of rapidly retrieving information, even beyond that now furnished by abstracting journals is sorely needed, but even if the central libraries only provided at cost rapid photocopy and transmittal of library material upon request, this would be an important service. The ultimate in library science may be instrumentation, whereby a Univac-like system with information properly coded and stored would enhance the rapidity and volume of retrieval of pertinent data. A less extensive service was envisioned earlier as a need in connection with reports on estuarine research[31].

Systematics. The science of systematics deals with the identification of organisms, with describing new species, and arranging all species within a comprehensive classification. There are not enough sufficiently trained systematists to handle the quantities of specimens now in marine collections, let alone provide the basis for expanded volume of collecting. Now that instrumentation has entered the field of language translation, further refinements may produce optical scanning devices that can assist the systematist in routine identification and quantification. Identification of most species may always demand the direct attention of a systematist, but it is possible that analysis by larger taxonomic categories, particularly of plankton groups, could be speeded up by instrumentation.

Synoptic Data. Summaries of atmospheric climatic data, relating one climatic factor to another, give a useful synopsis of prevailing conditions. This same approach has not been widely applied to summarization of oceanographic data. The chief drawback is the lack of sufficient data from which to prepare the synopsis. Instrumentation can be the key to the collecting of data for synoptic treatment. Several parameters of the hydrospheric climate can be continually sampled from stationary platforms, such as Texas Towers, light houses, or anchored or floating buoys. As we gain experience and sophistication in handling the data collected, this data, when run through computers, should quickly provide answers to problems concerning the differential or combined effect of environmental factors upon life in the ocean.

Acknowledgements

Colleagues at the University of Delaware have aided the preparation of this paper. Their help is gratefully acknowledged. Dr. G. Fred Somers specifically suggested the placement of a light absorbing unit in the plant community presentation of Figure 6. Mrs. Nancy S. Smith assisted in the preparation of the illustrations.

Figure 5 was redrawn and the indices converted into the metric system from the article on "The 'liquid jungle.'" in Vectors, 3(3): 12-16 (Mr. Chester N. Hess, Editor).

As a participant in a recent meeting of oceanographers at Washington, D. C., the National Science Foundation "Conference on Marine Laboratories" (February, 1962), I have benefited from the exchange of ideas with colleagues. Where this article falls short of representing items discussed with others or learned at the conference and elsewhere, the responsibility is mine.

References

There is an ample non-technical and technical literature for anyone interested in the various aspects of oceanography. The following selection of books, listed in relative order of increasing technical presentation, will provide a reasonable introduction to oceanography in its many facets: Carson[5], Walford[36], Cowen[9], Coker[7], and Moore[20]; Sears[29], Buzzati-Traverso[4], Hedgpeth[16], Kuiper[19], von Arx[35], Sverdrup *et al*[33], and Defant.[13] The *Scientific American* and laboratory publications, *Oceanus* (Woods Hole Oceanographic Institution), *Sea Frontiers* (University of Miami Institute of Marine Science), and the *Estuarine Bulletin* (University of Delaware Marine Laboratories), provide excellent general reading. If the reader has access to libraries specializing in marine sciences literature, a better impression of oceanographic research can be gained by consulting the collected reprints of the large laboratories such as the Woods Hole Oceanographic Institution and Scripps Institution of Oceanography and journals such as *Marine Research*, *Deep-Sea Research*, and *Limnology and Oceanography*. There are also many excellent foreign publications which should be consulted.

One publication[27], still in preparation, promises to be of such significance and interest to the engineer that it is listed here. It will include, in addition to reporting the formal presentations and questions and answers concerning that position of the Naval Oceanographic Program related to instrumentation, an extensive listing of: U. S. laboratories doing oceanographic work; companies interested in making oceanographic equipment; government agencies having oceanographic programs; and, instruments that various agencies need to have developed. A similar publication[32] is also nearing completion.

Information on oceanographic institutions can be found in an annotated listing of hydrobiological laboratories[17] which is being revised.

1. Anon. 1961. The "liquid jungle." Vectors, 3(3): 12-16.

2. Anon. 1962. Aluminum submarine for oceanic research. *Commercial Fisheries Review*, *24*(2): 33.

3. Anon. 1962. Navy announces design plans for a seagoing acoustic research device longer than a football field. *Navy Oceanographic Newsletter*, *1*(3): 5.

4. Buzzati-Traverso, A. A. (Editor). 1958. Perspectives in Marine Biology. University of California Press.

5. Carson, Rachel L. 1950. The Sea Around Us. Oxford Press.

6. Chapman, G. 1958. The hydrostatic skeleton in the invertebrates. *Biological Reviews*, 33(3): 338-371.

7. Coker, R. E. 1954. This Great and Wide Sea: An Introduction to Oceanography and Marine Biology. University of North Carolina Press. (1962. Torchbook edition, Harper & Brothers).

8. A Committee Report. 1959. Introduction and summary of recommendations of the Committee of Oceanography. Amer. Sci., 47(2): 234-249.

9. Cowen, R. C. 1960. Frontiers of the Sea. The Story of Oceanographic Exploration. Doubleday & Co., Inc.

10. Dansereau, P. 1957. Biogeography: An Ecological Perspective. Ronald Press Co.

11. D'Arcy Thompson, W. 1942. *Growth and Form*. Cambridge University Press.

12. Dean, D. L. 1958. Design, construction and testing of a plywood hyperbolic paraboloid lattice structure. *University of Kansas*, *Engineering and Architectural Bulletin*, *No*. *41*: 1-19.

13. Defant, A. 1961. Physical Oceanography (2 vols.). Pergamon Press.

14. Emiliani, C. 1958. Ancient temperatures. *Scientific American*, 198(2): 54-63.

15. Finch, V. C. and G. T. Trewartha. 1942. Elements of Geography: Physical and Cultural. McGraw-Hill Book Co., Inc.

16. Hedgpeth, J. W. (Editor). Treatise on Marine Ecology and Paleoecology: Volume 1, Ecology. Geol. Soc. Amer., Mem. 67.

17. Hiatt, R. W. (Editor). 1954. Directory of Hydrobiological Laboratories and Personnel in North America. University of Hawaii Press.

18. Hull, C. H. J. 1960. Discussion: Oxygen balance of an estuary. J. Sanit. Engr. Div., Amer. Soc. Civil Engr., 86(SA 6): 105-120.

19. Kuiper, G. P. (Editor). 1954. The Solar System, II. The Earth as a Planet. University of Chicago Press.

20. Moore, H. B. 1958. Marine Ecology. John Wiley & Sons.

21. Neeson, Margaret G. 1962. Flip, the sea-going pencil. Skipper, 22(1): 30-31, 44.

22. Odum, E. P. and H. T. Odum. 1959. Fundamentals of Ecology (2nd Ed.). W. B. Saunders Co.

23. Odum, H. T. 1956. Primary production in flowing waters. Limnol. Oceanogr. 1(2): 102-117.

24. Odum, H. T. 1961. Ecological potential and analogue circuits for the ecosystem. Amer. Sci., 48(1): 1-8.

25. Owen, G. 1953. The shell in the Lamellibranchia. Quart. Jour. Microsc. Sci., 94(1): 57-70.

26. Pochapsky, T. E. 1961. Exploring subsurface waves with neutrally buoyant floats. Hist. Soc. Amer. Jour., 8(10): 34-37.

27. Rockwell, J. -- Editor. (In preparation). Proceedings of the Government-Industry Oceanographic Instrumentation Symposium. Miller-Columbia Reporting Service (931 G Street N. W., Washington 1, D. C.). ((300 to 400 pages, lithoprint; about $5.00 per copy)).

28. Ruttner, F. 1953. Fundamentals of Limnology. University of Toronto Press.

29. Sears, Mary (Editor). Oceanography. Amer. Assoc. Adv. Sci., Publ. 67.

30. Shuster, C. N. Jr. 1957. On the shell of bivalve mollusks. Proc. Nat. Shellfish Assoc., 47: 34-42.

31. Shuster, C. N. Jr. 1958. Suggestions for a national program in estuarine research, education, and conservation advisement. Minutes, 17th Ann. Mtg., Atlantic States Marine Fisheries Commission, Appendix MA-2: 13-21.

32. Smith, M. (Publisher). Hydrospace Buyers' Guide. Data Publications (1831 Jefferson Place, N. W., Washington 6, D. C.). (($10.00 per copy)).

33. Sverdrup, H. U., M. W. Johnson, and R. H. Fleming. 1946. The Oceans: Their Physics, Chemistry, and General Biology. Prentice-Hall, Inc.

34. Swallow, Mary. 1961. Deep currents in the open ocean. Oceanus, 7(3): 2-8.

35. Von Arx, W. S. 1962. An Introduction to Physical Oceanography. Addison-Wesley Publishing Co.

36. Walford, L. A. 1958. Living Resources of the Sea. Ronald Press Co.

37. Yonge, C. M. 1936. The evolution of the swimming habit in the Lamellibranchia. Mémoires Musée Royal D'Histoire Naturelle de Belgique, 2 ser., Fasc. 3: 77-100.

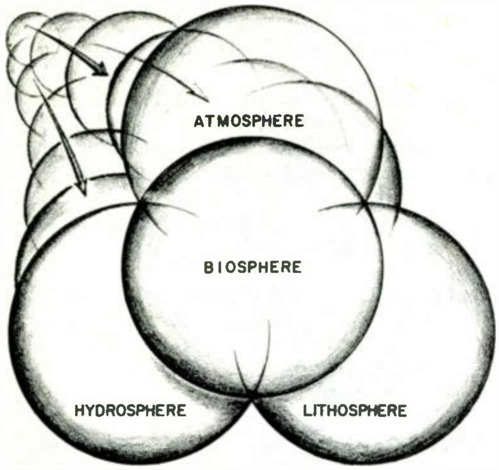


Fig. 1. Environmental processes have affected organic evolution throughout most of geologic time on the Earth. This diagram portrays the four major intersecting spheres within which the environment processes are continually interacting.

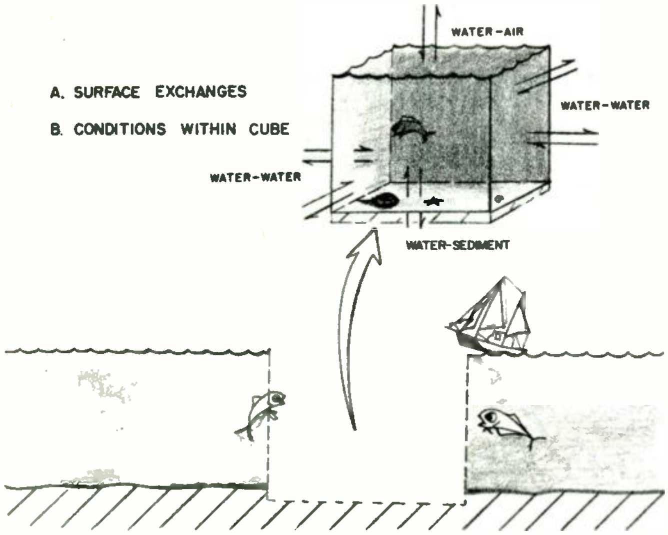


Fig. 2. An intensive study of a cubic section of the Earth's surface, measuring all of the exchanges between the biota and layers of the cube (air, water, and land) and between the cube and its surroundings, would be of great theoretical as well as practical interest.

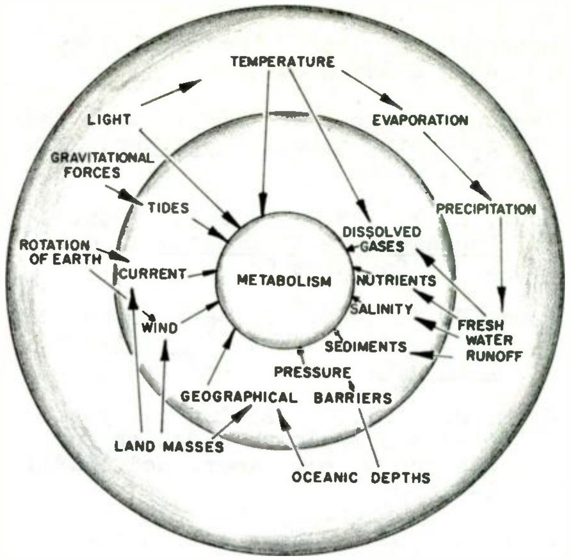


Fig. 3. Celestial and planetary conditions and processes (outer set) produce or moderate environmental conditions in the hydrosphere (inner set) that influence the metabolism of aquatic organisms. Based upon Dansereau's [10] treatment of Finch and Trewartha's[15] concepts.

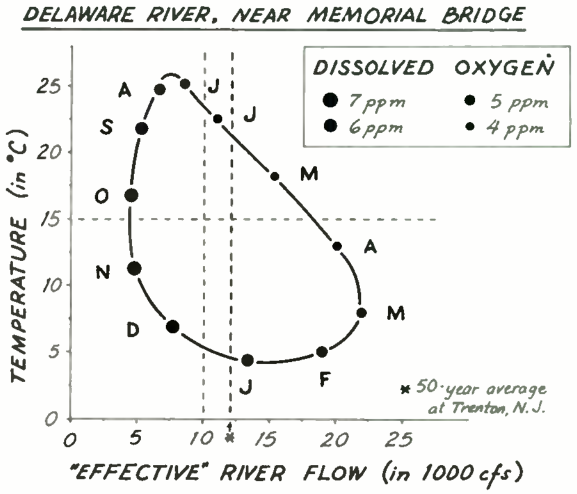


Fig. 4. This hydrospheric climatograph shows average monthly dissolved oxygen-temperature-river flow relationships during an average year, 1949 through 1958 (based on U. S. Department of Health, Education and Welfare, Public Health Service data).

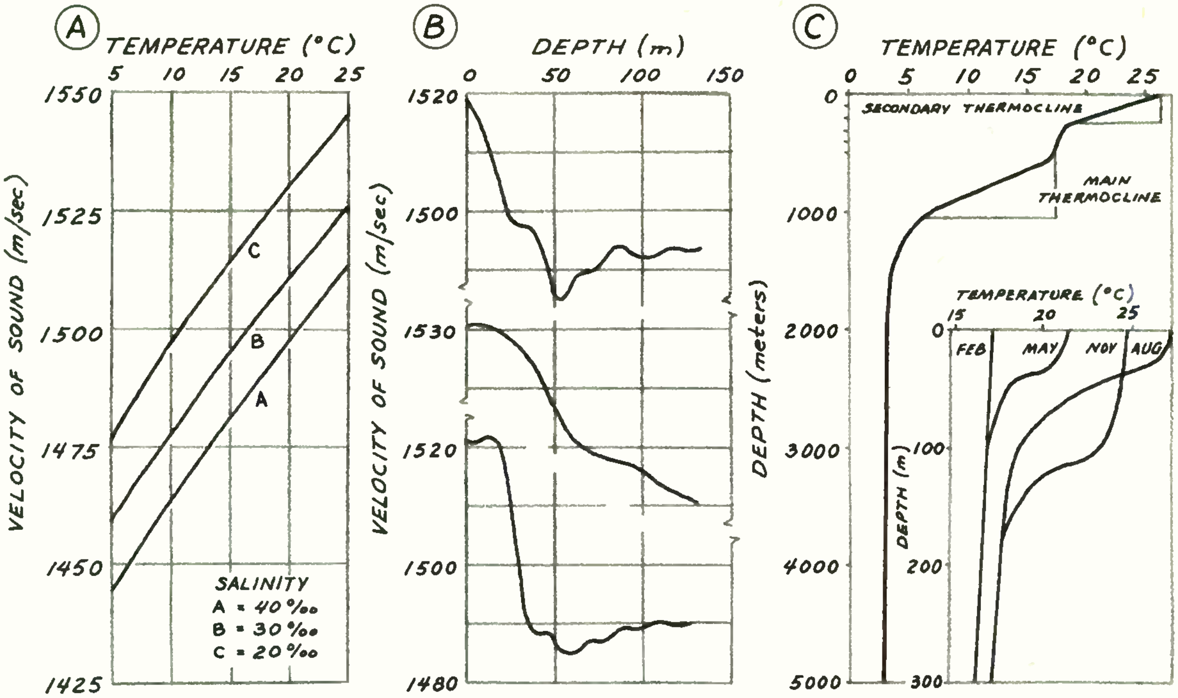


Fig. 5. The velocity of sound in sea water is not uniform since it varies with temperature, salinity, and pressure.

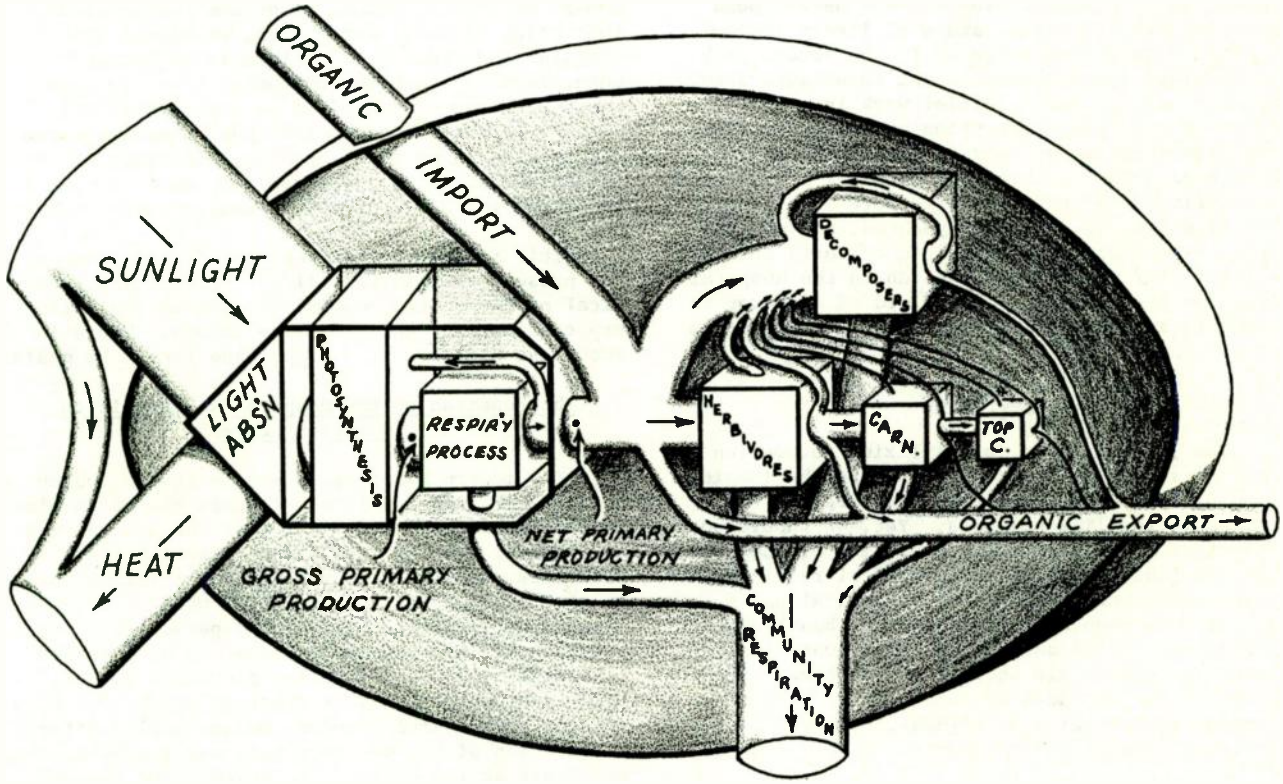


Fig. 6. Energy flows in one direction through a community of organisms. Measurement of the energy flow indicates the state of the metabolism of the community. If the energy flow is in equilibrium, the imports (light and organic material) equal the energy losses and exports (heat, respiration, and organic matter).

AUTOMATION IN HOSPITAL CARE

W. A. Spencer, C. Vallbona, and L. A. Geddes
Baylor University College of Medicine
and
Texas Institute for Rehabilitation and Research
Houston, Texas

Abstract Summary

Because attempts to automate hospital care are potentially important contributions of engineering, it is necessary to reveal the steps in this process. The avoidance of failures and confusion in this field require initial differentiation of two basic areas: a) routine hospital automation needs in management, business functions, and certain communication and information problems, which can be served now, and b) research into the fundamental medical sciences relative to patient symptoms and the integration of hospital care services with individual patient needs, which constitute a longer process. The routine uses depend upon efficient planning, and will likely follow hospital "operations research." The second problem involves learning what ought to be automated, what can be, and how. Initial work is now addressed to medical information procurement. The long-term learning process is just beginning and depends on joint activities in medical science and engineering directed toward the more subtle problems of diagnosis, disease dynamics, and medical care. Experimentally determined criteria to expedite fuller usage of automation in the hospital situation is expected. This field of endeavor is presently an outstanding challenge to engineering and medicine.

Introduction

The possibilities for utilizing automation facilities in hospital functions fire the imagination probably more than any other area of engineering in the life sciences. Yet, we should not lose sight of the necessary evolution of knowledge that must occur before practical realization is achieved. Automation in the hospital has many connotations such as data gathering, handling, processing, and computing. Applications of computers and electronic equipment have been tried and are being extended in the areas of patient diagnosis, monitoring of illness, scheduling of services, medical record keeping, fiscal control, etc.

Automation in hospital functions in the fullest sense means utilization of automatic information procurement, decision-making, and self-control of the major component activities that occur in care of the ill person, with appropriate safeguards and individualization. Automation engineers are indispensable components of these activities, because they help procure scientific knowledge and assist in making many practical applications of it. If the present status of automation in the hospital was ready for exploitation and full-scale utilization, and if it was of proven value, the present challenge to engineering would not be quite so crucial. In the minds of many in the field of health, the apparent situation is one in which people in medicine stand on the periphery of an expanding scientific technology and, on hands and knees, beg to see its application for humane purposes. Before we become too critical of this point of view, and others, we should look at hospital and medical objectives in an attempt to understand at least the component tasks we presently perceive. We need to be more sympathetic with ourselves and our colleagues in medicine and engineering as we face together challenges of the future and as we become receptive to the problems uncovered in our joint search and research efforts.

This review will attempt to identify some of the present work in this field, some of the practical problems, and hopefully point up realistic expectations for the immediate future. There is much less difficulty in conceiving long-term goals.

The Function of the Hospital and Physician in Patient Care

The modern hospital is an institution which renders services to ill man to cure him or to minimize his incapacitation. It has the unique task of insuring that the ill person entering its portals gets, under direct medical supervision, what treatment and assistance he needs when he needs it. Besides an identification of the person's illness (diagnosis), his disease process may be arrested or limited and even sometimes prevented by definitive treatment, that is medical or surgical. When this is not possible, modern medicine also affords a reduction of the person's pain and suffering and has means at its disposal to restore the chronically ill (restorative medicine or rehabilitation). The major task of the hospital thus entails the provision of personal medical and related services, which require resources of personnel and facilities. It must be organized so that it is efficiently, effectively, safely, and economically operated within the specifications of the services it must provide.

The utilization of hospital services depends upon the decisions or choices that a patient's physician makes in diagnosis and treatment. The physician must make many judgments and estimations, some of which are based on scientifically objective data or the results of precise tests, and others which use empiric knowledge, impressions that come from his special senses, clinical memory, and imagination. He must continuously improve the accuracy of these decisions by applying his specific knowledge about the processes of illness, and by using his knowledge about the nature and behavior of disease agents and pathological consequences of particular injuries. He must include all new sources of information and utilize instrumentation which improves his perception of the patient's status. He must periodically evaluate the results of his treatment choices. He must anticipate complications where possible and forestall deterioration which will render the patient unresponsive to treatment and recovery. He must plan in the face of extreme individual variation. Variability and complexity are hallmarks of dynamic responses of ill people, even among those with identical diagnoses. The physician's decision-making tasks,and the provision of services by the hospital which the physician orders, are related because the hospital is organized to carry out these orders.

Medical Problems Stimulating Application of Automation Technology

From the point of view of accuracy, a strict historical accounting of the developments in automation in the hospital would be in order, but this does not provide understanding of either the achievements or the problems that exist. It is more helpful to look at the questions which are posed during the physician's decision-making and use of hospital services. These questions can then be turned into needs and framework will be provided for rationalizing current activities in automation. Some of the typical physician's questions are:

1. What is the patient's illness?

2. How ill is the person and does he need hospitalization?

3. If he needs hospitalization, has the patient been in the hospital before? What were the diagnoses, course of treatment, and results at those times?

4. Is proper space available in the hospital?

5. What needs to be done now? What can be done later? What does not need to be done?

6. How is the person responding to the selected treatment or to natural recovery? Is he receiving what has been ordered? Did he receive it at the prescribed time?

7. How can his treatment be tailored to his changing requirements as he gets better or worse?

8. How can the manifestations of his illness and the results of his care be recorded with permanence and retrieved with ease?

9. How is his illness the same or different from others who have been seen?

10. How can what has been learned about individual patients be retained for future guidance in care of patients with similar problems or for research purposes.

11. How can the personal and economical significance of his care be determined and accurately measured?

Use of such simplification carries a hazard in which a simply stated problem is often not fully comprehended by either the physician or the engineer. Thus, in regard to physiological monitoring of ill patients, Doctor Blumberg is quoted in an article in the Wall Street Journal of December 19, 1961, as saying:[1]

> "In the electronics field, engineers don't know what the hospital needs and the hospitals often don't know themselves. All too often the result is a machine built to the hospital's specifications that the hospital finds somewhat useless."

All is not lost, because this process of probing uncovers basic problems which in themselves may unify elements and efforts that are now random. These typical questions point up routine uses of automation and at the same time indicate a definite need for long-term research and coordinated efforts between engineering and medicine. It is helpful to contrast the current medical information procurement process and that which may be imposed by automation.

Current Information Procurement and Usage in Medical Care

Two events alone or together initiate the process of information gathering and utilization in the activities of medical care. First, there is the perception of the phenomena of illness and the development of ideas or concepts which are formulated in the physician's mind. Usually, these ideas are related at first to treatment and secondly to cause, or vice versa, if knowing the cause clearly alters the plan of treatment. A series of processes commence which may be idealized as follows. There is elaboration of signs and symptoms and historical information by questioning and examination which progressively focuses on the problems that are emerging, and subsequently on the immediate therapy needs, if any. A rather loose recapitulation of these gleanings is made in a written record that also affords time, by its formulation, for the physician to digest findings and to consolidate the needed decisions. Lastly, there is an execution of immediate decisions, and later there may be incorporation of results from laboratory and other specialized tests into the evolving plan of management. Not just once, but in an individualized continuing process, these activities reappear in the course of the evolution of illness,treatment and recovery. Alternate choices in turn may have to be made. Unfortunately for automation this can be a unique constellation of events that depends upon the fact that the same disease in two individuals may call forth different responses which have different therapeutic significance. The problem, as it were, is that two suits of clothes of the same material and identical configuration and size may fit one person and not another. The imponderables presented by the patient's own personal reactions to his disease, and his expectations, may enhance or interfere with treatment resulting in a need for further modification of therapy. Thus, there is not often a hard and fast sequence in this process that may have many continuing independent, interdependent, and simultaneous aspects.

By contrast, as engineers are quite aware, when complex electronic and other instruments are used in such activities, an orderly sequence of events must be followed that are in accord with the logical consistency and design of the instruments and the functions to be measured. This can be helpful in some situations and highly frustrating in other circumstances. Beginning with the phenomena or event to be measured, there is usally a step-wise sequence starting with detection of the natural event. Herein, there is presently an important assumption that the data thus acquired are biologically representative. Next,there is conversion into some common form such as electrical signals for processing, temporary or permanent display of raw or partially processed data, presumably pure and unmodified by measurement. Then, discrimination and selection of data for additional alteration and conversion to a form that can be mathematically treated are made. Computation or biomathematical and biostatistical analyses are done for information extraction or representation. There may be display of the data or some of its attributes in a permanent and retrievable form. Finally, there is utilization, repetition, validation, etc. Thus, in the highest form of automation, it could be conjectured that detection, diagnosis, and measurement of illness, selection of therapeutic decision alternates, data processing, and execution of control actions and in-course revision of continuing actions would be continuously generated in an integrated system. This system would be to some extent self-governing, since it would incorporate response patterns for control of the process of care itself (Cybernation). In this rigorous sense, we must admit much of our current activity is limited indeed, and coupled mostly with physiological and biochemical data acquisition and written record processing. Data acquisition, processing, computing, data display, and mathematical analysis for personal usage rather than machine action are our major immediate objectives.

The parallel between the two operational systems, one manual and one automatic, should be examined. In reality, there is a larger task which must be accomplished first, and that is to understand precisely both of these systems and to establish by careful experimentation the extent to which they can be merged. What is practically desirable has to be known in terms of what parts of the human operation should be replaced if greater accuracy and safety can be proved.

Information Needs in Hospital Care Which May Be Influenced By Automation

Assuming that a major application of automation in the hospital is in the general area of the hospital and physician's information search and manipulation, then the prime initial purpose of automation is defined. This would be to gain useful clinical information and make it rapidly available for personal decision-making. In this context, a detailed review of the hospital information problem areas is helpful.

The areas of information search, some of which may be elegantly executed at the present time by the human "computer," are grouped as follows:

1. Getting information about the patient's status or situation:

a. In regard to his illness, its diagnosis, its cause, manifestations of the signs and symptoms, and their time course which are captured presently only in verbal,and to a lesser extent,in quantitative symbolization.

b. In regard to treatment choices and dosages.

c. In regard to hospital services that involve admission, the execution of treatment and final disposition or discharge of the person.

2. Getting information about the process of medical and hospital care:

a. The procurement and use of laboratory and technical diagnostic and nursing services.

b. The conduct of therapy; specification and control of drug administration; use of surgical, medicinal, and physical elements of care.

c. The way in which prediction and anticipation for early warning is made to head off complications and harmful conditions posing a threat to life or producing chronic incapacitation.

d. The current hospital communication process which uses the bedside medical record and order book as the permanent recording of descriptive information on patient responses, laboratory data, etc.

e. The operation of the hospital "system" itself in terms of patient flow, supply logistics, and personnel flow.

3. Procuring information on the medical, social, and economic significance of illness:

a. Hospital management, facility utilization, and development.

b. Procurement of statistics on prevalence of illness, etc., for preventive medical and epidemiological use.

c. Coordination in the use of all kinds of health facilities such as general or special purpose hospitals, public and private health agencies, etc.

d. Centralized individual health record keeping, including the record of previous illnesses, results of pre-illness testing, and other data that would be available rapidly on demand in the event of illness, anywhere, at any time.

4. Provision for the management of research in hospitals where care and research are either combined or separate activities:

a. Scientific instrumentation in discrete research tasks and shared usage of data processing and computing facilities in the institution or in regional centers.

b. Systematizing information flow, bibliographic recording for collaborative research tasks carried out in the hospital, among several institutions, and nation-wide.

5. Getting information on the need for, and ways to obtain and train, professional people to engage in the field of health services.

Hopefully, a frame of reference now exists in which the discrete and important automation attempts in the hospital can be studied with a perspective which takes into account these admittedly broad "design parameters" of hospital care automation.

Description of Current Research and Applications in Hospital Automation

A description of both contemplated and actual attempts at partial automation in the hospital setting is considered here in relation to information problem areas such as getting information about the patient's disease. We will ignore, partially, an important but fairly well developed usage of automation in fiscal management of the hospital.

Computer-Aided Diagnosis

Several significant efforts have been made in this area. The mathematical basis for diagnosis has been formulated by Ledley and Lusted. [2,3] Automated diagnosis, or answering by computer methods, the question of "What disease does the person have?" has stimulated the imagination of physicians, mathematicians, and engineers because diagnosis is a logical process which presumably could be carried out rapidly with the assistance of computers. There are problems in mathematizing both the semantic or "meaning" content of diagnostic signs and symptoms and personal observations which need to be surmounted. Problems also exist in quantifying signs and symptoms by identifying their time and magnitude characteristics. Matrices of symptoms of disease have been set up in a simplified mathematical approach to diagnosis in congenital heart disease.[4] In this instance, application of rigorous mathematical probabilistic theory has been made in computer-aided diagnosis. Similarly, diagnosing the probability of certain disorders of the blood, and programing missing diagnostic cues or tests needed to increase diagnostic accuracy, were reported at the first Rockefeller Symposium on Computer-Aided Diagnosis in 1959.[5] These and other intriguing applications of computer technology to diagnosis have a basic requirement of clarifying the information content of some of the data which has been or should be accumulated in the human process of diagnostic evaluation. This kind of work has also stimulated, directly, or entirely independently, studies of the diagnostic process itself. Critical evaluations are being made on the dynamics of the process which the physician actually performs but rarely teaches.[6]

Simulation of the diagnostic process, beyond application of probability statistics in matrices of signs and symptoms, must take into account the fact that not all individual diseases have mutually independent signs and symptoms. Diagnostic information and its content may appear in the absence of certain signs and symptoms, in their time course of appearance, relative magnitude, clustering, etc. These uses are not realized as yet, since they have to be based on extensive and accurate recordable and retrievable medical data of known information content not yet available. Computer diagnosis attempts have stimulated pursuit of such objectives. Considerable work is proceeding in both a diagnostic context and in fundamental research on automatic analysis of the electrocardiogram [7,8,9] and the electroencephalogram.[10,11] Digital computer analysis of components of the analog recording of these bioelectronic events has been made with success. Practical application has been achieved in the case of the electrocardiogram.

The general interest and activity are reflected in the fact that there were ten papers presented at the 1961 International Conference on Medical Electronics on computer diagnosis alone.

Continuous and Intermittent Physiological Monitoring

We have had some experience in this area. It is both important and provocative because there is unquestionable value in determining the course of the gravity of an illness and the dynamic responses to therapy. These activities have begun with the detection and recording (usually graphic) of the so-called "vital signs" to monitor the physiological status of the patient. The transduced events have included body temperature, pulse rate and volume, electrocardiogram (ECG) and heart rate from the ECG, respiration rate and recently depth (volume), skin electrical resistance changes (GSR), electroencephalogram (EEG), oxygen saturation of the circulating blood, either singularly or in various combinations. Beginning in 1953, the authors with Doctor Hebbel E. Hoff worked on instrumentation and analysis of periodic physiological measurements in individuals with chronic respiratory and other impairments.[12] Our joint experiences suggest that we are still at the threshold of identifying the information content of physiological data as it relates to illness. We have worked in gap areas such as transducer development for acquiring and transmitting analog data such as temperature, blood pressure, heart rate from the ECG, and respiration rate and depth.[13] Continuous application and usefulness have not been established. Many similar systems have been developed for processing and display of such data and are identified in the previously cited article in the Wall Street Journal for use in surgery, in surgical recovery rooms, and at nursing stations attached to or remote to hospital bed stations. Failures have occurred in interpretation in usage, and in coupling to the patient rather than in technical engineering and in hardware capability. Data discrimination for information content and data reduction to meaningful summary form depend upon research into the information content of physiological data. What aspects of such data should be processed must be established. For example, is severity of illness reflected in absolute magnitude values, of vital signs, or in time series changes, or in rate or slope changes, or in inter-relations of data sets, etc.? In spite of these gaps in knowledge, it should be pointed out that there has been empirical value to the usage of electrocardiographic and heart rate monitoring during the conduct of anesthesia and during intracardiac surgery where hypothermia is used and the surgeon must know the status of heart conduction as he repairs heart chamber defects.

At this stage of experience with physiological data acquisition, it seems important to provide a permanent analog magnetic tape record. Thus, as design requirements evolve and the fleeting but possibly meaningful variations in physiological phenomena are recorded, discriminate selection of information-bearing data or data components could be made from permanent reproducible records. Then the future programing for periodic or continuous monitoring, for continuous analog-to-digital conversion, and for determination of useful display modes would be greatly facilitated, and valuable guidance data would not be lost.

The preceding application areas are, therefore, directed towards finding more precise ways to diagnose illness, to refine choice of treatment, and to more accurately anticipate treatment needs in a particular situation. Obviously, too, this is in its infancy and will be studied with great interest by physicians and engineers.

Application of automation technology in procurement and usage of hospital services is at present a contemplative procedure which will soon be activated. Several techniques seem promising.

On Line Computer Interrogation

At the recent Scientific Session of The Heart Association of Southeastern Pennsylvania on The Application of Computers in Cardiovascular Disease, J.J. Baruch[14] proposed the feasibility of tying a large scale digital computer into the operation of a general hospital nursing station, laboratories, and pharmacy. With appropriate terminal equipment, a two-way system of communication between hospital personnel and a remotely located computer is visualized. Information could be rapidly obtained on the record number by name and residence, previous admission data, etc. upon the appearance of a patient in a hospital station. Recording and disposition of prescriptions,entry and interrogation to determine the drugs that are elected for administration could be systematized and presumably controlled. Hand written messages, orders, and to some extent, "vital sign" measurements could be actively entered into the computer memory. Continuous updating of working patient records could be automatized. Contradictions and lapses in this process could be identified and safety controls formalized. Ultimately, operational data for computer simulation and study of hospital function could be gained.

While not conceptually unique, this potential application of computers in hospital functions could be effective in several important problem areas. Development of workable time sharing of large computers tied into "satellite" input and output stations is needed in the hospital situation for reasons of economy. Shortening of the time interval for record look-up, updating, and display of wanted information on demand should be practically evaluated. Technical problems of rapid memory access to variable data in variable locations in the often unpredictable medical application must be solved. The influence of current attitudes of medical and hospital personnel in daily usage of automatic systems needs to be established.

Another approach to the automation of hospital record keeping has actually been started in several places. This work is concerned with current hospital recording techniques.

Application of Automatic Data Processing in Medical Records

Two of the authors have been concerned with digitizing part of the hospital medical record for the past four years.*[15,16] Nearly 100 source

* Office of Vocational Rehabilitation Project #318 entitled "Coordinated Physiological and Biophysical Studies for the Evaluation of Rehabilitation Procedures Employed in the Care of Individuals with Severe Neuromuscular, Respiratory and Circulatory Disabilities," 9/1/58 - 8/31/62.

documents have been designed and tested to provide for key-punched entry of: time ordered diagnoses, physician and nursing bedside observations; results from manual measurement of temperature, pulse, blood pressure, respiration, fluid intake and output and biochemical laboratory data; functional test data such as electrocardiogram analysis and interpretation, radiographic interpretations, and cardiopulmonary function test results; the kind of physical treatment and its duration, results of muscle testing; and data describing age, sex, family composition, etc. At first, business machine data processing equipment has been used, and recently the computing facilities of the Baylor Biomathematical Research Laboratory have been utilized. This practical experience has shown that the process of replacing manual record methods must first be prepared to deliver regularly and on demand useful output formats without apology for machine or technical problems. Even in a specialized research hospital with greater uniformity of services than the general hospital, gaining acceptance and use by the personnel originating the data are essential. It has been possible to demonstrate in modest special applications a great versatility and saving of time in assembly of relatively large data sets which are medically useful. Periodic evaluation and review techniques of industry (PERT-Program Evaluation Review Techniques) can be performed manually for selected individual patients. It should be possible to obtain retrospectively, patterns of treatment response and time estimates to give desired results in given disease situations. This will require availability of rapidly retrievable record data, progressively increased recording of both patient responses and the nature of hospital services used for each patient. Sufficiently large experience of this type should be susceptible to computer prognosis (prediction) of the time and kind of hospital effort that is required for a desired outcome of hospital care. Simultaneously, it should be possible to explore dynamic programing of hospital service utilization for each individual patient. This study has not explored the problem of digital recording of the narrative content of the medical record which is being done by Doctor Schenthal and co-workers of Tulane University Medical School[17] at the Hutchinson Memorial Clinic. Here, in conjunction with the Tulane University Computer Center, a mark-sense punch card history, physical examination, and laboratory record was generated, which could be machine punched, and then the data rapidly located in a large collection of records.

At the Lovelace Foundation, Schwichtenberg[18] has used a mark-sense punched card original medical record with some success in medical screening of pilot candidates.

There are several other groups working in this area who will not be identified in this brief review.

There is a pertinent kind of hospital systems research being conducted which should be mentioned, although it is not presently automated. This activity should provide essential information needed for automation.

Hospital Operations Research

Flagle[19] has been studying the hospital operation at Johns Hopkins Hospital. The valuable objects of operations research are grouped by him in the following way: a) demands for services--the flow of patients, b) logistics--the flow of equipment supply, etc., c) communications (records, orders, etc.), and d) organization of services--flow of human resources of the hospital. The disciplining of study imposed by operations research has been valuable in finding critical problems in these operational activities, in trouble shooting, and in suggesting policy revisions. It is expected to find logically the fundamental mechanisms of hospital operation and their characteristics. This kind of research is desperately needed in advance of the sweeping changes that may be imposed by automation. Automation ought to replace at first those things that are not safe and efficient. In turn, when safe and effective automation usages have been established, then economic considerations become the determining factor.

Specialized Instrumentation in Hospital Laboratories

To avoid incompleteness, a large field of specialized technical instrumentation development should be mentioned. The hospital laboratory shares in technical advances on a broad front. Direct carry-over from medical and other scientific research and industrial application is going on at a rapid pace. Sophisticated equipment is available for automatizing biochemical laboratory analysis, for safety monitoring where forms of radiant energy are being used such as in x-ray therapy and in radioactive tracer laboratories; for many diagnostic applications in radiology such as fluoroscopic image intensification, cinefluorography; for highly specialized application, such as automatic blood cell counting, densitometry, and even automatic instrument mass screening for abnormal cancer cells in specimens obtained from body cavities.

Potential application of special instrumentation should be mentioned in regard to work going on in the use of digital computers in pattern detection which has potential value in x-ray screening for pathology, in automatic computer analysis for EEG screening, and more. It is possible that these important and discrete efforts will lend themselves to incorporation into integrated information systems in the hospital by means of systems research applications in the distant future. At the present, as in industry, instrument compatibility, standardization, etc. are recognized as very practical problems in medical application, but substantial research on these applications is needed.

In the larger activity of inter-institutional sharing of experience and coordination of function which should benefit by automation technology several groups are quite active.

Inter-hospital Information Processing

Slee[20] has established the feasibility of merging hospital record face sheet data from several institutions. Thus, hospitals would contribute to a generally available pool of experience with disease incidence, results of therapy, etc. Several large joint studies using computing center facilities for data handling are under way in cancer detection, evaluation of the nature and results of surgical treatment of cerebral vascular disorders, in cancer chemotherapy for surveillance and drug screening, and in cancer radiation dosimetry calculation to name but a few. This is an activity which will mushroom and will become a highly significant and disciplining force that will undoubtedly influence hospital practices.

Automated Centralized Health Record Keeping

As hospital input data is progressively improved in quality and quantity, a major anticipated usage will be the study of technical requirements for centralized health record keeping. This health usage of automation will likely follow in time, centralized and full automated record keeping of income for tax purposes, and population finger printing identification, etc. Dr. Frederick Moore, at the School of Medicine of the University of Southern California has pointed out that besides the patient's experiences in the hospital itself, the patient's future course out of the hospital and his movements among health agencies must be established if we are to evaluate or improve hospital operations.[21] This cannot be done with our present record systems because a problem exists that is not unlike automatic bibliographic search of the scientific literature. The greatest stimulus for improvement of our information process will come from real scientific use of records which will show their present inadequacy.

Although we may appear to be overwhelmed by the formidable incompleteness and "softness" of our medical data, the rigors of automation will create standards, policing, and proper morale.

Therefore, it can be conjectured that progressive improvement in quality of medical information both in and out of the hospital, provision of efficient health services at all levels, and establishment of bench marks for evaluation and guidance of health services, may presently merge on the ground swell of automation.

Aspects of Automation and Professional Training

As we explore and prepare for hospital automation, we cannot ignore the need to train ourselves and those who will follow in these endeavors. With the hazard of oversight, we would like to identify briefly some personal activities. In the institutional complex of the Texas Medical Center, video, audio, and direct wire data telemetry is employed in conjunction with joint teaching activities in the basic medical sciences. Laboratory and clinical demonstrations are conducted between Baylor University College of Medicine and the Texas Institute for Rehabilitation and Research. This opportunity to expose the young medical student, the engineer, the instructor, and the professor, to the application of scientific techniques to clinical problems, and to jointly explore the phenomena of experimental physiology in the laboratory has been most rewarding. We are seeing besides the student's technical proficiency, a readiness for independent critical thinking, for contribution of research knowledge, and for genuine respect of the many disciplines composing our present health endeavors. These and other experiences encourage us to expect that the young physician will place many of the aspects of scientific medicine in the hospital of the future in a new and fresh perspective. Their likelihood for accelerated contribution to hospital care and automation is very probable.

Discussion

It is necessary to be circumspect to specify the significant contributions of automation to the hospital care process and thus, to leave it largely unanswered. This, of course, reflects the unsteady state of our knowledge. It is also a consequence of our current desire to have solutions before even the problems are known. This review has identified many activities, trials, applications, and pursuits. Most are incomplete. Many are becoming more interdependent. Some are conjectural; several are anticipatory of the future; most have unknown practicality if we separate off purely business management aspects. This is not unexpected as medical care in the hospital becomes more scientific and yet remains patient-centered. For a long time it appears the physician must be like the key person in the space flight control center, who in spite of all that instrument technology can offer, must still personally decide "go" or "no go." We do not believe this is unrealistic or avoiding the challenge of the future. At present, we have more needs than accomplishments. Breakthroughs are not expected and we have to spend a great deal of undramatic, diligent, carefully planned, and executed searching, researching, and study application. We have to find what is quantifiable and what ought to be quantified before full automation. We can probably accomplish the needed experimental and simulation research before automating a large part of hospital care.

We now have and may have further unequal evolution of knowledge and technological sophistication in this field. Currently, we are almost in the situation of trying to fathom the precise design, the components, and their inter-connections, in a huge operating computer to which we can attach, on the outside, only two electrodes.

Since we have voluntarily chosen to undertake the large task of automation in hospital care, it is essential to prepare now for the eventualities. Physicians must be trained who can think like engineers and, indeed, do some things the engineer does. We have to train engineers to experience the variability, continuity, self-controlling, self-replacing, reactive, adaptive, dynamic characteristics of bio-systems. The levels of organization and the modes of biological coupling and control are unique aspects of mammalian structure and function. These should be intriguing to automation engineers, who by training and talent are ideally prepared to assist in developing knowledge in this area.

There are some simple signposts from our own limited experience in particular, which we believe to be worthy of repetition.

At this juncture of merging engineering and medical activity, as in many such interfaces, we believe we are in a situation where the engineer must obtain design data for himself. For example, he must acquire personal knowledge of the physiology of the event to be measured, this in turn being one of the design parameters. He ought to have thorough understanding of the situation of measurement and the significance of variation in these states. He should bear in mind some of our ancient heritage on instrumentation, such as: "addition of the measuring instrument must not alter that which is being measured."

Although it may seem rigorous to place some of this responsibility on the engineer , it seems to be appropriate that often in any new field of endeavor, daily work must progress ahead of knowledge to obtain missing design information. In this we expect much sympathy from the engineers. Frequently, the correct design data can only be obtained from a model synthesized from a combination of definite data and approximately of the parameters not available at the time. Of course, the true engineering approach also demands immediate quantification and analysis of the newly found information. In addition to and adhering to the law of instrumentation, irrespective of the physiological data to be explored, Rein's criterion is worthy of repetition;[22] "maximal efficiency in the transducers, a minimum of electronics."

Conclusion

Practically all of the current accomplishments in the development and application of hospital care automation have required much assistance from the engineer, physician, and research scientist. Joint and individual tasks have already been undertaken. Nearly always joint experience has been required. Pursuit of appropriately large tasks in a highly integrated, closely coordinated effort with good specification is a reasonable future expectation.

The initial work will be addressed to medical information procurement and rapid utilization in hospital operations as presently constituted. A fairly routine automation job exists in the simpler parts of hospital operation and management. Concurrently, the engineer and physician must engage in a long-term learning process in applying automation to the more subtle problems of diagnosis and medical care. Finally, depending on the outcome of this, experimentally determined criteria will be found to expedite the fuller application of automation technology in the hospital situation. This composes one of the outstanding challenges to engineering in the life sciences which by good fortune is emerging as a productive field.

References

1. Carley, W.M.: Ailing Medical Aids. Wall Street Journal, December 19,1961.

2. Lusted. L.B. and Ledley, R.S.: Reasoning Foundations of Medical Diagnosis. Science, 130: 9,1959.

3. Ledley, R.S. and Lusted, L.B.: Use of Electronic Computers in Medical Data Processing. IRE Transactions on Medical Electronics, ME-7: 31,1960.

4. Warner, H.R., Toronto, A.F., Veasey, L.G., and Stephenson, R.: A Mathematical Approach to Medical Diagnosis. Journal of American Medical Association, 177: 177,1961.

5. Proceedings of Conference on Diagnostic Data Processing. IRE Transactions on Medical Electronics, ME-7: 1960.

6. Eden, M: Personal communication.

7. Cady, L.D., Jr., Woodbury, M.A., Tick , L.J., and Gertler, M.M.: A Method for Electrocardiogram Wave-Pattern Estimation. Circulation Research, 9: 1078, 1961.

8. Caceres, C.A. and Rikli, A.E.: The Digital Computer As An Aid in the Diagnosis of Cardiovascular Disease. Transactions of New York Academy of Science, 23: 240, 1961.

9. Pipberger, H.V., Freis, E., Taback, L., and Mason, H.L.: Preparation of Electrocardiographic Data for Analysis by Digital Electronic Computer. Circulation 21: 413, 1960.

10. Adey, W.R.: The Modeling of Cerebral Systems from Computation of Brain Wave Records. IBM 3rd Medical Symposium, October 9-13, 1961, Endicott, N.Y.

11. McCarthy, C.E.: Digital Computer Program for Automatic Analysis of EEG. Presented at the Scientific Session of the Heart Association of Southeastern Pennsylvania, entitled The Application of Computers in Cardiovascular Disease, February 27,1962, Philadelphia, Pennsylvania.

12. Geddes, L.A., Hoff, H.E., and Spencer, W.A.: The Center for Vital Studies--A New Laboratory for the Study of Bodily Functions in Man. IRE Transactions on Bio-Medical Electronics, BME-8: 33, 1961.

13. Geddes, L.A., Hoff, H.E., Spencer, W.A., and Vallbona, C.: Acquisition of Physiological Data at the Bedside. A Progress Report. The American Journal of Medical Electronics, 1: 62, 1962.

14. Baruch, J.J.: Hospital Automation for Administrator and Comptroller. Presented at the Scientific Session of the Heart Association of Southeastern Pennsylvania, entitled The Application of Computers in Cardiovascular Disease, February 27,1962, Philadelphia, Pennsylvania.

15. Spencer, W.A. and Vallbona, C.: Digitation of Clinical and Research Data in Serial Evaluation of Disease Processes. IRE Transactions on Medical Electronics, ME-7: 296, 1960.

16. Spencer, W.A. and Vallbona, C.: A Preliminary Report on the Use of Electronic Data Processing Technics in the Description and Evaluation of Disability. Archives of Physical Medicine and Rehabilitation, 43: 22, 1962.

17. Schenthal, J.E., Sweeney, J.W., and Nettleton, W., Jr.: Clinical Application of Large-Scale Electronic Data Processing Apparatus. Journal of American Medical Association, 173: 6, 1960.

18. Schwichtenberg, A.H., Flickinger, D.D., and Lovelace, W.R.: Development and Use of Medical Machine Record Cards in Astronaut Selection. U.S. Armed Forces Medical Journal, 10: 1324, 1959.

19. Flagle, C.D.: Operations Research in the Health Services. Presented at The First Joint National Meeting, The Operations Research Society, and The Institute of Management Sciences, San Francisco, California, November 8-10,1961.

20. Slee, V.: Automation in the Management of Hospital Records. Presented at the Scientific Session of the Heart Association of the Southeastern Pennsylvania, entitled The Application of Computers in Cardiovascular Disease, February 27,1962, Philadelphia, Pennsylvania.

21. Moore, F.J.: Summary Address. Presented at the 3rd IBM Medical Symposium, Endicott, New York, October 9-13, 1961.

22. Rein, H., Hampel, A.A., and Heinemann, W.A.: Photoelektrische Transmission Manometer zur Blutdruckschreibung. Pflug. Arch. ges. Physiol. 243: 329, 1940.

A PHASE-CHANNEL COMBINER FOR THE NRL SPACE SURVEILLANCE SYSTEM

M. G. Kaufman
U. S. Naval Research Laboratory
Washington, D. C.

Summary

The Space Surveillance System developed at NRL for the detection of earth satellites forms a fence across the southern part of the United States. Four receiving sites are alternated with three transmitting sites which illuminate satellites with radio energy as they traverse the fence. The angle of arrival of the reflected signals is measured at each receiving station by a compound radio interferometer. These signals yield a multiplicity of channels which are normally recorded on paper and analyzed manually by visual aids and special slide rules.

With the increasing satellite population and the repetitive nature of their orbits, the number of fence crossings has increased considerably. Automation in the detection process is needed to facilitate identification and sorting of satellites from each other (and from refuse), if a large backlog of data is to be avoided. The purpose of this report is to describe an electronic system which automatically combines the phase channels into one unambiguous channel and depicts the angles of arrival of the radio energy from the satellites.

The basic technique used in the phase-channel combiner is as follows. Since complete a priori information on the incoming signals is known, simulated signal waveforms, locally generated, are correlated channel by channel with each signal channel from the radio interferometer. At the moment of coincidence between each simulated and live signal channel, a "marker" pulse is generated. This process is applied simultaneously to all of the output channels of the multi channel radio interferometer.

At the instant that the individual marker pulses, so generated, are coincident with each other, an index coincidence pulse is formed. This pulse controls phase-measuring equipment whose output is a recorded electrical analog of the space-angle position of the satellite. The resultant space angle so determined is derived from several phase channels in real time, and the resolution in angle is a function of the most accurate interferometer channel in the combination.

It is seen therefore, that the necessary complexity in multi-antenna combinations used in the radio interferometer detection system is not passed on to the data processing.

Introduction

The U.S. Navy Space Surveillance System for the detection of earth satellites forms a fence across the southern part of the United States(1). The fence consists of four receiver sites and three cw transmitter sites, the latter alternately located between the receiver sites. Both transmitters and receivers use fantype coplanar antenna beams wide in the east-west plane and narrow in the north-south plane. The cw transmitters illuminate the satellite, and radio-interferometer techniques are incorporated at the receiver sites to determine the satellite's position in the east-west and north-south planes. The receiving sites are located at San Diego, California, and Elephant Butte, New Mexico, in the west, and at Silver Lake, Mississippi, and Fort Stewart, Georgia, in the east (Fig. 1). A data-transmission system links the surveillance receivers to NRL, Washington, D.C., and the Operations Center at Naval Weapons Laboratory, Dahlgren, Virginia.

The angle of arrival of the reflected radio energy is measured at each receiver sited by a compound radio interferometer (one that has many antenna pairs) which yields a multiplicity of channels of phase data. These are fed to the data-transmission line (2) and transmitted to the Operations Center at Dahlgren, Virginia, for processing. At the Operations Center the data are recorded on paper in real time and are subsequently analyzed manually by visual aids and special slide rules. The Naval Ordnance Research Center computer, located at that facility, is being instrumented into the system to automatize this operation. NRL is rapidly instrumenting digital techniques to marry the outputs of the surveillance system directly to the computer.

Due to some practical limitations in antenna spacings, the phase channels are individually ambiguous in depicting the angle of arrival of the signals. Therefore it is necessary to make at least some cursory analysis of the data from several phase channels to establish whether the signals thereon are bona fide. With the increasing satellite population and the repetitive nature of their orbits, the number of fence crossings has increased considerably. Automation in the detection process is needed to facilitate identification and sorting of satellites from each other (and from refuse), if a large backlog of data is to be avoided.

The purpose of this report is to describe an electronic system which automatically combines several phase channels into one unambiguous channel. This single channel depicts the angle of arrival of the radio energy from the satellite in real time. The resolution in angle therefrom is a function of the most accurate interferometer channel used in the combiner. When the phase-channel combiner is used at the receiving sites there is the added advantage of saving channels on the data-transmission line from that site to the Operations Center. Other data, such as doppler and phase rate, can then be added to the overall information received from each site by utilizing the data-transmission channels so released.

Brief Description of the Space Surveillance System

The Navy's Space Surveillance System operates as follows. Radio energy radiated into space is reflected from objects to the receiving sites. The transmitting antennas provide illumination in a narrow fan-shaped beam which is coplanar with similar receiving-station beams. The position of a reflecting object in the common antenna patterns is determined by measuring the angle of arrival of the reflected signals by means of interferometers at two of the receiving sites. At each of the receiving stations the data from the phase-measuring equipment are transmitted, in real time, on phone lines to NRL and the Naval Weapons Laboratory.

The antenna field associated with a receiving site is arranged in pairs of antennas with selected spacings to provide the desired degree of accuracy and eliminate ambiguity. By means of amplifiers and converters, signals from an antenna pair are reduced in frequency, with phase preservation, and the phase difference is compared to a precision stable oscillator as reference. After being filtered, the phase-carrying signals are applied to analog phase meters. The outputs of the phase meter are recorded along with universal time to obtain a permanent record of the time and position of the satellite. The signals are phase coherent with respect to each other, varying in phase with respect to the stable reference oscillator as a function of the satellite's instantaneous position in space (phase rate) and the distance (baseline) between the antenna pairs (Appendix A).

A typical pair of antennas and the geometry of the situation are shown in Fig. 2. Signals are shown arriving from satellite S. Because of its great height as compared with the baseline d between antennas, the rays are essentially parallel. It is clear from the figure that energy reaching antenna B will arrive later than at A. The amount of this phase-front delay is

$$\phi = d \sin \Theta$$

where

ϕ is the electrical phase delay (degrees)
d is the electrical spacing between the antennas ($n\lambda/2$)
Θ is the angle of arrival measured from zenith (space degrees).

By instrumenting the antennas with phase-measuring receivers, the space angle Θ is measured as a voltage analog in ϕ. It is noted that when d, the antenna spacing, is equal to or greater than $\lambda/2$, determination of Θ becomes ambiguous; on the other hand, the angular resolution is proportional to d. For example,

$$\phi = d \sin \Theta$$

$$d\phi = d \cos \Theta \, d\Theta$$

$$\frac{d\phi}{d\Theta} = d \cos \Theta$$

noting that

$$\frac{d\phi}{d\Theta} = d$$

as $\cos \Theta \rightarrow 1$, near zenith.

In order to take advantage of the resolution improvement obtained by long baselines while circumventing the problem of ambiguity, many pairs of antennas are used, spaced so that signals from each successive pair can be correctly deduced from the signal of the previous pair, when taken in order from the shortest baseline to the longest.

In summary, it is noted that a pair of antennas provides phase signals for one data channel and that the channels must be analyzed sequentially for an unambiguous solution. The surveillance system incorporates many pairs of antennas to achieve the required angular resolution. Figure 3 illustrates where the phase-channel combiner fits into the surveillance system. Figure 3a shows the signals from the various antenna pairs going directly into the data-transmission system to be relayed to Washington and Dahlgren, Virginia, whereas Fig. 3b shows four channels passing through the combiner, where they are consolidated and then fed to the data transmitter.

Description of the Phase-Channel Combiner

The type of electronic instrumentation used for measuring the phase of the signals from each antenna pair generates a sawtooth voltage waveform for each cycle of electrical phase variation. The maximum amplitude of this sawtooth waveform is constant, and its frequency is proportional to the baseline length and electrical phase rate of the signal from one antenna pair. Figure 4 is an idealized drawing of several channels of these signals. The baseline lengths increase as the wave period decreases from channels 2 through 6, respectively. This set of waveforms represents the signals expected from the surveillance system as a satellite passes through the antenna beam from the western to the eastern horizon (staying at all times in the beam). This is not a practical case, but is used here for purpose of illustration.

As can be seen in Fig. 4, for every space angle read along the top of the figure, there are corresponding values for the phase readings of each channel, as read vertically and directly below. It can be shown mathematically, or cursorily from the figure, that there is a unique set of phase readings for each space angle. This of course is a necessary condition if the surveillance system is to function unambiguously.

Figure 4 can be drawn to the exact dimensions as the paper used in the surveillance system's data recorder, made into a transparent overlay, and placed squarely upon the data. An observer could monitor the data as it passed beneath the overlay and read off the space angle the moment all of the phase-data channels were individually coincident with the overlay's phase lines. However, this would be quite a formidable task. The phase-channel combiner to be explained does the above task electronically 100 times per second.

The combiner incorporates electronic means for simulating the phase channels exactly as shown in Fig. 4. A block diagram of its instrumentation is shown in Fig. 5. Thus with complete a priori information on the phase signals, it is only necessary to generate simulated signal waveforms and compare them electrically with the incoming signals. A voltage-comparator circuit for each channel instantly indicates, with a pulse, when the received signal is equal to the simulated signal. The pulse outputs of the comparator circuits are fed to a coincidence circuit, and at the moment that all of these pulses are in time coincidence an index pulse is formed by the coincidence circuit. The index pulse is used to control an analog phase meter whose output is read on a recorder. This reading then is an electrical analog of the space-angle, as a result of combining the received phase signals by simulation and comparison techniques.

The other blocks shown on Fig. 5 outside of the sawtooth generators, comparator circuits, the coincidence circuit, recorder, and the analog meter previously mentioned above, are used functionally as follows. The master oscillator, a precision 1-kc tuning fork counted down to 100 cps, is used to synchronize the combiner system. The various delay circuits are used for setting the phase (or time) relationships between the sawtooth generators and the frame gate. Since all of the received signal channels are compared simultaneously, it is the function of the frame gate to turn the overall system on and off at the time when all sawtooth generators are in proper synchronization. A pulse generator and counter are shown at the bottom of Fig. 5 which provide a running time train of pulses proportional to the space angle being measured. A computer, which may be used for further calculations, is shown fed by the pulse counter. The coincidence indicator in the lower right of the block diagram flashes with the coincidence activity in the combiner system. Random flashes are coherent channel-to-channel noise bursts, whereas a prolonged steady glow indicates a satellite pass. A panel meter calibrated in space angle is also included. The blocks shown dotted are the only additional components required to combine another channel.

In order to demonstrate perhaps more clearly the time relationship between the received phase signals, the simulated sawtooth waveforms, and the pulse coincidences, reference is made to Fig. 6, which represents three channels of the combiner action. The abscissas have been laid out as the natural function of cos θ (θ = space angle) and the ordinates as voltages which are a function of the electrical phase shift. The sawtooth waveforms represent the locally generated simulated signals, whereas the horizontal dash lines depict the instantaneous level of the received (or live) phase signals. Considering the abscissa shown at the bottom of the figure as a function of time, the frame gate starts at t_0 and lets each simulated sawtooth waveform run on until time t_2. During this period, which is approximately 0.01 second, all possible combinations of phase relationships between the three channels shown are generated. Calling attention to the bottom channel only, it is noted that there are six points of coincidence pulses per frame for this channel. Like action will occur on the other two channels, with a correspondingly greater number of coincidence pulses being generated, due to the greater number of cycles in the respective sawtooth waveforms. All of these coincidence pulses are fed to an "and" gate, which generates an index pulse only at the time all three of the individual channel coincidence pulses occur simultaneously. This is illustrated at the bottom of the figure on the time axis at t_1. Although there are many individual channel coincidences, there is only one time during a frame when the coincidences are coincident with each other, as shown by the vertical line t_1.

Carrying the analysis further, it is noted in Fig. 6 that the time index pulse (point 6) occurs when cos θ equals minus 0.56. The space angle θ in this case therefore is 56° 56' above the western horizon, or 33° 4' west of zenith. This space angle is automatically represented as an electrical analog by the analog meter shown in Fig. 5, which operates as follows. A reference pulse is generated at the beginning of the frame gate, which triggers a one-shot multivibrator circuit in the analog meter. The index pulse, mentioned above, resets the same one-shot multivibrator. The period of the square wave so generated is therefore proportional to the space angle. The square wave is subsequently integrated and filtered into a dc level which is fed to a recorder for viewing and processing. Returning to Fig. 4, the dc level or space-angle analog is shown as a dashed line on channel 4. It is noted that this combined resultant is unambiguous, with only one value for each space angle.

The train of pulses shown at the top of Fig. 6 is also used for monitoring the space angle. Their total number is proportional to the space angle, since they commence at time t_0, being triggered on by the frame gate, and end at time t_1 which is controlled by the index pulse. By using a high pulse-repetition rate, good resolution can be obtained by monitoring the pulse train with an electronic pulse counter.

It is noted therefore that the combiner has two signal outputs which are proportional to the space angle. The first, the analog phase meter, whose action is represented in Fig. 7, shows a variable pulse width controlled by the frame gate and the index pulse. This action can be stated mathematically as

$$\theta \;::\; K\,360^\circ \left| \frac{T_1}{T_1 + T_2} \right|$$

where

θ = space angle in degrees
T_1 = pulse width in seconds
$T_1 + T_2$ - period in seconds
K = constant of proportionality.

Similarly, the second, the pulse-train output, represented in Fig. 7, can be expressed as

$$\theta \;::\; K(\mathrm{PRF}) \left| \frac{T_1}{T_1 + T_2} \right|$$

where

PRF = the pulse repetition frequency.

Figure 8 illustrates the action of the coincidence pulses. The abscissa is time, and the ordinate is pulse amplitude. Four channels are shown. The wider pulses are used for the channels carrying the smaller baseline data, since the resolution is less on these, as explained earlier with reference to Fig. 2. By tailoring the coincidence pulses to approximately 20 percent of a sawtooth cycle for each baseline, more noise per channel can be tolerated before the overall four-channel coincidence is lost. The index pulse is shown crossed-hatched on the bottom trace, and it appears electrically only when there is coincidence between all four of the individual channel pulses shown directly above. For the purpose of illustration, consider a satellite located at 60 degrees west of zenith. As shown in Fig. 8a, the channel pulses will line up vertically, producing the index pulse. An inspection of the remaining pulses will show that there is not another case of four-channel coincidence. Continuing this check in ten-degree steps toward zenith, Figs. 8b through 8g illustrate the pulse action for the case of a satellite staying in the beam over the above range of space angles. Similarly, this process carries for space angles east of zenith.

The accuracy with which the combiner determines the space angle is related directly to the longest baseline being combined in the set. Another factor is the width of the coincidence pulses used, assuming equal phase noise on each channel. Referring again to the basic system equation,

$$\phi = d \sin \theta$$

or

$$\sin \theta = \frac{\phi}{d}$$

where ϕ is the phase (electrical degrees)
d is the spacing between antennas ($n\lambda/2$)
θ is the angle of arrival measured from zenith (space degrees).

Adding sky and system noise, N, to the electrical phase gives

$$\sin \theta = \frac{\phi + N}{d}\,.$$

In effect, the error in sin Θ is inversely proportional to the length of the baseline. This is illustrated in Fig. 9. Figure 9a shows one cycle of a typical sawtooth waveform, as seen on one output channel of the surveillance system.

Using time as analogous to sin Θ on the abscissa and the fact that sides of similar triangles are proportional, the figure shows that

$$\frac{\Delta t}{\tau} = \frac{N}{E\phi_1}$$

or the time error

$$\Delta t = \frac{N\tau}{E\phi_1}$$

where N is the noise voltage
Δt is the time error
τ is one period of the phase signal
$E\phi_1$ is the electrical phase signal

Now consider Fig. 9b, which shows a second phase signal from the surveillance system derived from a baseline between antennas twice as long as that of Fig. 9a. The slope of the phase signal will now be increased by a factor of two. Assuming the noise level on this channel to be equal to that on the other, the time error becomes

$$\Delta t = N\,\frac{\tau/2}{E\phi_2} = \frac{1}{2}\left[\frac{N\tau}{E\phi_2}\right].$$

Thus equal noise levels on two channels will be less degenerative on the one which has the greater signal phase rate. In general the error in sin Θ is inversely proportional to the length of the baseline. The coincidence pulse width in the combiner is set for 20 percent of a sawtooth for each channel. Noise peaks causing variations greater than this value will result in no combiner output, since pulse coincidence between channels will not occur. The 20-percent pulse width was chosen as a compromise between the assurance of having coincidence with weak and noisy signals and the problem of width so great that ambiguity results.

Allowing for ±5 percent noise, miscellaneous system tolerances in the combiner electronics, and the coincidence pulse widths used, a one-percent error in the combiner output can be expected for angles near zenith. For this reason current practice is to record several signals from the longer baselines concurrently with the combiner output. The former are used for the final resolution of the space angle.

An internal calibrator was incorporated in the combiner so that the system could be initially set up in the field and periodically checked during operations. This unit furnishes calibration signals to all channels, causing the combiner output to vary step-wise from 0 to 100 percent in 10-percent intervals, as shown in Fig. 10. The abscissa is time, and the ordinate is electrical phase level proportional to the sine of the space angle. The speed of the chart was 5 mm per second. In order to facilitate reading out the value of the space angle from the combiner analog channel, Table 1, "Space Angle vs Percent Deflection" has been prepared. Considering the lower edge of the combiner channel recording as zero percent (western horizon) and the upper edge 100 percent (or due east), the intermediate values are readily determined from Table 1. Station zenith, of course lies at the 50-percent line. In essence, the phase-channel combiner generates a virtual unambiguous baseline approximately $\lambda/2$ in length from several multiwavelength-long baselines.

Results of Laboratory Tests

An electromechanical phase-signal simulator was designed and used to check out the phase-channel combiner. The simulator consists of four motor-driven 360 degree potentiometers with the coupling-gear ratios chosen to provide rotations proportional to the antenna baselines. With a dc voltage applied to each potentiometer, slowly varying sawtooth waveforms are readily generated which are identical to those which are formed by a complete surveillance system. These simulated signals were then fed to each of the corresponding signal inputs of the phase-channel combiner. Figure 11 shows a recording of these simulated waveforms, labeled 1, 2, 3, and 4. The upper trace is the combiner output. It is shown repeatedly covering the space-angle range from horizon to horizon as the signal-simulator waveforms vary through several excursions from west to east and conversely from east to west, when the motor reverses. The abscissa is time and the ordinate is electrical phase. The chart speed is 1/4 mm per second.

Noise Test

During the laboratory tests on the combiner, 20-percent wide-band noise was added to each channel. No appreciable degradation of the combiner output signal was noted.

Stability Tests

These tests indicated that a calibration check should be made once every twelve hours to keep drift down to one percent.

Response Time

Response time was found to be very good, as can be deduced from the calibration steps shown in Fig. 10. A closer examination using a faster chart speed indicated a 25-millisec delay for full deflection, with 10-millisec attributed to the inertia of the recording stylus. This gives about an 8 to 1 safety factor over the fastest signal rise time expected.

Sensitivity Test

An rf signal is fed to all of the antenna pairs for periodic system calibration. Since all signals are in phase, the recorded phase channels will appear to read a zenith pass, or 50 percent, on each trace, as shown in Fig. 12. As the rf signal strength is reduced, system noise appears on each channel. Channel 4 is the combiner output, and the corresponding signal strength is monitored in channel 1. It is noted that the combiner channel gives zenith pass indication well into the noise of both the phase channels and the agc.

Life Test

The combiner was operated continuously for three days without failures, signal-output degradation, or excess heating of individual components. Upon completion of this final test, it was shipped to the Silver Lake, Mississippi, site for field testing with live signals.

Photographs of the combiner are shown in Fig. 13a and b. Figure 14 shows a typical station setup of two combiners flanking a common power-supply rack. Calibration and maintenance are performed on the standby unit, thus minimizing data interruptions.

Results of Field Tests

The overall results of these tests were quite satisfactory. Several clips from recordings of satellite signals are presented. The first is Fig. 15. When reading the record from the bottom up, the channels are:

1. AGC
2. Phase signal E-W (shortest baseline)
3. Phase signal E-W
4. Phase signal E-W
5. Combiner to
6. Phase signal E-W
7. Phase signal E-W (longest baseline)
8. Phase signal N-S (baseline in quadrature).

Timing is shown, coded, below the agc channel. Another auxiliary stylus carries the alert-pulse signal, whose trace is shown located between the agc and the first phase-signal channels. The combiner signal is the resultant of combining channels 2 through 5. (The phase signal for channel 5 is not shown, since this channel is now instrumenting the combiner output.) With reference to the combiner signal on channel 5, the normal no-signal position of the pen is at the top of the channel chart. It is noted that random phase noise occasionally becomes coherent between channels, causing sporadic combiner action. The upper edge of the channel represents 90 degrees east of zenith, the center line zenith, and the lower edge 90 degrees west of zenith. The pen is usually offset about 6 percent above the upper edge, so that there will be a distinct difference between no signal and a reading of due east. In Fig. 15, at the time that agc reaches maximum, the combiner channel reads 61 percent (zero percent is at the bottom of the channel). From Table 1, this indicates an object located at a space angle of 12.71 degrees east of zenith. When two receiving sites receive signals from the same object simultaneously, its height and location can be calculated. Another signal from an aircraft crossing the surveillance fence due west is shown in Fig. 16.

It may be noted that phase-noise quieting extends beyond the period of the agc signal, indicating a lower threshold sensitivity for these channels. Likewise, the combiner channel is not inhibited by low signal strength. Two examples of this are shown in Figs. 17 and 18. The first shows a solid signal on the combiner channel at 10.37 degrees west of zenith, with an insignificant rise in agc and the second shows the combiner reading well off each side of the agc main lobe. Since agc essentially renders a trace of the antenna pattern, it may be suspected that the combiner is reading coherent phase in the antenna side lobes. This is shown to be true in Fig. 19, where a strongly radiating satellite causes the agc to trace out several side lobes on each side of the main beam. The combiner is seen to follow through these lobes at approximately 46 degrees west of zenith. Figure 20 shows another very strong signal and the corresponding combiner action.

Many things can be deduced from the records as experience is gained in reading them. In order not to go too far afield from the combiner development, only a few aspects of record reading are considered apropos in this paper. Namely, the slope of a phase-channel trace indicates the direction the satellite is traveling as it crosses the fence. Considering the E-W channels, derived from the system's main antenna field a positive slope represents west-to-east travel, and vice versa. The surveillance system has several antenna pairs placed in quadrature to the main east-west field. They are known as the north-south set.

Using slopes from these in conjunction with the system's east-west slopes, it is possible to determine the relative direction the satellite is traveling as it crosses the fence (Appendix B). Figure 21 illustrates the various combinations involved. This method is more accurate for low-altitude passes near zenith, since the phase rate falls off toward the horizon, due to the searchlight* effect of the antenna beam and the cosine law variation of phase rate (Appendix A). Some of the slope combinations shown in Fig. 21 are physically impossible with the present geometry of the surveillance system and the laws of orbital mechanics; however, it is not unusual to have an extraneous signal behaving so as to mock one of these conditions.

In conclusion, therefore, with the slope information, a list of satellite predictions and the quick-look combiner reading on the space angle, the task of identifying satellites from space debris, meteorites, and airplanes is simplified.

Conclusions

A technique for combining phase channels in a compound interferometer has been demonstrated. In summary, some of the characteristics of the combiner are:

1. Gives a quick reading of the space angle
2. Releases data-line channels for other data
3. Helps identify false targets, such as meteorites and airplanes
4. Is more sensitive to phase quieting than the human eye and less tedious to read
5. Gives a "line" readout rather than a point; i.e., the space angle is resolved over the interval of time that the satellite signal is under surveillance
6. Operates in real time
7. Additional channels can be readily combined.

Work is underway to combine more of the higher resolution phase channels in the E-W set, as well as those in the N-S set. Some work has also been done towards a vector resolution of both sets, which if accomplished successfully will give the satellite's direction. An important by-product of the above will be phase-rate instrumentation, which can be used to define a relative-velocity vector of objects passing through the detection system.

Since the completion of the first combiner, several avenues have opened up which are leading to new designs. These will be described in future reports.

Acknowledgements

For the support and advice of my supervisor, Mr. R. L. Easton, head of the Space Surveillance Branch, the technical council of Mr. L. O. Hayden and the excellent system engineering of Mr. C. H. Weaver, the author is particularly indebted.

References

1. Easton, R. L. and J. J. Fleming, "The Navy Space Surveillance System," Proc. IRE 48:663-669 (1960).
2. Kaufman, M. G. and F. X. Downey, "Data Transmission for the NRL Space Surveillance System," NRL Report 5522, August 1960.

*Widening of the antenna beam at great distances, causing the phase rate to reduce for a given satellite velocity.

Appendix A

Derivation of Phase and Phase Rate

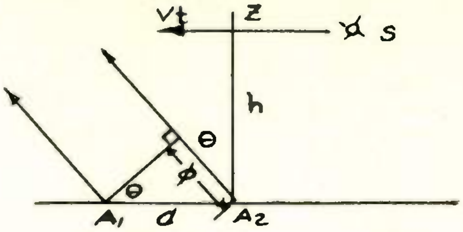


Fig. A1

From Fig. A1:

$$\Theta = \tan^{-1}\left(\frac{Vt}{h}\right) \qquad \text{(A1)}$$

$$\Theta = \sin^{-1}\left(\frac{\phi}{d}\right) \qquad \text{(A2)}$$

where t = time
Θ = space angle $(0 < \Theta < \frac{rr}{2})$
ϕ = elec. phase angle
V = velocity

Combining Equations (A1) and (A2),

$$\sin^{-1}\left(\frac{\phi}{d}\right) = \tan^{-1}\left(\frac{Vt}{h}\right)$$

$$\frac{\phi}{d} = \sin\left[\tan^{-1}\left(\frac{Vt}{h}\right)\right]$$

$$\phi = d\sin\left[\tan^{-1}\left(\frac{Vt}{h}\right)\right]$$

(used in graphing ϕ, see Fig. A2)

Let $\phi = d \sin\alpha$

$$\dot\phi = d\cos\alpha(\dot\alpha) \qquad \text{(A3)}$$

Since $\alpha = \tan^{-1}\left(\frac{Vt}{h}\right)$

$$\dot\alpha = \frac{1}{1+\left(\frac{Vt}{h}\right)}\left(\frac{V}{h}\right)$$

Putting α and $\dot\alpha$ into Equation (A3),

$$\dot\phi = d\cos\left[\tan^{-1}\left(\frac{Vt}{h}\right)\right]\left\{\left[\frac{1}{1+\left(\frac{Vt}{h}\right)^2}\right]\left(\frac{V}{h}\right)\right\}$$

$$\phi = \frac{dV}{h}\left[\frac{1}{1+\left(\frac{Vt}{h}\right)^2}\right]\cos\left[\tan^{-1}\left(\frac{Vt}{h}\right)\right]$$

Let d, V, and h = 1

$$\dot\phi = \left(\frac{1}{1+t^2}\right)\cos(\tan^{-1} t)$$

(used in graphing $\dot\phi$, see Fig. A2)

Appendix B

Determining the Angle of Crossing Through the Fence

Antennas with baselines in quadrature but patterns superimposed are shown in Fig. B1 and B2.

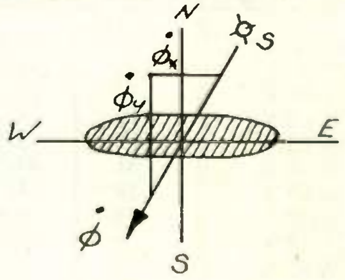


Fig. B1

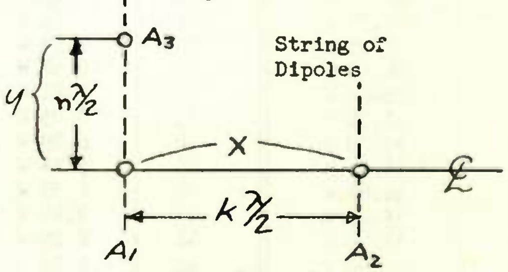


Fig. B2

From vector diagram, Fig. B3

$$\dot\phi_x = \dot\phi \sin\psi$$

$$\dot\phi_y = \dot\phi \cos\psi$$

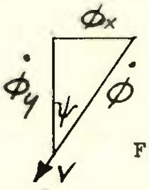


Fig. B3

$$\tan\psi = \frac{\dot\phi\sin\psi}{\phi\sin\psi} = \frac{\dot\phi_x}{\phi_y}$$

Angle of crossing,

$$\psi = \tan^{-1}\left|\frac{\dot\phi_x}{\phi_y}\right|$$

valid when equal length baselines are involved, (n = k), or if the rates are corrected for unequal lengths. The assumption is made that the earth's velocity vector is small. Instrumentation is shown in Fig. B4.

TABLE I

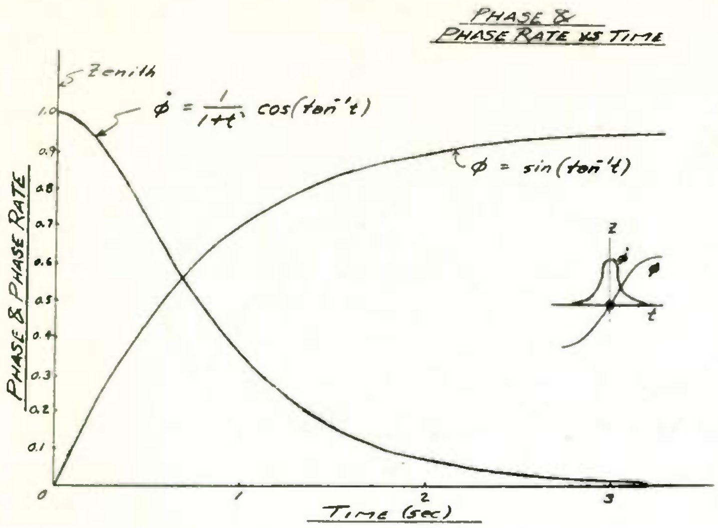
SPACE ANGLE vs. PERCENT DEFLECTION

%	Space Angle (West)	%	Space Angle (West	%	Space Angle (East)	%	Space Angle (East)
0	90.00 W	26	28.69 W	51	1.15 E	76	31.33 E
1	78.52 W	27	27.39 W	52	2.29 E	77	32.68 E
2	73.74 W	28	26.10 W	53	3.44 E	78	34.06 E
3	70.05 W	29	24.83 W	54	4.59 E	79	35.45 E
4	66.93 W	30	23.58 W	55	5.74 E	80	36.87 E
5	64.16 W	31	22.33 W	56	6.89 E	81	38.32 E
6	61.64 W	32	21.10 W	57	8.05 E	82	39.79 E
7	59.32 W	33	19.88	58	9.21 E	83	41.30 E
8	57.14 W	34	18.66 W	59	10.37 E	84	42.84 E
9	55.08 W	35	17.46 W	60	11.54 E	85	44.43 E
10	53.13 W	36	16.26 W	61	12.71 E	86	46.05 E
11	51.26 W	37	15.07 W	62	13.89 E	87	47.73 E
12	49.46 W	38	13.89 W	63	15.07 E	88	49.46 E
13	47.73 W	39	12.71 W	64	16.26 E	89	51.26 E
14	46.05 W	40	11.54 W	65	17.46 E	90	53.13 E
15	44.43 W	41	10.37 W	66	18.66 E	91	55.08 E
16	42.84 W	42	9.21 W	67	19.88 E	92	57.14 E
17	41.30 W	43	8.05 W	68	21.10 E	93	59.32 E
18	39.79 W	44	6.89 W	69	22.33 E	94	61.64 E
19	38.32 W	45	5.74 W	70	23.58 E	95	64.16 E
20	36.87 W	46	4.59 W	71	24.83 E	96	66.93 E
21	35.45 W	47	3.44 W	72	26.10 E	97	70.05 E
22	34.06 W	48	2.29 W	73	27.39 E	98	73.74 E
23	32.68 W	49	1.15 W	74	28.69 E	99	78.52 E
24	31.33 W	50	0.00 Z	75	30.00 E	100	90.00 E
25	30.00 W	Zenith					

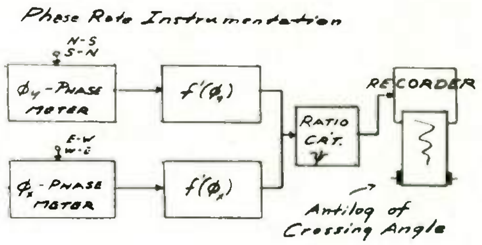
TABLE I-B

Table II, Phase-Rate Ratios vs. Angle ψ of Crossing

Ratio	Angle	Ratio	Angle	Ratio	Angle
0	0°				
0.01746	1°	0.60086	31°	1.8040	61°
0.03492	2°	0.62487	32°	1.8807	62°
0.05241	3°	0.64941	33°	1.9626	63°
0.06993	4°	0.67451	34°	2.0503	64°
0.08749	5°	0.70021	35°	2.1445	65°
0.10510	6°	0.72654	36°	2.2460	66°
0.12278	7°	0.75355	37°	2.3559	67°
0.14054	8°	0.78129	38°	2.4751	68°
0.15638	9°	0.80978	39°	2.6051	69°
0.17633	10°	0.83910	40°	2.7475	70°
0.19438	11°	0.86929	41°	2.9042	71°
0.21256	12°	0.90040	42°	3.0777	72°
0.23087	13°	0.93252	43°	3.2709	73°
0.24933	14°	0.96569	44°	3.4874	74°
0.26795	15°	1.00	45°	3.7321	75°
0.28675	16°	1.0355	46°	4.0108	76°
0.30573	17°	1.0724	47°	4.3315	77°
0.32492	18°	1.1106	48°	4.7046	78°
0.34433	19°	1.1504	49°	5.1446	79°
0.36397	20°	1.1918	50°	5.6713	80°
0.38386	21°	1.2349	51°	6.3138	81°
0.40403	22°	1.2799	52°	7.1154	82°
0.42447	23°	1.3270	53°	8.1443	83°
0.44523	24°	1.3764	54°	9.5144	84°
0.46631	25°	1.4281	55°	11.430	85°
0.48733	26°	1.4826	56°	14.310	86°
0.50953	27°	1.5399	57°	19.081	87°
0.53171	28°	1.6003	58°	28.636	88°
0.55431	29°	1.6643	59°	57.290	89°
0.57735	30°	1.7321	60°	∞	90°



A2



B4

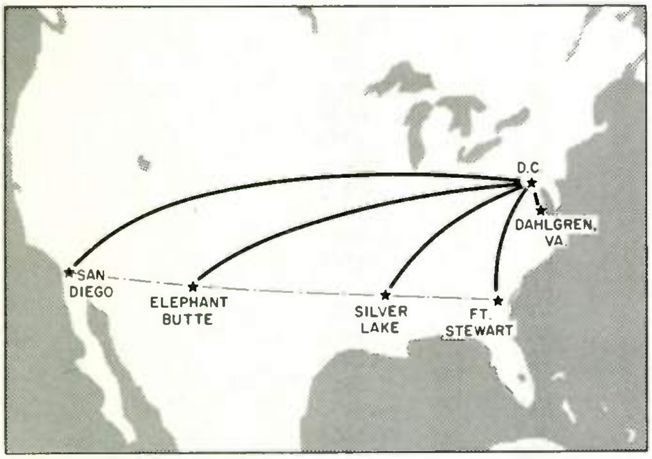


Fig. 1. Location of the space surveillance receiving stations showing the data-transmission lines to Washington, D.C. and Dahlgren, Va.

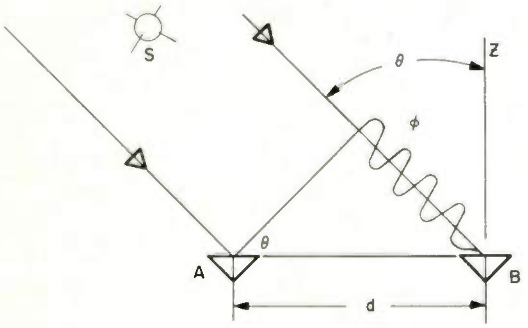


Fig. 2. Basic geometry involved in the space surveillance system with respect to a pair of receiving antennas.

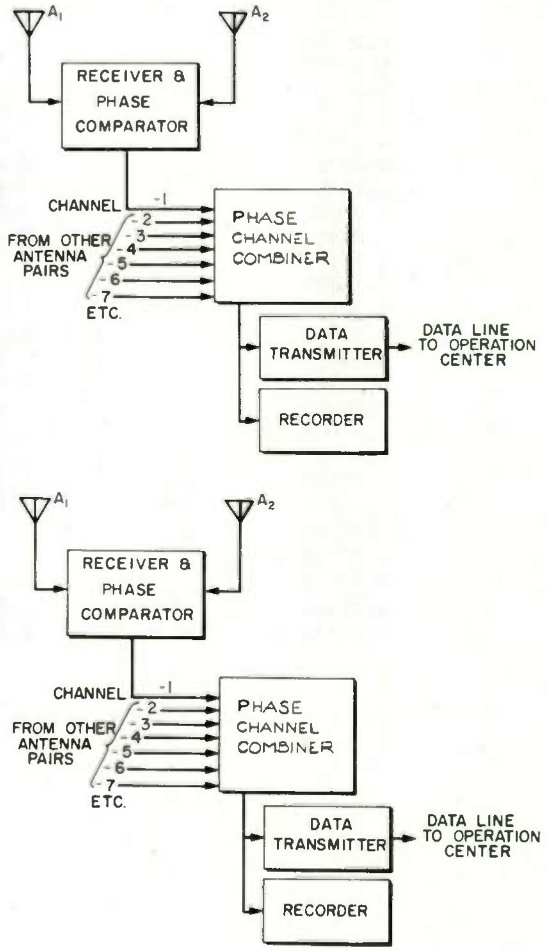


Fig. 3. Sketch showing how the phase channel combiner fits into the space surveillance system, (a) original system (b) original system with phase channel combiner added.

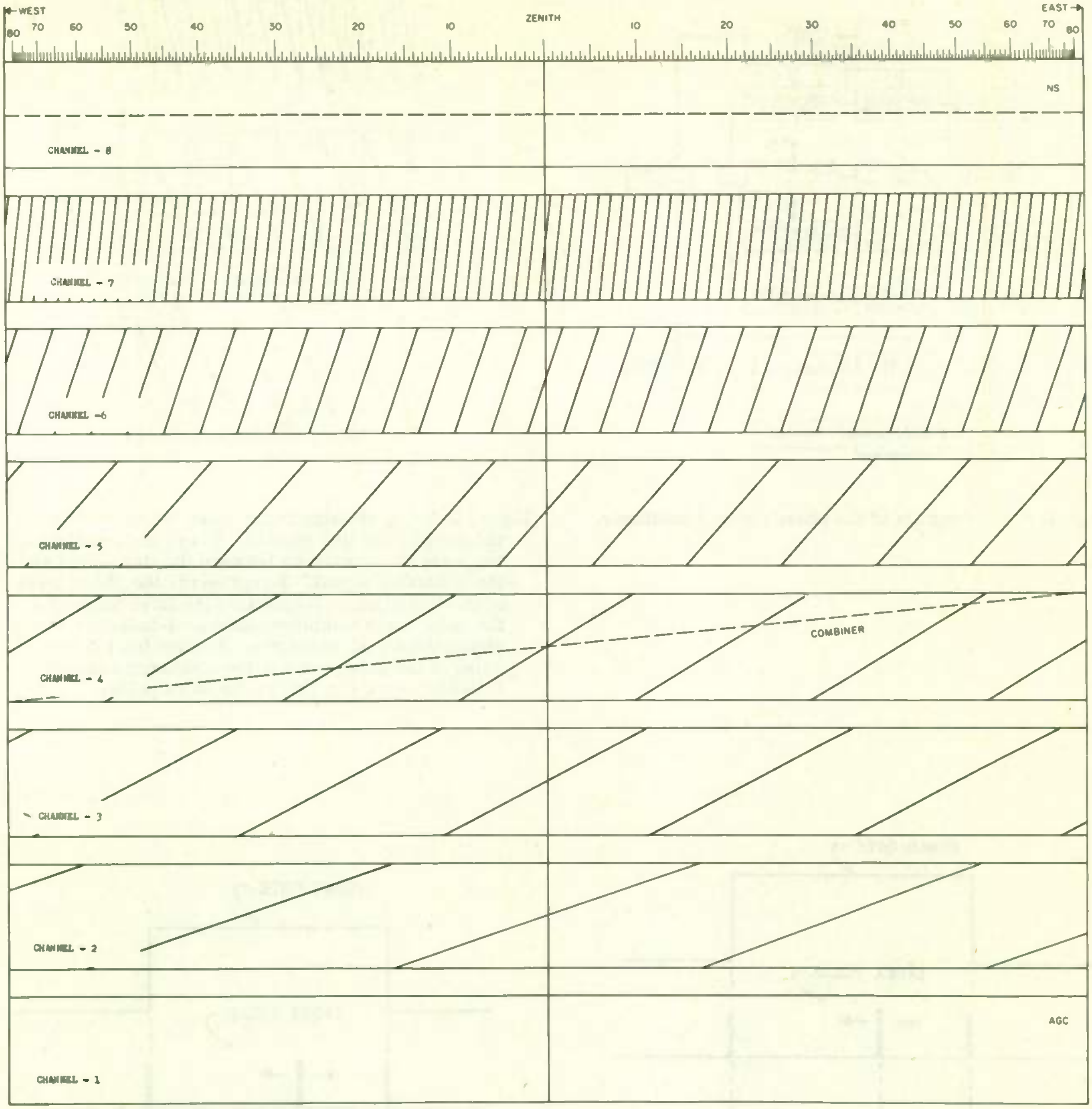


Fig. 4. Drawing showing a frame of simulated signal waveforms to illustrate the phase relationship between channels.

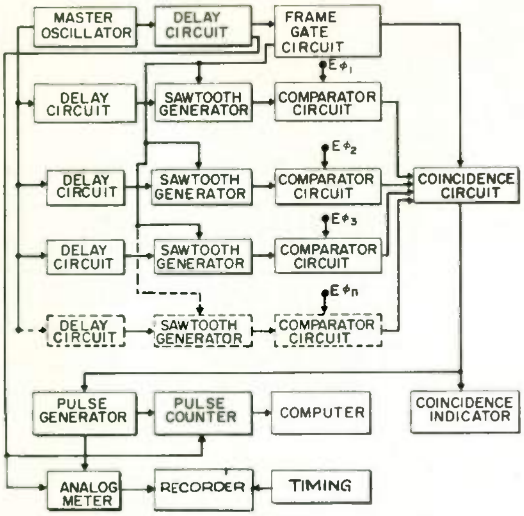


Fig. 5. Block diagram of the phase channel combiner.

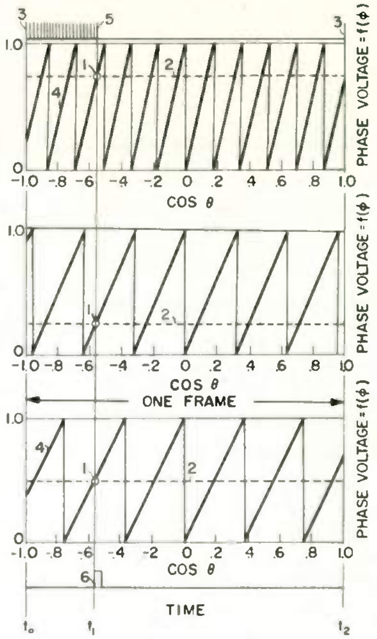
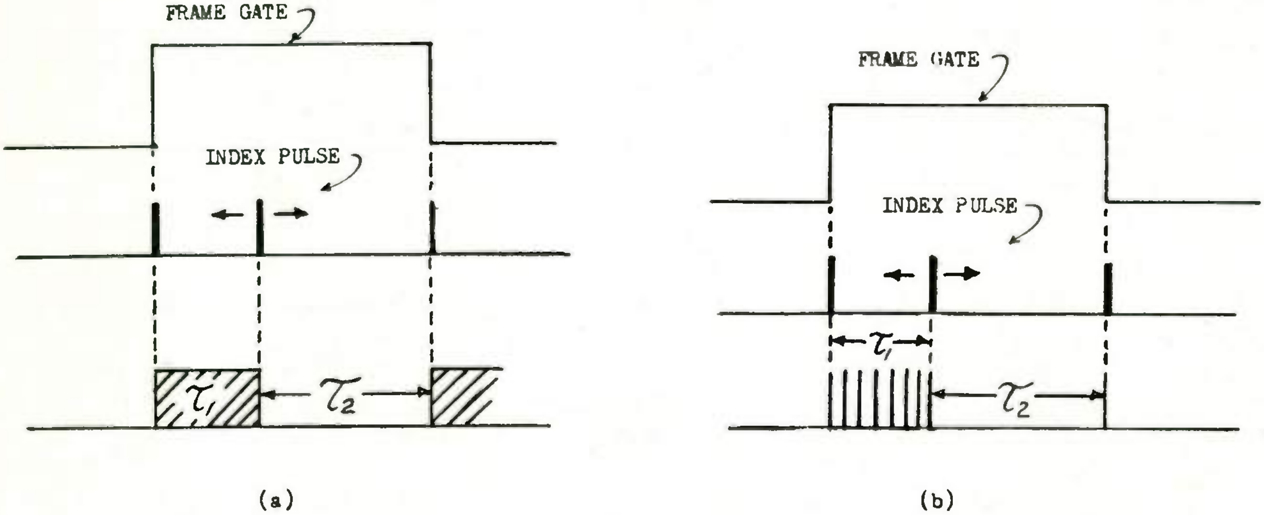


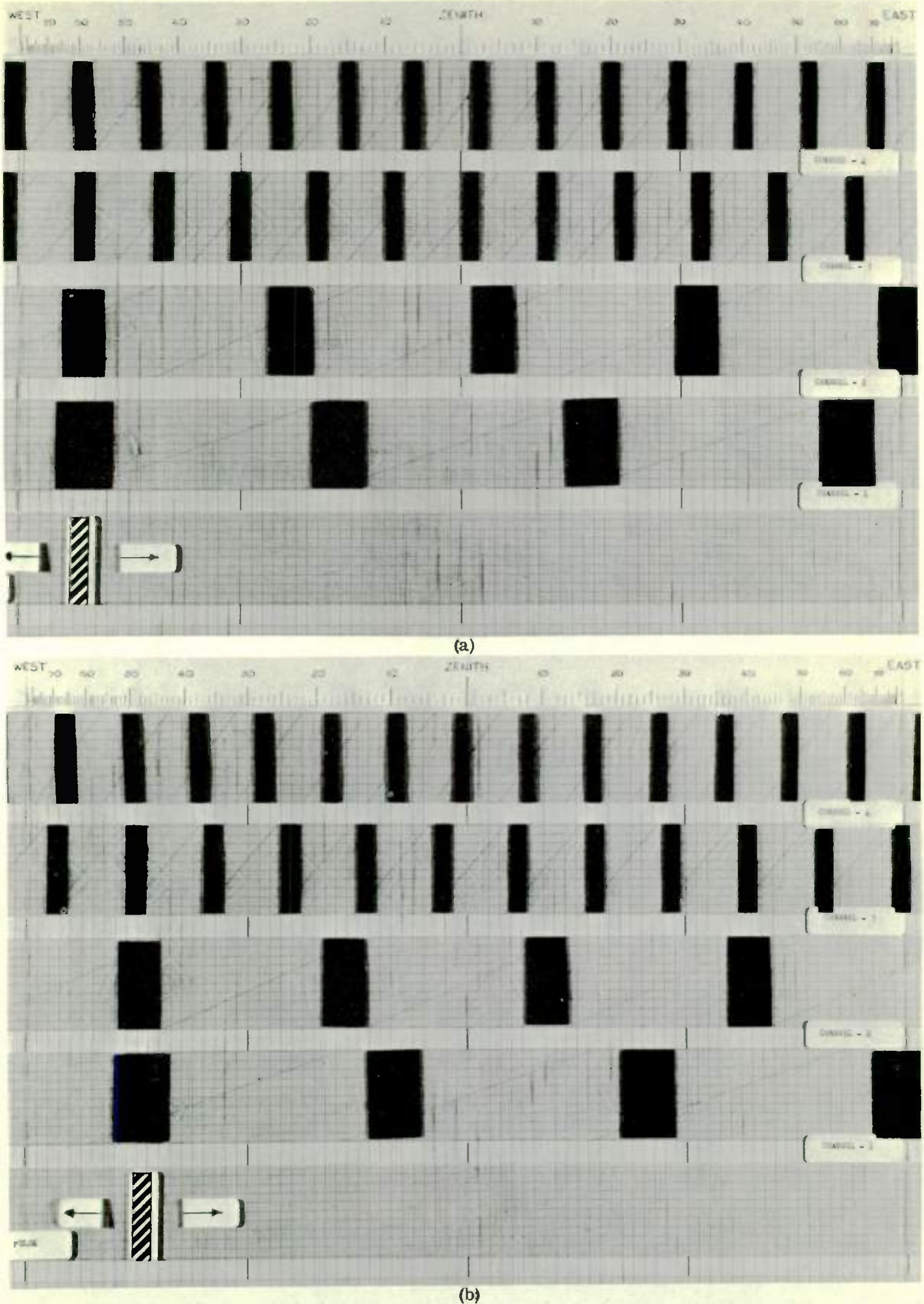
Fig. 6. Drawing showing coincidence between simulated signals and live signals. Wherein: 1-indicates the point of coincidence between the live signal and the simulated signal. 2-represents the phase level of the live signal. 3-indicates the first pulse of the pulse train combiner output. 4-indicates the simulated signal waveform. 5-indicates the last pulse of the pulse train in the combiner output. 6-indicates the position of the index pulse.



(a)

(b)

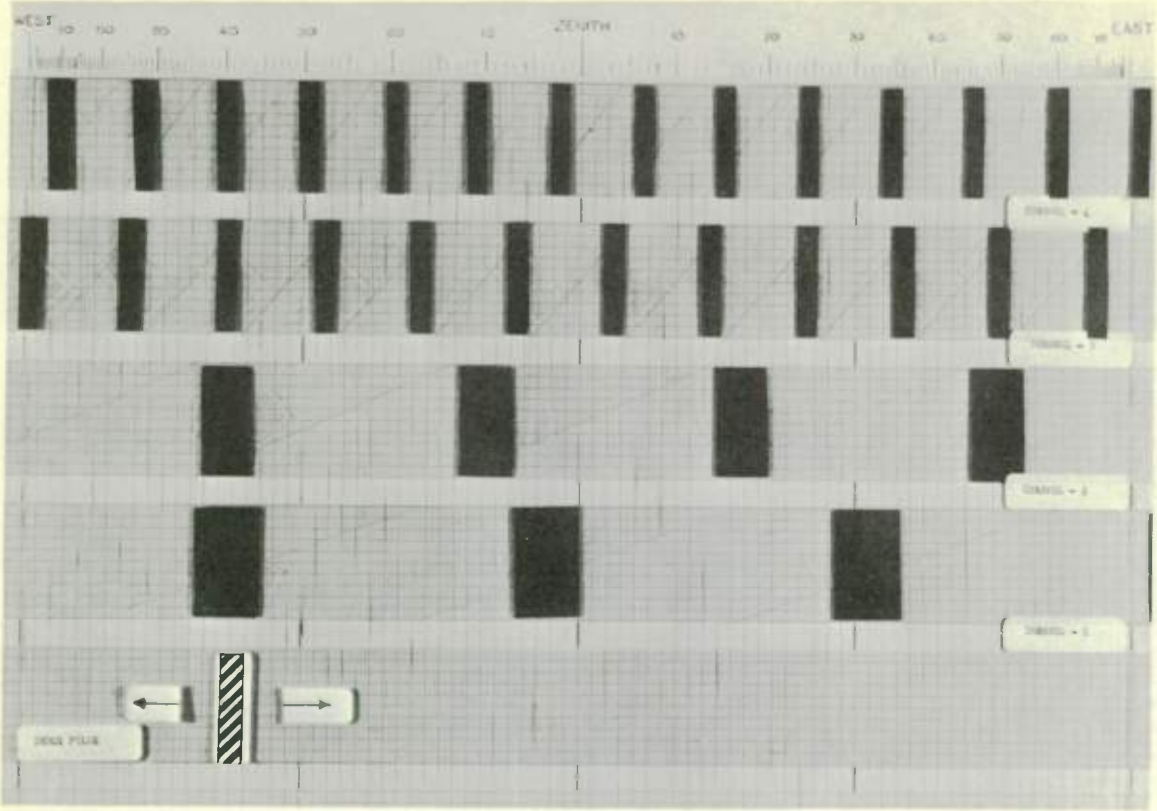
Fig. 7. Waveforms pertaining to the two combiner outputs (a) the analog phase meter and (b) the pulse count.



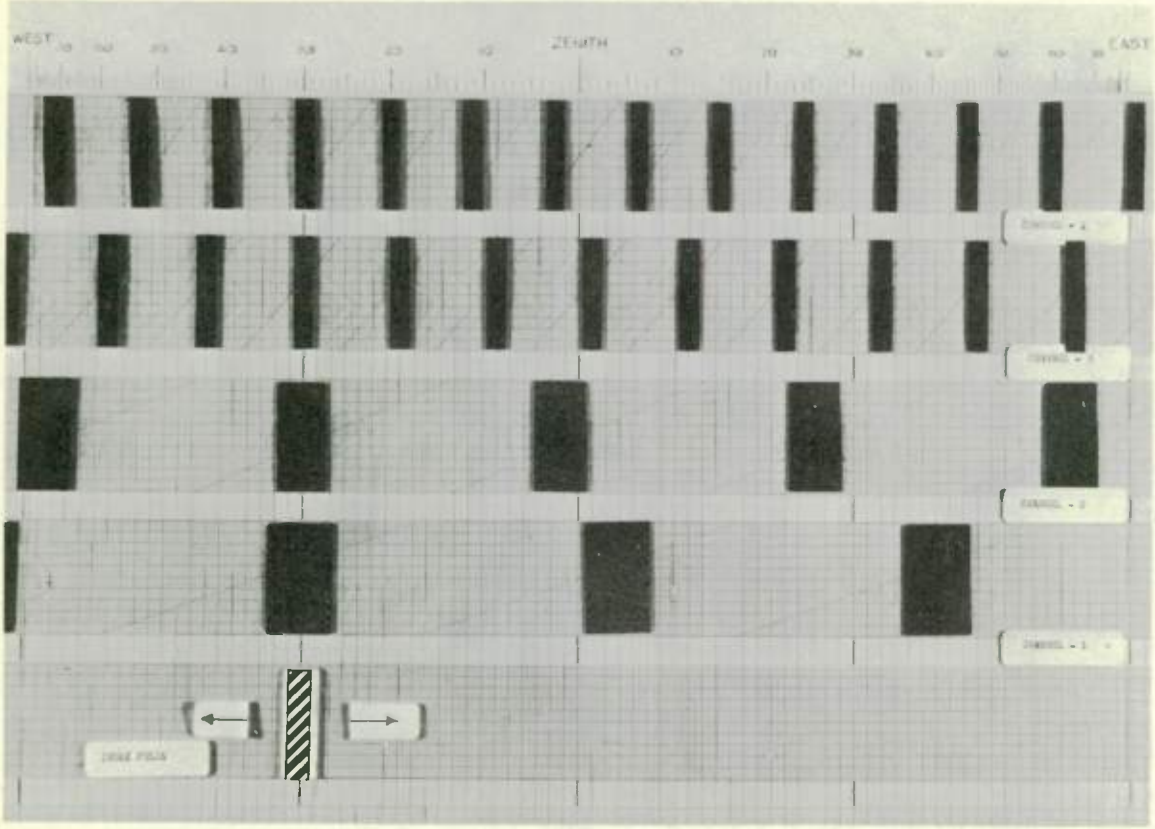
(a)

(b)

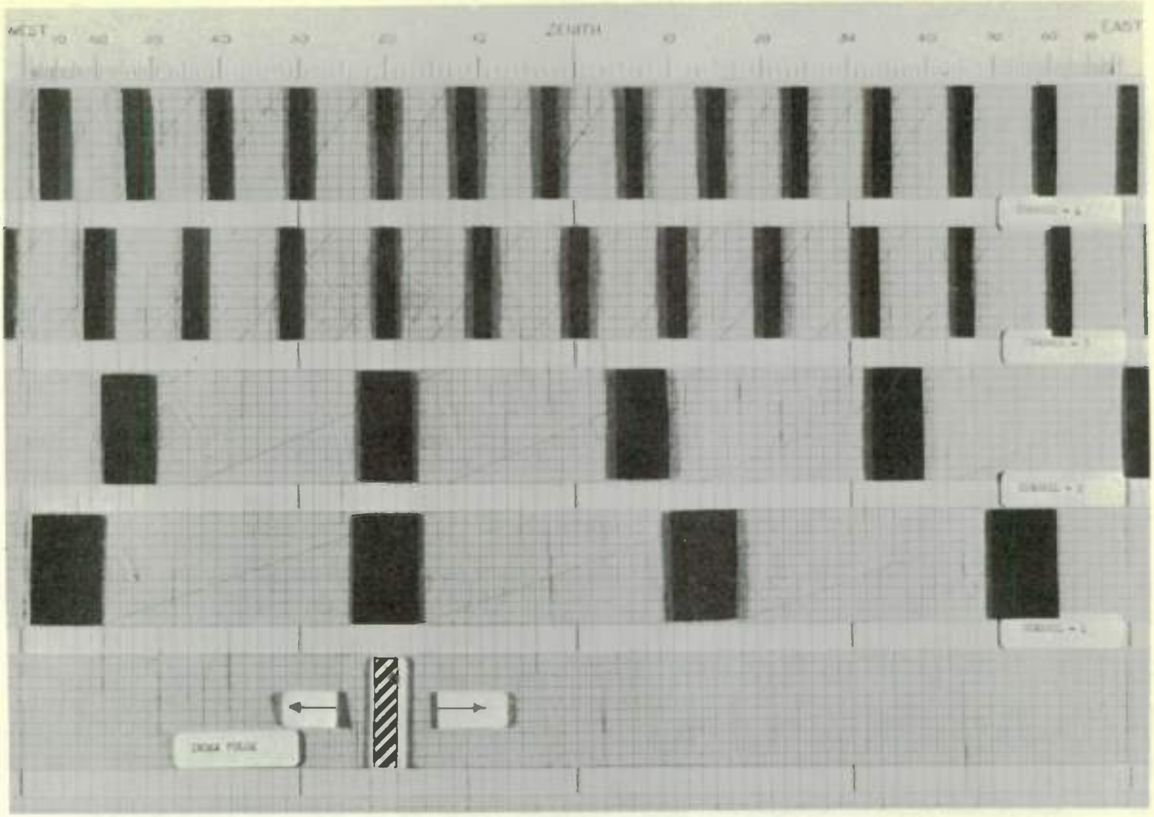
Fig. 8. Photographs of a plexiglass model used to demonstrate the action of the coincidence pulses. Coincidence is shown on figure (a) at 60 degrees west of zenith. Figures (b) through (g) show coincidences at intervals of ten degrees. Similar action occurs for signals arriving east of zenith.


WEST
70
60
50
40
30
20
10
ZENITH
10
20
30
40
50
60
70
EAST
CHANNEL = 4
CHANNEL = 3
CHANNEL = 2
CHANNEL = 1

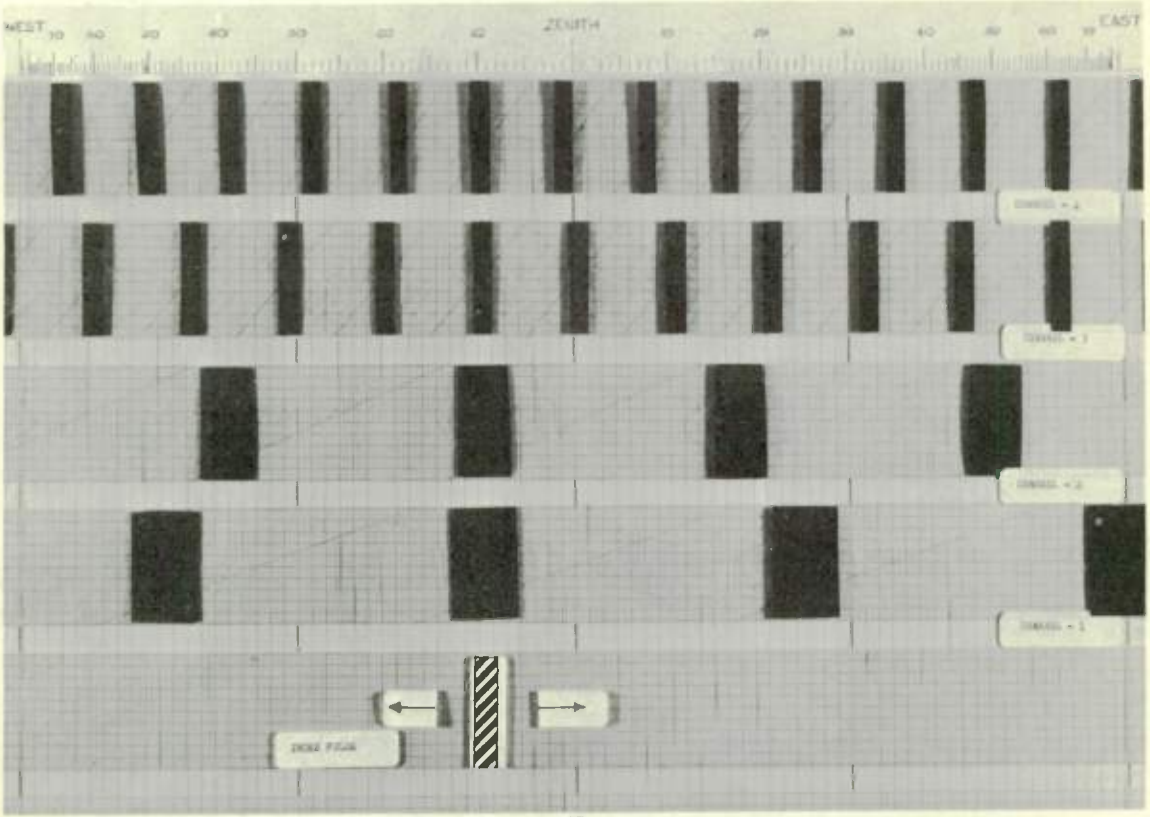
(c)


WEST
70
60
50
40
30
20
10
ZENITH
10
20
30
40
50
60
70
EAST
CHANNEL = 4
CHANNEL = 3
CHANNEL = 2
CHANNEL = 1

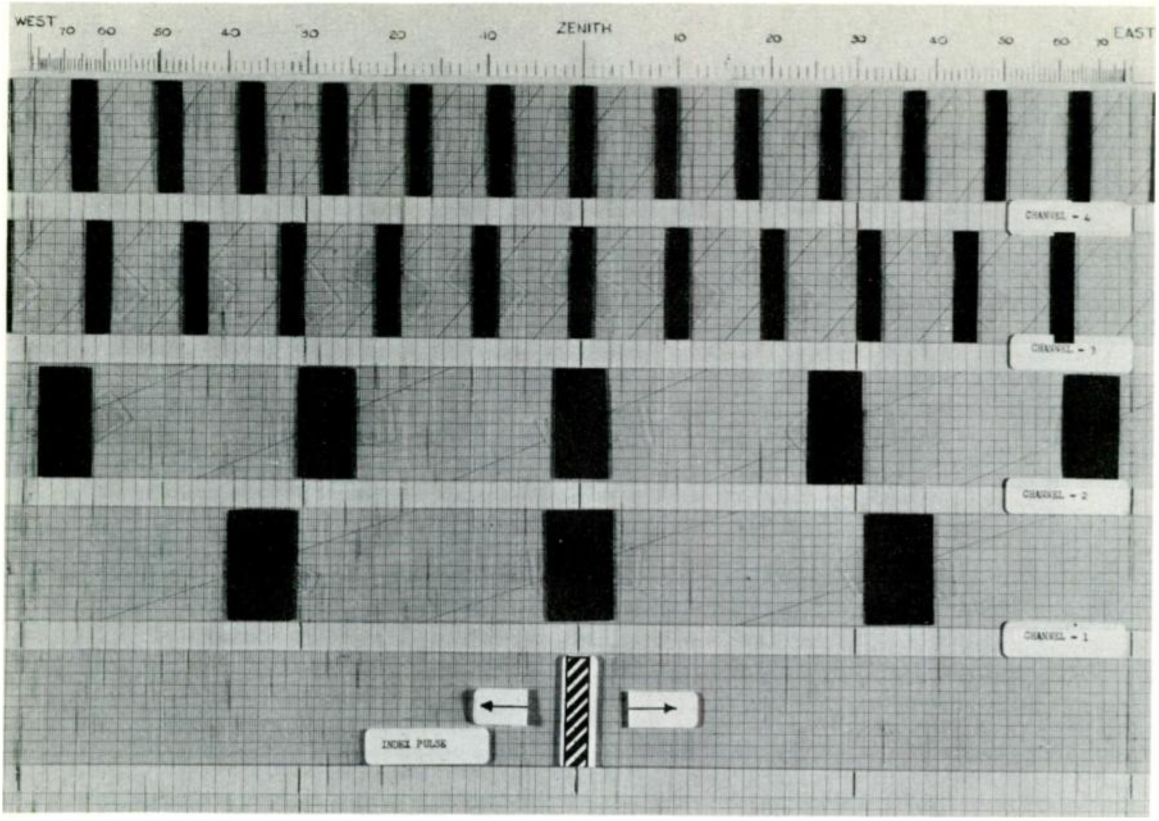
(d)


WEST
70
60
50
40
30
20
10
ZENITH
10
20
30
40
50
60
70
EAST
CHANNEL = 4
CHANNEL = 3
CHANNEL = 2
CHANNEL = 1

(e)


WEST
70
60
50
40
30
20
10
ZENITH
10
20
30
40
50
60
70
EAST
CHANNEL = 4
CHANNEL = 3
CHANNEL = 2
CHANNEL = 1

(f)



(g)

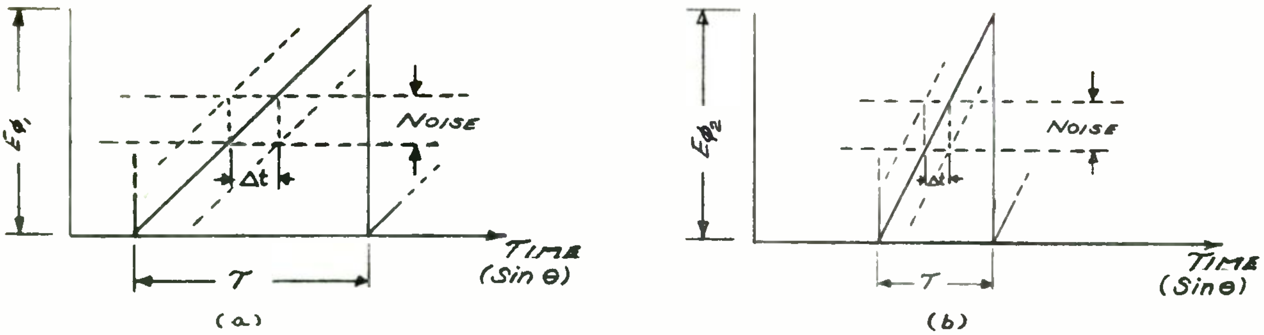


Fig. 9. Sketch of two sawtooth signal waveforms derived from baselines having lengths in ratio of 1 to 2.

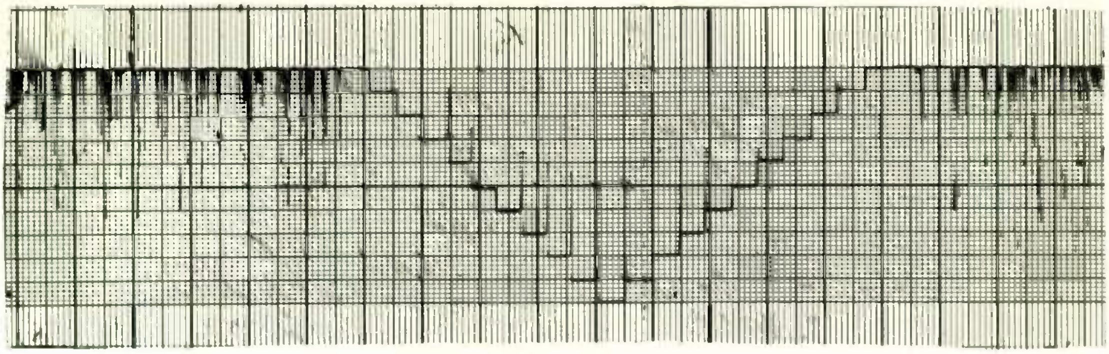


Fig. 10. Sample recording of the combiner calibration waveform.

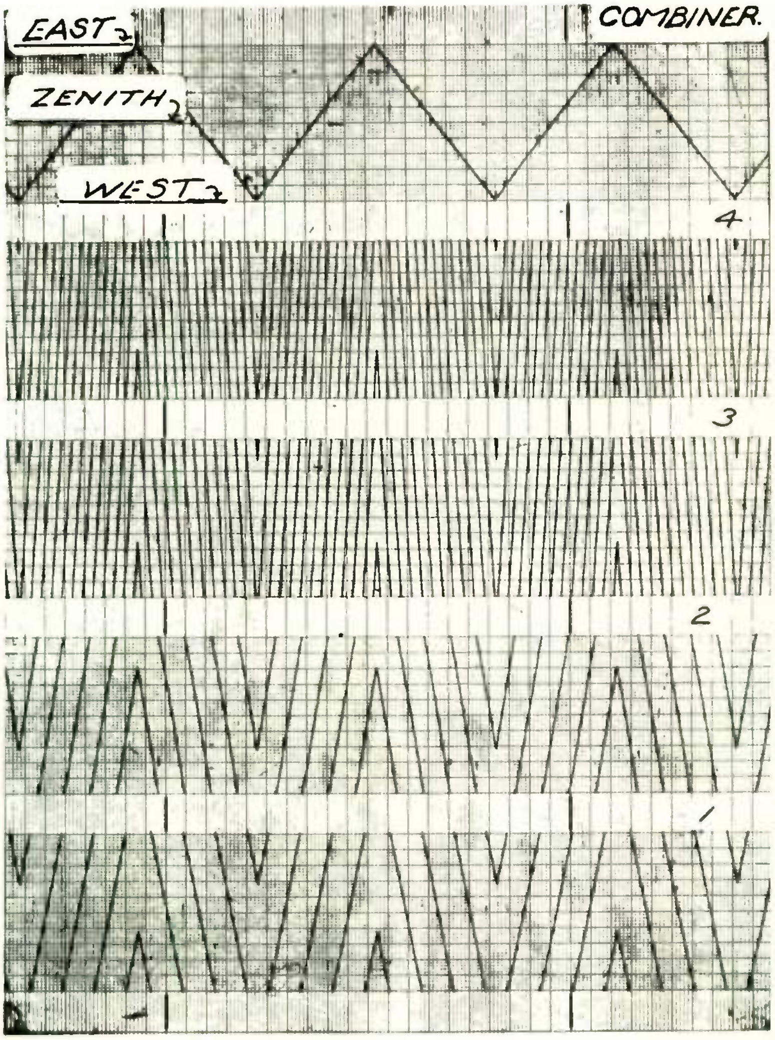


Fig. 11. Recording showing waveforms generated by a laboratory signal simulator.

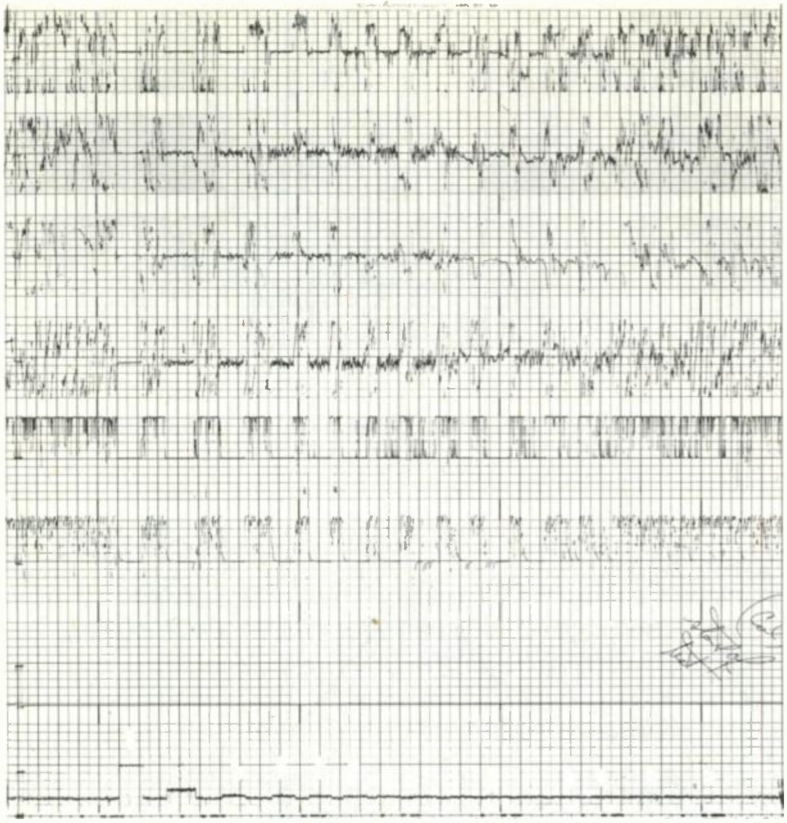
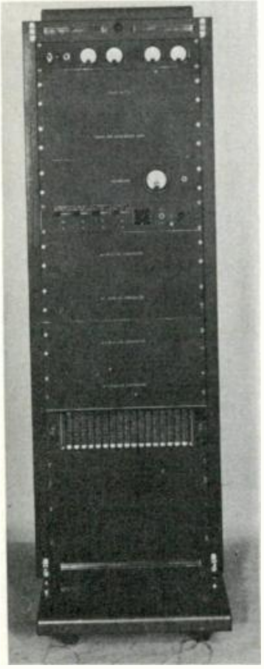


Fig. 12. Recording of sensitivity test. The relative signal strength is monitored by the agc voltage 1 level, channel 1. The combiner output, recorded on channel 4 shows a zenith reading (midscale) for each succeeding lower level of agc, until the signal disappears into the noise. Channel 3, 5, 6 and 7 show the corresponding phase measuring channels. (Channel 2 has no information.)



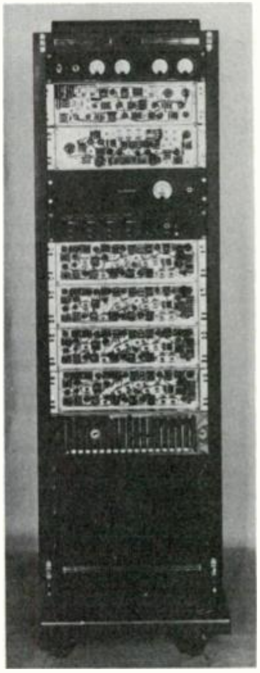


Fig. 13. Photographs of the combiner rack.

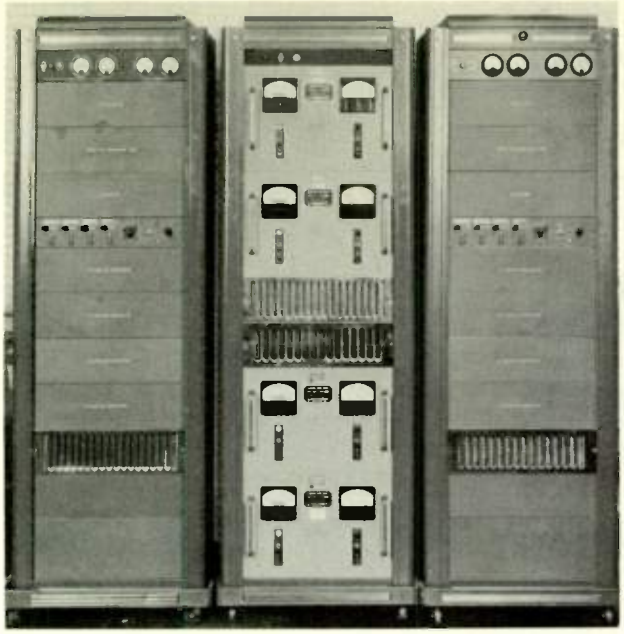


Fig. 14. Typical space surveillance station combiner installation, consists of two combiner racks and a power supply rack. One combiner is on standby for the other.

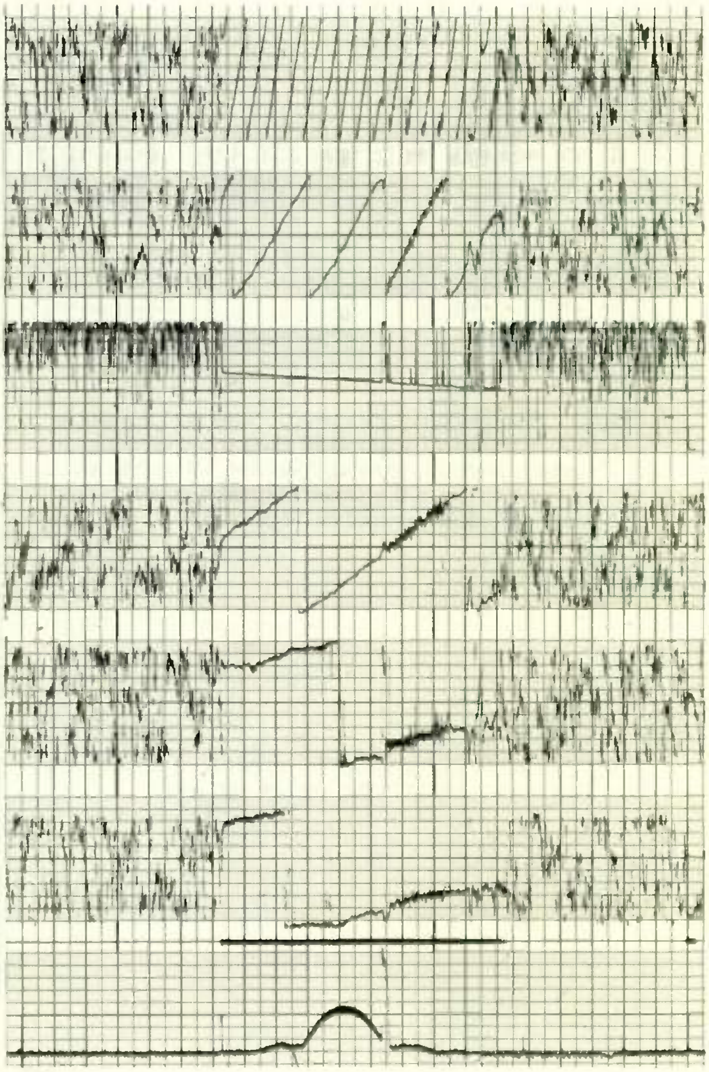


Fig. 15. Ordinary bounce signal from a satellite passing station zenith at 12.71 degrees east. Combiner output on channel 5 shows improved signal to noise over original phase signals, channels 2, 3, and 4.

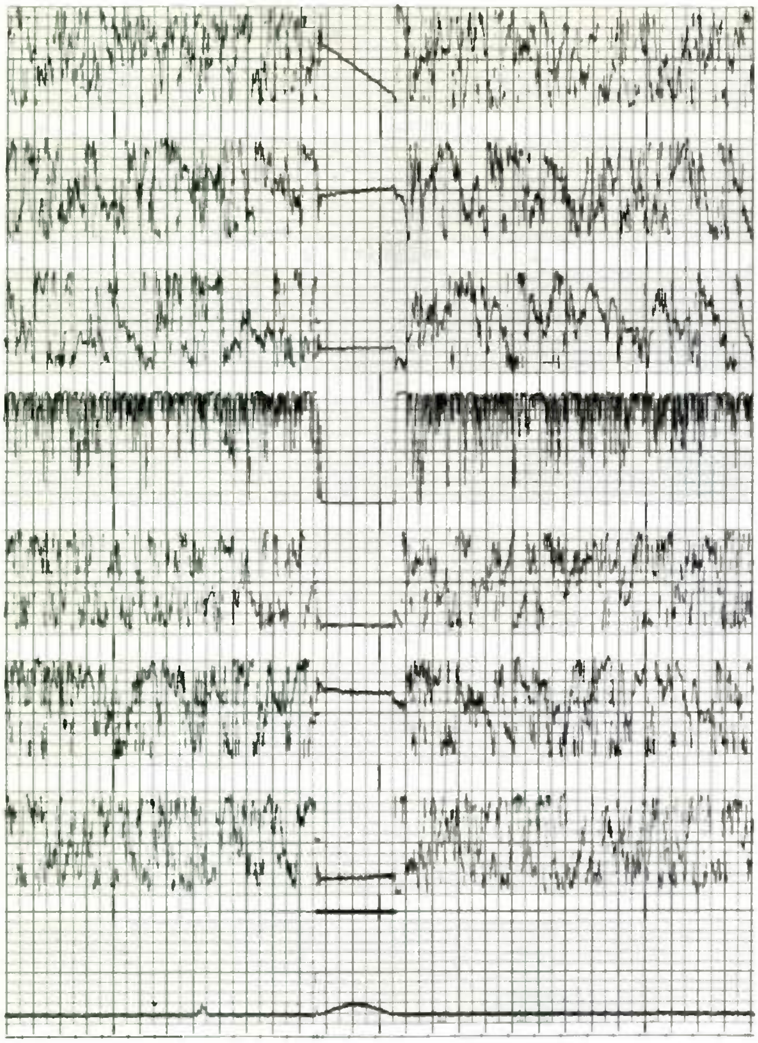
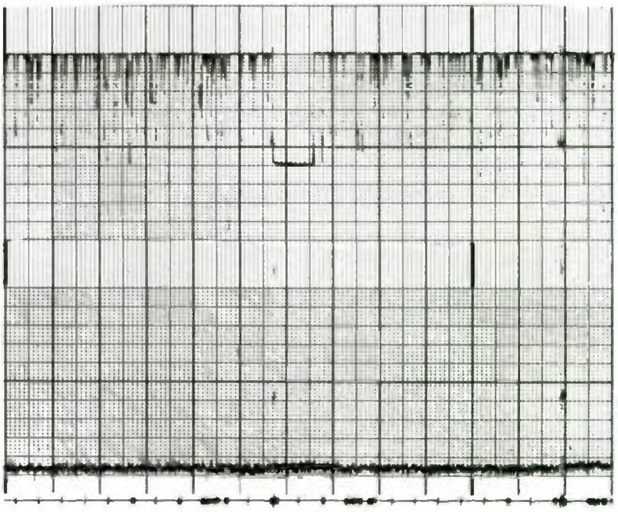


Fig. 16. Typical airplane signal. Note that the phase rate is practically zero and the target is on the western horizon, passing from the NW to the SE as determined from the phase slopes of channels 7 and 8 and the diagrams on Figure 21.

Fig. 17. Combiner action under conditions of very low signal strength as noted by low agc level.



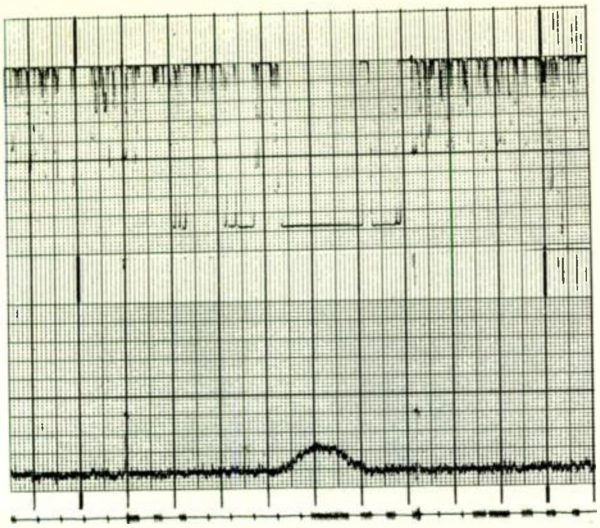


Fig. 18. Combiner action reading out well beyond each side of main agc lobe.

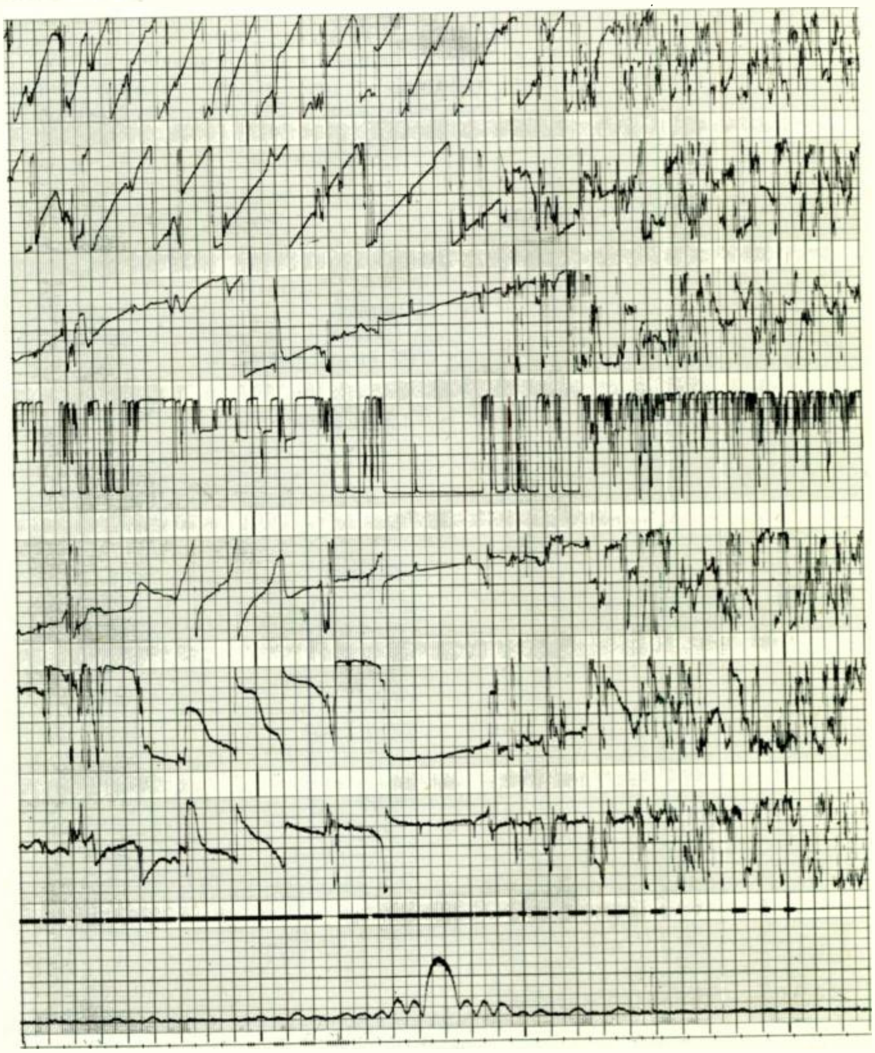


Fig. 19. A typical record of a radiating satellite denoted by the violent break-up of the sawtooth waveforms due to phase reversals probably caused by the satellite spin. The combiner output is recorded on channel 5. Note the consistent combiner readings through the antenna sidelobes, the latter are clearly seen on the agc trace, channel 1.

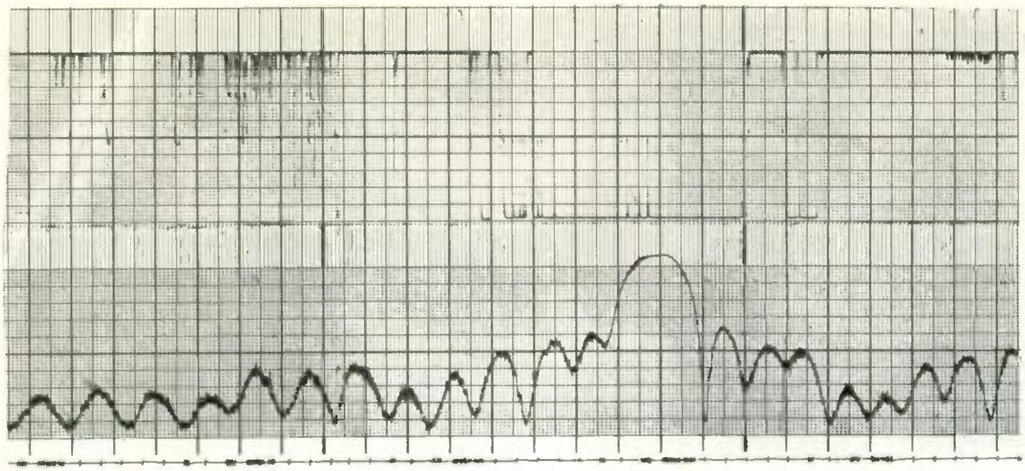


Fig. 20. A very strong bounce signal from a satellite. Agc on channel 1 and combiner output on channel 2. Combiner channel indicates satellite passed receiving site at 73.74 degrees west of zenith.

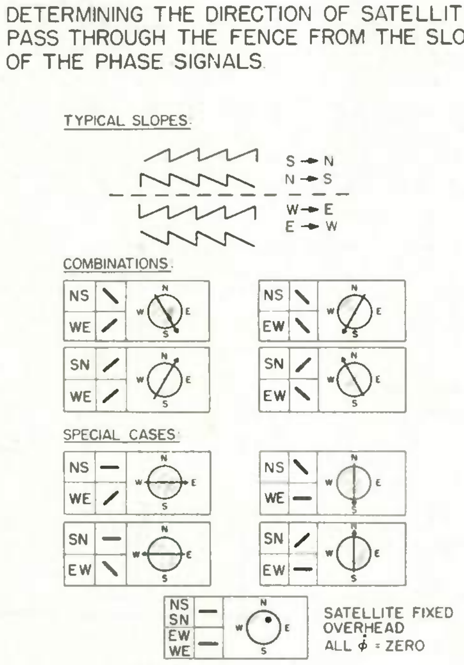


Fig. 21. Diagrams for determining the direction a satellite passed through the surveillance system.

THE EXPERIMENTAL DYNAMIC PROCESSOR DX-1

C.M. Walter
Electronics Research Directorate
Air Force Cambridge Research Laboratories
Bedford, Massachusetts

SUMMARY

A concept is outlined which involves the use of a pair of interconnected, medium size, digital computers as the nucleus of an experimental dynamic data processing system. The primary objective is to achieve a processor configuration which supports not only an extensive research program on various methodologies of on-line dynamic data processing but which also permits the alterations necessary to try new real-time data processing techniques, without inhibiting the continuity of the methodological investigations. This goal is achieved through a three phase plan, in which extensive modification of only one of the dual central processors is permitted, at any one time, in order to implement new hardware techniques. Particular stress is placed on the use of the central digital processors for exercising logical control over electro-optical and other highly parallel analog processing media.

INTRODUCTION

The concept underlying the Experimental Dynamic Processor DX-1 is an outgrowth of extensive researches in the surveillance processes area at AFCRL over the past four years. This concept arose from the necessity for treating the entire surveillance data processing domain from a more unified point of view.

A great variety of fundamental measurement and interpretive processes are involved in this area. At one end of the process spectrum great quantities of raw sensor data are involved, with little significance to be attached to individual items of the data. From these data, through appropriate transformations, patterns of significant information must be extracted, and interpreted. This process leads progressively toward smaller amounts of highly evaluated information, ordered in importance relative to various criteria for judgement, from which increasingly more important decisions must be made.

Clearly the most elementary research into basic processes covering such a wide spectrum, calls for an extremely flexible dynamic processor capable of evolving with the research effort itself. The requirement of flexibility dictates that the central processor be a programmed digital processor and controller. This can effectively handle extensive decision making operations and manipulate a reasonable amount of digital information. The problem of filtering and extracting significant information from truly staggering amounts of sensor data, however, dictates the use of flexible analog processing devices operating under the general control of the digital machine.

Since no commercially available processors satisfy all of the necessary requirements, it has been necessary to specify the more general requirements and to guide the development of a systems configuration which can be effectively used as a research tool in a rapidly changing environment.

THE GENERAL PROCESSOR CONFIGURATION

System Configuration

The overall systems configuration is indicated in Figure 1. The Central Digital Processors are modified Digital Equipment Corporation Programmed Data Processors (PDP-1's). The general organization within these processors is indicated in Figure 2. Extensive provision has been made for maximum flexibility in rearranging the various peripheral devices into configurations best suited for the study of particular classes of problems.

Operating Characteristics

Some of the more important characteristics of the basic central processing units and of certain of the peripheral devices are outlined under the following headings:

General Physical Features.
Control Console and Central Processor: 2'w x 7'l x 6'h (one for each PDP-1);
On-line Typewriter Console: 2'w x 2'l x 3'h (at least one per PDP-1);
Digital Magnetic Tape Units: 2' x 2' x 6' (four units on initial system);
Color Display Scope: 3'w x 5'l x 4'h;
Standard Scope: 3'w x 4'l x 4'h;
Precision Photographic Scope: 3'w x 7'l x 4'h.

General Electrical Features.
Solid state, 5 mc logic circuitry;
Built in marginal checking to facilitate maintenance;
Random access core storage, 18 bit words, 4096 word modules, 5 μ sec. cycle time, field switching (three modules on initial system, maximum of 16 modules on each PDP-1).

General Logical and Arithmetic Features.[1]
Single address, single instruction, 18 bit parallel machine;
Fixed point add time, 10 μ sec;
Fixed point built in multiply time, 25 μ sec,, divide time, 40 μ sec.;

Memory reference instruction execution time, 10 μ sec.;
Non-memory reference instruction time, 5 μ sec.

Other Logical Features.
One unit indexing and branching on contents of any memory location in 10 μ sec.;
Indirect addressing (single level mode for referencing any of 16 possible core modules, multi-level mode for multiple indirect addressing within any module);
Microprogramming on operate instructions;
Sequence break for concurrent operation of several in-out devices and main program sequence;
inter-machine buffer system for transfer of data between central processing units.

General Input-Output Features.
Paper tape input-output for general program work;
Digital Magnetic Tape Units, compatible with IBM 729II tape format for large data handling and buffer storage (four tape units on initial system);
Automatic Tape Control for automatic transfer between main memory and tape in blocks of words in IBM format;
Programmed Tape Control for data transfer one character at a time, using any format;
Analog magnetic tape input through multiplexed A-D converter.

Man-Machine Communication Features.
Typewriter and console control for program monitoring;
Both color, black and white and photographic scope outputs for visual process monitoring and recording;
Light pencil for graphic communication with machine through display scopes;
Analog parameter adjustment capability using A-D converter and sequence break system.

OPERATING PHILOSOPHY AND DX-1 SYSTEM RATIONALE

The Dynamic Processor as an Integral Part of the Dynamic Data Processing Research Program

It is now possible to obtain reliable, and reasonably economical, digital processors which are tailored to specific requirements and yet have all the flexibility of general purpose machines. The Experimental Dynamic Processor Concept is based on using this sort of device as the nucleus of an experimental facility which can be uniquely fitted to the needs of a rapidly advancing research effort. The difficulty associated with the continual construction of one highly special piece of processing apparatus after another, to test various tentative hypotheses, can be minimized. Moreover, by limiting the domain of principal users to those faced with common categories of problems, the most effective intimate use of the facility, with a minimum of friction, can be assured. This is particularly true of problems requiring extensive on-line real time process control and graphic observation of various phases of the processing operation. Also, long range planning for updating the facility can be intelligently carried out when the parties involved have a common mutual appreciation of each others requirements, and interest in a common goal.

Flexibility vs. Continuity of Research: The Dual Central Processor Concept

In order to provide the maximum in flexibility and ease of facility updating with the minimum of inconvenience in interrupted research efforts, we have adopted the dual central processor arrangement as an integral part of the DX-1 concept.

Probably the most frustrating and difficult problem which faces a facility which is built around any kind of programmed data processor or digital computer, is that of when to make irreversible changes in machine logic. Such changes will generally invalidate large classes of programs and hence necessitate rewriting them or terminating the research effort which made use of them.

By starting with central processors which are basically compatible and which can share common memory and peripherial units, a very high degree of flexibility and adaptability can be achieved. The system can then be updated in phases, with very little interruption in the continuity of extended research efforts, and with a minimum of rewriting on non-utility programs.

Ease of Operation: Man-Machine Interaction

Following the experience gained on WHIRLWIND I, TX-0, TX-2, and on other machines which stressed the concept of close man-machine interaction, every effort has been made in the initial DX-1 configuration to provide the nucleus for an increasingly more effective capability in this direction.

In addition to the usual visual and photographic CRT displays with light pencil communication, a color CRT display for advanced attribute extraction and pattern recognition studies is being provided. The use of color provides an additional dimension in data display capability and has already proved of great utility in filter design studies which required overlaying graphic information corresponding to different ways of handling the same underlying process. Its potential for the description of complex processes is almost without limit.

In investigations into the nature of almost any physical process, an appropriate pictorial or graphic description is often literally worth many thousand words (or tables of data). This is particularly true when a systematic and effective means exists for transforming the graphic descriptors. At present, the necessity for going back and forth between graphic and analytic modes of description is very cumbersome. In the Section "Prospects For Processor Evolution", some of the prospects for devices which achieve a more effective synthesis of both graphic and analytic modes of description will be discussed.

A concerted programming effort is underway, in conjunction with other owners of the same basic digital machine, to evolve as rapidly as possible the basic utility and executive programs needed to make the system into an effective research tool. Great effort by all users is being put on flexible, problem oriented compilers and interpretive programs which make maximum use of the relatively unique man-machine communication capability of the central processor.

Through the dual central processors, with sequence interrupt features, a capability exists for extensive time sharing and program interleaving. In this context, however, it is vital to note that the time sharing problem is a very sophisticated one. An while it holds great potential for the future, it is more a subject for programming research than an effective working tool, at the present time. Our goal here is simply to make use of this capability whenever feasible, to enhance the operating effectiveness of the system. Obviously, any drastic alteration of either of the central processors will generally be incompatible with existing time sharing procedures involving both machines.

The Concept of Critical Size for an Experimental Processor

The minimal size for a dynamic processor facility of the above type is to some extent dictated by the amount of digital storage needed to operate utility programs and problem oriented (e.g., FORTRAN like-) languages which are adequate to take the burden of basic programming off the users. At present this is of the order of 4,000 words of high speed storage, backed up by a minimum of two magnetic tape units, or a drum, or disc file storage.

An even more important factor influencing the minimum amount of high speed storage is the requirement for implementing moderately intricate process control and graphic display procedures. The use of effective electro-optical storage and display media, when available, will greatly enhance the display generating capability of modest size digital processors.

The upper limit on high speed storage is determined by a number of factors, both scientific, economic and administrative. The scientific utility increases with size, but this is gradually offset by increasing administrative problems attendant upon the economic necessity of sharing the facility among an increasingly more diverse group of users. Each group then has increasingly less to say about the overall processor configuration. For some categories of problems, size and speed of the central digital processor is all important, and is worth buying, even at the price of having very little to say about the basic processor configuration.

The mode of operation envisaged in the DX-1 concept, however, in which the processor is an integral part of the very means for studying the underlying dynamic processes, requires that complete control of all aspects of the processor operation be in the hands of the research scientists. In this context, W. Clark and others at Lincoln Laboratories are evolving a very small, laboratory bench size, programmable processor, the LINC computer,[2] which can be used as the nucleus for flexible real time process control and data reduction operations needed to aid in laboratory experiments. Particular stress is being placed on its potential utility in biomedical research.

In the on-line dynamic data processing area, the proper balance between low significance analog processing elements, and the manner in which transition to increasingly higher significance digital processing is to be carried out, is a very delicate matter and usually requires much on-line trial and error operation. This mode of operation is usually completely incompatible with the efficient use of existing large scale computer facilities.

System Limitations

The short word length, with attendant limitations on instruction repertoire, and on size of storage directly addressable at one time, makes the DX-1 type of system somewhat ineffective for certain large scale logical processing efforts in advanced programming, artificial intelligence and information storage and retrieval systems. This system configuration is also somewhat inefficient for routine analytical work involving extensive manipulation of numbers requiring high precision. These are, however, precisely the domains which are usually best handled by a large scale central facility.

PROSPECTS FOR PROCESSOR EVOLUTION

With the processor in the proper limited environment, centered about the investigation of statistically corrupted measurement processes, involving sophisticated attribute extraction and interpretation problems, the following reasonably clear phases of improvement can be delineated:

Phase 1: Primarily Digital Operation. With the DX-1 configuration essentially as specified in the "Introduction", programs are being written to carry out the simulation and evaluation of a variety of attribute extraction procedures involving close graphic monitoring of all phases of the on-line processing operation. These procedures will be applied to specific categories of recorded surveillance data, and to many other types of statistically corrupted data, having both known and unknown attributes, to assess their signal extraction capabilities.

Large quantities of low precision measurement data from a wide variety of sensors, both ground based and satellite based, are being recorded and fall precisely in the above category. The use of appropriate displays and light pencil communication, for purpose of altering and improving the attribute extraction process at various intermediate stages of the operation, are essential to the rapid and successful investigation of the data.

We are now in an era in which the data collection capability of the primary sensors often far exceeds the capabilities of the apparatus which can be assigned full time to carrying out the processing operation. The relative costs are also often completely incommensurate.

Hardware for Phase 1 of the DX-1 system is now being debugged and will be in operation by April 1962. Figures 3 and 4 illustrate the compact nature of the system.

Phase 2: Logical Control of Electronic Analog Processing Elements. In order to rapidly filter and examine very large amounts of low precision data, a requirement exists for large, very rapid access storage and for very wide band multipliers. Since these requirements can be satisfied by low precision devices, the use of analog media, such as single and double ended storage tubes, looks feasible here, at least as an interim measure.

Means are under investigation fcr having the central digital processor simply act as a highly flexible programmable controller for gating low precision data on and off the storage media and channeling it through appropriate wide band analog arithmetic devices. Of particular interest here, is the "pulse analog" concept[3] evolved by M. Connelly and others at the Electronic Systems Laboratory, MIT, and implemented on the TX-O computer. Other efforts are underway, in this vein, at NBS, IBM, and elsewhere.

The extent to which further development is desirable along the lines of lumped parameter analog devices, under the control of digital processors, hinges heavily on the speed with which media cited in Phase 3 can be evolved.

Phase 3: Logical Control of Electro-optical Processing Elements. Through the use of electro-optical techniques, which show great promise for highly parallel processing and for the manipulation of very large stochastic matrices, the possibility exists for the real time investigation of many dynamic processes utilizing very sophisticated processing methodologies.

Suitable electro-optical devices are just beginning to evolve from basic materials research. The principal bottleneck is in the development of rapidly variable optical transmitivity film media, in which the transmitivity can be varied either electronically, or by photochromic means, and is reversable. Media satisfying some of these requirements have been evolved, and within two years appropriate film media should be available having most of the desired properties, except wide dynamic range.

The potential of even a modest dynamic range medium, which can be written on, read, and erased optically, is almost without limit for implementing such statistical processing and attribute extraction techniques as factor analysis, and many other methods based on multivariate analysis. In those problems in which the nature of the raw measurement data is not clear-cut, and which call for adaptive filtering techniques, it is highly desirable to keep the data in analog form as long as possible. Premature digitalization, and the use of sequential digital processors, almost invariably leads to the adoption of highly ad hoc information extraction techniques.

A particular goal of the third phase of the DX-1 system is to explore a number of intrinsic attribute extraction schemes based on the work of P. Greene [4], W. Huggins [5], and others. These methods almost invariably involve the manipulation of large stochastic matrices and determination of the eigenvectors of some of these matrices.

ACKNOWLEDGEMENT

The author wishes to acknowledge the fine cooperation received from personnel of the Digital Equipment Corporation and to others who have been instrumental in implementing the first phase of the DX-1 system.

REFERENCES

1. Programmed Data Processor-1 Manual, F-15B, Digital Equipment Corporation, Maynard, Mass.

2. Quarterly Progress Report of Division 5 on Information Processing, Lincoln Laboratories, MIT, 15 September 1961, Report No. AFESD-TN-1019.

3. Connelly, Mark, "Real Time Analog-Digital Computation," to be published in the IRE Transactions on Electronic Computers, March 1962. See also Binsack, J.H., "A Pulsed Analog and Digital Computer for Function Generation," Electronic Systems Laboratory, MIT, Report No. AFCRL-TN-60-1111, October 1960.

4. Greene, P.H., "An Approach to Computers that Perceive, Learn, and Reason," 1959 Proceedings of the Western Joint Computer Conference.

5. Huggins, W.H., et al, "Representation and Analysis of Signals," Parts I to VIII. This is a series of monographs published by The Johns Hopkins University, Department of Electrical Engineering under Contracts AF19(604)-1941 and Nonr-248(53).

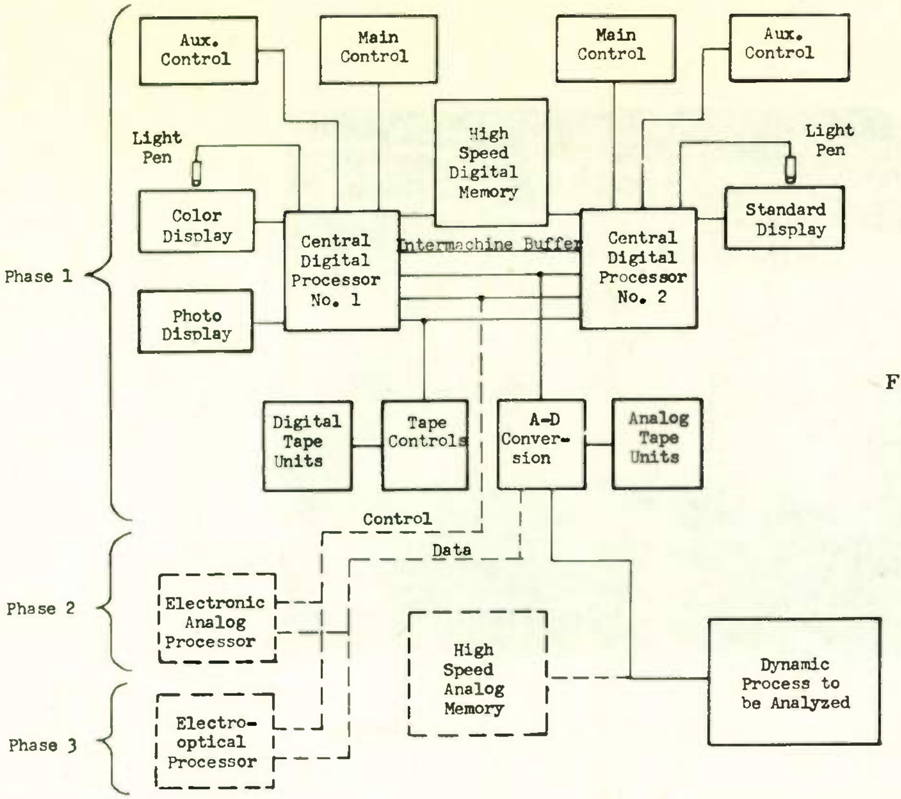


Fig. 1. DX-1 system block diagram.

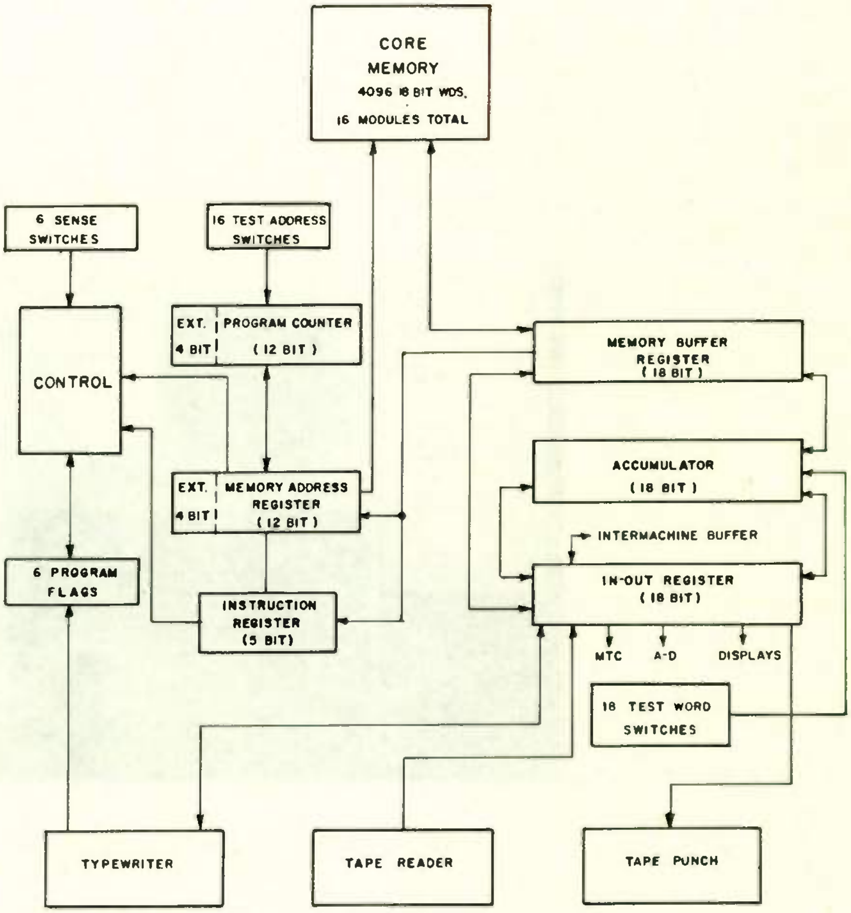


Fig. 2. Central processor block diagram (PDP-1 system).

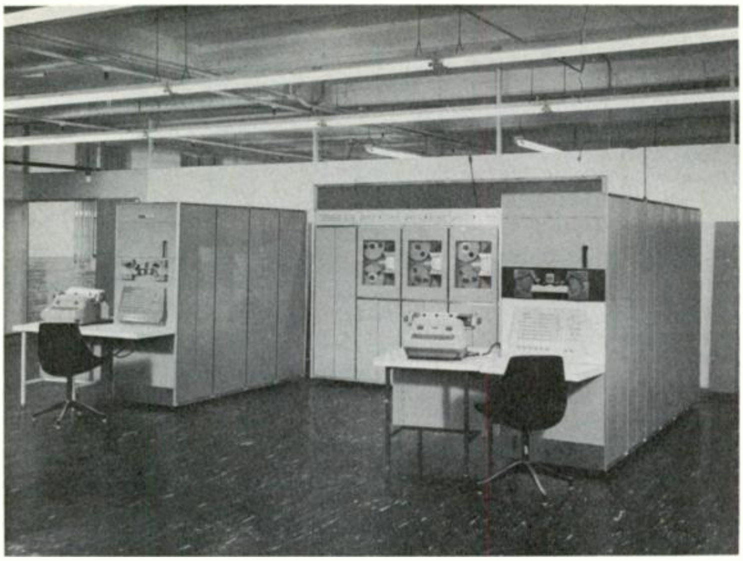


Fig. 3. Dual central processor and magnetic tape transport configuration.

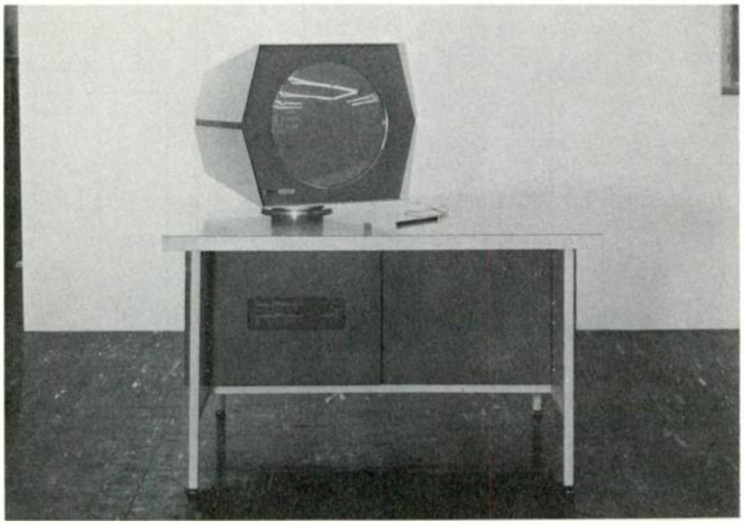


Fig. 4. Display oscilloscope and light pen unit.

REDUCTION OF TAPE SKEW
IN MAGNETIC INSTRUMENTATION RECORDERS

Finn Jorgensen and Irving Moskovitz
Mincom Division, Minnesota Mining & Manufacturing Company
Los Angeles 25, California

Magnetic tape instrumentation has undergone a tremendous development over the past decade in the areas of magnetic, electronic and mechanical performance. With such elaborate developments, it is difficult to make any prophecies about the future of the art, but in one area a limit is being reached -- the pure mechanical tape transport. The skill of building fine mechanical instruments and machinery is as old as civilization itself and the refined techniques and tooling have been fully applied by the manufacturers of tape recorders. Along with refined mechanical tape transports, tremendous improvements have also taken place in the area of magnetic heads and tapes and we have seen the packing density increased every year. This high packing density, which often runs as high as 25,000 bits per inch in a linear recording also places more stringent requirements on the mechanical transport. While the higher packing density offers tape economy, it is most often used to obtain finer details of the data, including time correlation between different events recorded on a multi-track tape. Two limitations, however, are present when we want to playback multi-track tapes and obtain a true time relationship between the various tracks.

First, there is the gap scatter that, within the present tolerances, presents a static time difference between the tracks. At 120 ips recordings, this time difference can be as much as 1 microsecond. In addition, non-uniform tension gradients are present in the tape as it passes through the tape transport, causing skew. This is illustrated in Figure 1.

Such gradients may originate from incomplete tape guidance, dimensional non-uniformity in the tape, or simply random flutter. Such skew will introduce errors and as shown in the lower portion of the illustration, it is evident that the exact time relation between the recordings on different tracks is destroyed. The nature of this skew is illustrated in Figure 2, showing the waveforms that originate from a phase detector coupled between two tracks, both recorded from the same frequency source. Several frequency components are present in this signal, and a closer analysis reveals the typical spectrum, shown in Figure 3. A rather large dc component shows up as a result of gap scatter. If plotted for a whole reel of tape, this dc component will slowly change value due to reel rate and an unavoidable, microscopic, change in the mechanical characteristics of the magnetic tape. Other rapidly varying skew components are distributed from a fraction of a cycle to several hundred cycles. Typical figures for skew on an instrumentation recorder operating at 120 ips are, between adjacent IRIG tracks, same head stack:

Static skew due to gap scatter: Less than .8 microseconds
DC component of skew: .1 to 1 microseconds
AC component of skew: ± .3 microseconds

It would be desirable to eliminate skew completely by proper mechanical design, but, as mentioned earlier, the state of the art of mechanical design is reaching its limit. Also, the tape itself is an elastic material, constantly subject to dimensional change. As a first order approximation, skew is a straight linear shear of the tape from edge to edge and it has been proposed to correct skew by mechanically moving the playback head in phase with the skew. This, however, necessitates that reference signals are recorded on adjacent tracks, filtered from the playback signal and compared in a phase detector with the output signal from the phase detector controlling servo amplifier for the playback head. This cumbersome method, investigated and tried several years ago, has never led to any practical or commercial application, to the author's knowledge.

Last year another system for reduction of skew was presented at the IRE Convention by applying electrically variable delay lines. Such delay lines are now commercially available, and can be applied in an instrument reduction of skew between the tracks. The equipment performing this feat is Mincom's TRACKLOK. Signals from two adjacent tracks, see Figure 4, are separately sent through to delay lines; the amplifiers at the ends of the delay lines are merely for isolation and to provide proper driving and terminating impedances. After the two signals have passed the delay lines they are fed to a phase detector and any phase difference, such as will arrive from skew between the two signals, will result in an error signal. This in turn is amplified and drives the delay lines. The delay lines are, as mentioned earlier, electrically variable. That is, the signal is sent through one signal winding on a Ferrite core, and around this core a modulator winding is located. By biasing the modular winding with dc, it is possible to vary the delay from 1 to 2 microseconds, and consequently by adjusting the unit about a delay for each signal of 1.5 microseconds we can correct ± 1 microsecond skew between tracks. It should be mentioned that the two delay lines are driven in push-pull -- that is, one signal is advanced, the other is delayed. The present skew reduction with a TRACKLOK exceeds a hundred times or 40 db.

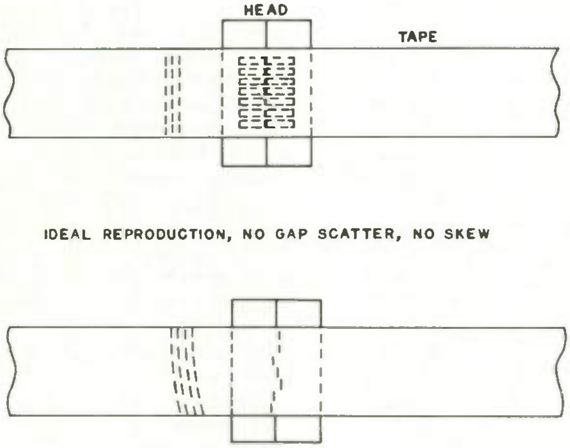
By all practical measures this has virtually eliminated skew, from an existing ± .3 microseconds to ± 3 nanoseconds. A photo of the TRACKLOK is shown in Figure 5. As you see, it is merely a black box to install in a system; no controls, just a matter of turning it on and it is operational.

Turning now to the application of TRACKLOK, the most obvious is in predetection recording. By this technique, it is understood that the data are recorded prior to detection. This is done by mixing the limited IF carrier from a receiver down so that with sidebands it falls within the bandpass of the recorder. See Figure 6. Upon playback the data can be detected either through a wideband demodulator or by mixing the carrier back up to the receiver's detector. This technique itself results in greatly improved data and is much less sensitive to tape dropouts. They do occur, however, and the only way to fully overcome them is by redundant recordings: i.e. record the same information on two adjacent tracks and add them upon playback. Without a skew compensating system this would actually result in a worse condition, where the two signals might phase each other out. See Figure 7. By using the TRACKLOK with predetection type recording where we already have the reference signals on tape, namely the modulated IF carriers, the result of adding tracks is superior to one track alone. Observations prove that dropouts are virtually eliminated by using the TRACKLOK. It should be added that the redundant technique in predetection not only eliminates dropout but increases signal-to-noise ratio with 3 db because the signals are added, giving a 6 db increase, while the random noise being added only gives a 3 db increase, resulting in overall 3 db improvement.

Figure 8 illustrates the superiority of redundant predetection recording over analog recording. Both traces display a 5 kc square wave, at a slow scan speed. The upper trace is the playback from a direct recording and displays higher noise, poorer low frequency response and dropouts. The lower trace is the detected square wave from a redundant recording, using the TRACKLOK and a wideband demodulator.

Another application for the TRACKLOK is in band splitting techniques. It is apparent that some band splitting techniques will not provide the reference signal, but this is easily overcome by merely multiplexing in a fixed carrier. This could be a clock signal, which later could be filtered in the TRACKLOK and used for skew correction. Also FM and analog splitting of a signal is possible and by recording the FM redundant on tracks 2 and 6 and the analog information on track 4, skew between all three tracks will be reduced as shown in Figure 9. The TRACKLOK operates the two delay lines in push-pull and since skew in a first order approximation is a straight shear of the tape, one track -- say 2 -- will be advanced while track 6 is delayed, with a mean value equal to the timing of track 4.

It is apparent that the method of skew reduction just described is a valuable asset to the field of data retention and reproduction. Only a few applications have been described. As the need for more accurate data reproduction increases, the role of TRACKLOK will become increasingly important.



IDEAL REPRODUCTION, NO GAP SCATTER, NO SKEW

ACTUAL REPRODUCTION, GAP SCATTER AND SKEW PRESENT (EMPHASIZED)

Fig. 1.

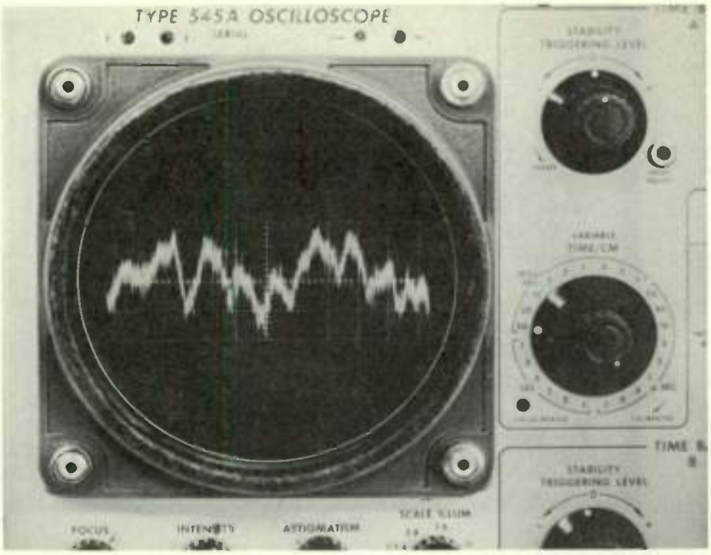


Fig. 2.

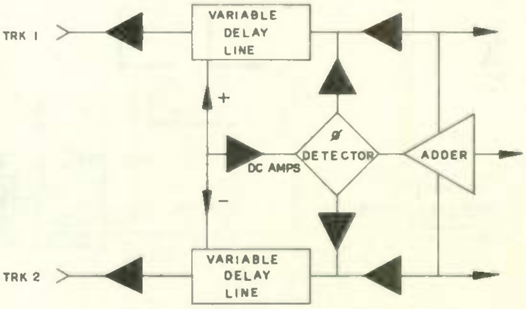


Fig. 4.

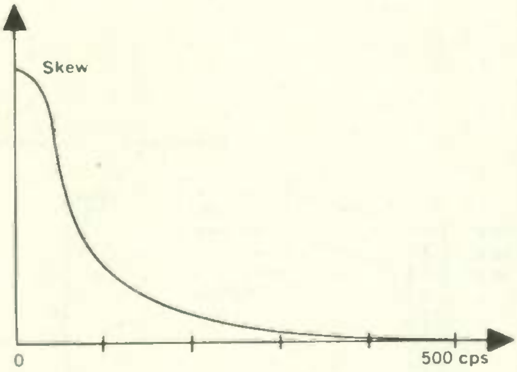


Fig. 3.

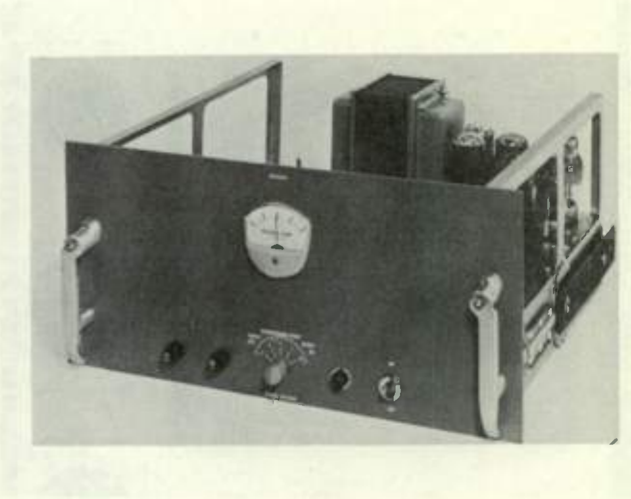


Fig. 5.

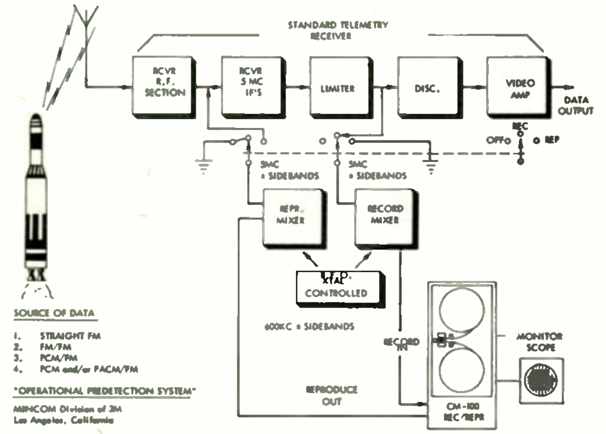


Fig. 6.

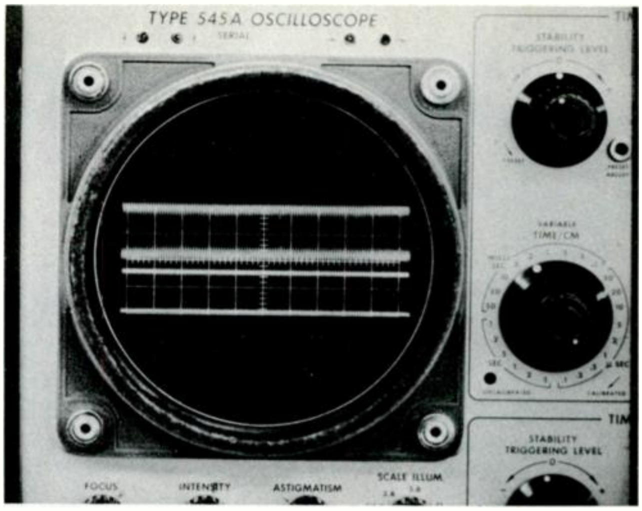


Fig. 8.

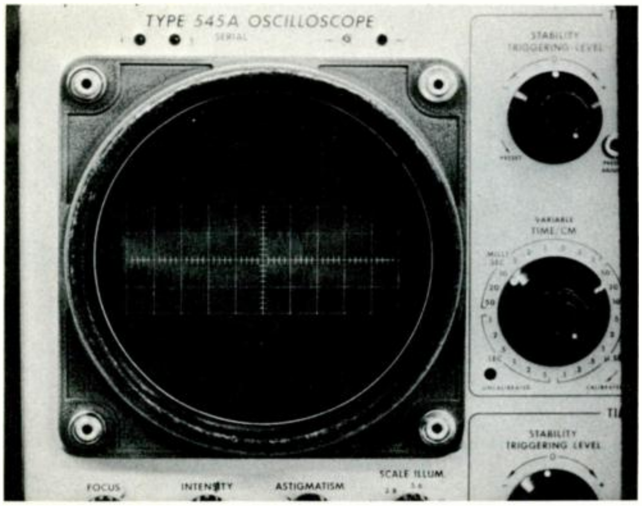


Fig. 7.

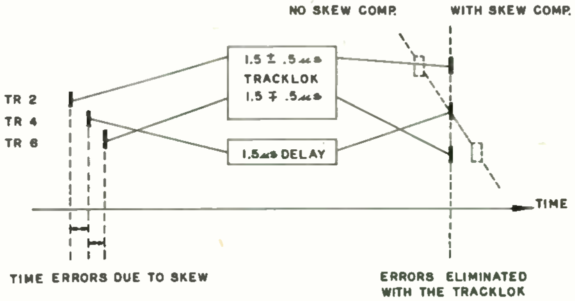


Fig. 9.

CALIBRATION OF DC AND AC DIGITAL VOLTMETERS

Charles Cimilluca* and Denis O'Leary
Electrical Measurements Laboratory
Sperry Gyroscope Company
Division of Sperry Rand Corporation
Great Neck, L.I., New York

Methods by which a digital voltmeter can be calibrated by the user are described. In the basic technique, the absolute accuracy of the digital voltmeter is measured at each full-scale range by (1) application of a known potential and (2) measurement of its linearity by comparison to the known linearity of a standard divider. The absolute voltage is determined by reference to a standard cell certified by NBS. Ranging is provided by suitable dividing networks whose accuracies are a function of one-to-one resistance matching and are thus self-calibrating. A vacuum thermocouple, also certified by NBS, is used to provide a known correlation between the a-c and d-c voltage levels. The effect of harmonics on the accuracy of average-reading AC-to-DC converters is discussed. A complete DC and AC calibration system is presented and reviewed.

Introduction

The digital voltmeter has won wide acceptance because of its high accuracy, speed, and ease of operation. The availability of laboratory accuracy levels in a general usage instrument has created a calibration problem. An indication of the magnitude of the problem is that the rated full-scale accuracy of many d-c digital voltmeters is two to four times as high as that of the typical laboratory volt box-potentiometer voltage-measurement system. The need for calibration has been accelerated by the insistence of the armed forces that all measuring instruments used in government product tests or evaluations be calibrated periodically and that these calibrations be traceable to the National Bureau of Standards (NBS).

What options does the owner of a digital voltmeter have to meet this latter requirement? A direct solution would be to submit the instrument to NBS for calibration. NBS, however, will not certify d-c digitals at this time. They feel that in the present state of development, d-c digital voltmeters should be calibrated at least every three months. Because of this, and because the accuracy of such instruments can often be affected by adjustments of uncalibrated controls, NBS recommends that d-c digital voltmeters be calibrated periodically by the user by utilization of a potentiometer and voltage divider (volt box) technique. The situation at NBS regarding the calibration of A.C.-to-D.C. converters for digital voltmeters is the same as for d-c instruments.

NBS is concerned about the problem facing the users of digital voltmeters and has taken an active part in sponsoring an American Standards Association Subcommittee to prepare an American standard for d-c digital voltmeters. (ASA Subcommittee No. C-39.6, "Digital Instruments," C. Stansbury, Chairman, National Bureau of Standards, Washington 25, D.C.). This subcommittee expects to issue a report sometime in the second quarter of 1962.

The user might purchase a complete or partial calibration system. If we specify a calibration system accuracy margin of three to one, a d-c system accuracy of ±0.0033 per cent and an a-c system accuracy of ±0.033 per cent are required. No complete system having these accuracies is known to be commercially available in the range up to 1000 volts.

It is probable that more accurate calibration systems will be available soon. The increased use of saturated standard cells and vacuum-thermocouple transfer standards coupled with the improved techniques and materials for fabricating stable resistors are signs of this. The advances made in the peak comparator method for making a-c voltage measurements may provide a new standard in this field. The purchase price of a high-accuracy calibration system, in all probability, will be very high, thus making it economically impractical for all but the larger users of digital voltmeters.

A third option is for the user to set up his own calibration system; this paper describes such a system. The general approach develops traceability to NBS via a reference-voltage source

* Mr. Cimilluca is now with North Hills Electronics Company

forD.C. and an A.C.-to-D.C. transfer standard for A.C. All operating levels are obtained by comparison to the reference voltage using dividing networks whose accuracies are a function of one-to-one resistance matching, and thus are essentially self-calibrating.

D-C Digital Voltmeter

Basic Circuit

Before discussing the calibration techniques it will be useful to review briefly the basic circuitry and operation of a d-c digital voltmeter. Most digital voltmeters utilize the potentiometer-volt box technique. A second type utilizes a voltage-to-frequency converter. The digital voltmeters referred to in this report will be those which use the potentiometer-volt box principle, although the calibration techniques are applicable to both types.

A schematic of the basic circuit is shown in Fig. 1. The potentiometer (R_1) is generally of the Kelvin-Varley or Wolff type, having a resolution and linearity of one digit in the least significant readout position. A fixed operating level, usually 10 volts, is supplied to the potentiometer by means of a Zener-diode-regulated supply (V_1). The output of the potentiometer is adjusted automatically to match the unknown voltage (V_x) to be measured. The potentiometer is standardized by means of a voltage reference (V_R), such as a standard cell. Unknown voltages greater than the operating level of the potentiometer are attenuated by a range divider (R_2,R_3,R_4). Range switching and polarity reversal are performed automatically or manually. The setting of the potentiometer at the balance condition is indicated by a digital readout.

The majority of digital voltmeters contain a standard cell (nominal value 0.0192) for self-calibration purposes. Most four-place instruments use 10 volts as the lowest full-scale range. The resolution of these instruments limits the calibration accuracy to ±0.1 per cent despite a rated accuracy of ±0.01 per cent for the standard cell.

The items that effect the accuracy of the voltmeter are:

1. The accuracy of the voltage applied to divider R_1.

2. The stability with time of the reference supply V_1.

3. The linearity of the divider R_1.

4. The ratio accuracy and stability with time and power level of the range divider comprised of R_2, R_3, and R_4,

5. The d-c offset, and offset stability with time of the d-c amplifier.

6. The last-digit-transfer sensitivity; the internal circuitry of the instrument should resolve ± half a decade beyond the last digit displayed.

Accuracy Ratings

The accuracy requirements of the d-c cailibration system are dependent on the rated accuracy of the instrument under test and the accuracy margin applied to the system. The most common accuracy statements applied to d-c digital voltmeters are:

±0.01% of full scale

±(0.01% of full scale + 1 digit)

±(0.01% of reading + 1 digit)

At full scale, the best digital voltmeter accuracy is ±0.01% or ±100 PPM. If we assign a nominal accuracy margin of 3:1, this will require a calibration- system full-scale accuracy of ±33 PPM. The linearity of a typical d-c digital voltmeter is ±100 PPM for a four-place unit and ±10 PPM.for a five-place unit, so the linearity requirements of the calibration system are ±33 PPM and ±3 PPM, respectively. Another requirement of the calibration system is that traceability to NBS be clearly established where required.

D-C Calibration Circuit

The basic calibration circuit, shown in Fig. 2, consists of three sections: a reference voltage, a ratio network to establish operating levels by comparison to the reference, and a dividing network to check linearity. Traceability to NBS is established by means of the reference device. All other voltages are determined by networks whose accuracies are a function of 1:1 resistance matching and, hence, are essentially self-calibrating. Each section will be considered.

Reference Voltage

The saturated standard cell is the only reference-voltage device certified by NBS to the required accuracy.[1] Accuracies of 1 to 5 PPM are obtainable with nominal stability of 1 or 2 PPM per year. A minimum of three cells should be available to permit intercomparison measurements. The cells require close temperature control because of their -50 PPM/°C temperature coefficient. An air oven operating at a nominal level of 30°C is frequently used to supply the regulation. Generally, saturated cells are not used as working standards because they are subject to permanent damage by a current drain; they are mostly used to calibrate working references that have good short-term stability. The reference-type Zener diode is recommended as the working standard. These devices meet rugged military environmental specifications, and, are capable of providing voltage stabilities of better than ±5 PPM over a period of several months under constant-current conditions.[2] At present there is a lack of information about their voltage hysteresis when the current is turned off for varying lengths of time, or when the diodes are subjected to temperature cycling, but this will not be a deterrent if they are calibrated before use. Nominal voltages for reference-quality Zener diodes range between 5 and 12 volts. An advantage of the Zener diode is that current may be drawn from the circuit incorporating the diode without causing permanent damage to the device. They can even be short-circuited for extended periods with no harmful effects. The temperature coefficient of reference Zener diodes is generally between ±5 to ±50 PPM/°C, and they should be temperature controlled for maximum stability.

For convenience, a 10-volt Zener-diode reference is preferred. The circuit of a reference "package" is shown in Fig. 3. It consists of a reference diode operating slightly above 10 volts. A ratio network comprised of R_x and R_y compare the voltage across the Zener with that of the saturated standard cell, V_R. The ratio is chosen so that:

$$\frac{R_x}{R_x + R_y} = \frac{V_R}{10 \text{ volts}} \qquad (1)$$

The 10-volt level is set by adjusting rheostat R_z for a null at the detector. A pre-regulator is used to stabilize the current through the reference diode, and all critical components are enclosed in a regulated oven. An added convenience to this circuit would be a divider or rheostat located at the junction of R_x and R_y which would permit the normal range of standard-cell values to be selected by a simple dial setting. The Zener package is used to set the 10-volt level directly and the 100-and 1000-volt levels by means of suitable dividing networks.

Ten unsaturated standard cells can be individually compared to a saturated standard cell and then connected in series to provide a nominal 10-volt reference. An exact 10-volt level can be established by comparison to this group. The possibility of damage due to current drain or mishandling makes the unsaturated cells less suitable for reference than the more rugged Zener diode.

Ratio Network

Accurately known ratios R_A/R_B are required to compare the 10-volt reference to the saturated-standard-cell reference and to compare the 100- and 1000-volt levels to the 10-volt reference. The first ratio is 10 volts: V_R or approximately 10:1.0181. The second and third ratios required are 10:1 and 100:1 respectively above the 10-volt level. Two approaches for the dividers are suggested. The first is a Kelvin-Varley divider having a linearity of ±1 PPM. The second uses groups of matched resistors in a series-to-parallel arrangement.

A one-part-per-million Kelvin-Varley voltage divider provides the most convenient form of a standard ratio. To set the 10-volt level, the value of the reference voltage divided by 10 is set on the divider controls and the voltage across the divider adjusted for a null. No calculations or auxiliary measurements are required. The accuracy of the voltage at the slider of the divider is limited to ±10 PPM by the linearity of the divider. The ratio accuracy of the divider at a nominal 10:1 setting is ±10 PPM. This adds up to an accuracy of ±20 PPM in setting the 10-volt, full-scale level. The 1-PPM divider can be used to calibrate an auxiliary divider having good short-term stability and adequate resolution. In this way the 1 PPM divider also can be used to check linearity while the 10-volt level is continuously monitored during the calibration.

In a similar manner the 1-PPM divider can be used to set the 100-volt level by comparison to the 10-volt reference. The accuracy of the 100-volt setting would be ±30 PPM. Unfortunately the Kelvin-Varley divider cannot be operated at the 1000-volt level because of power limitations.

The 1-PPM divider and a saturated standard cell provide the user with a simple means of calibrating a d-c digital voltmeter through the 100-volt range with a 3:1 accuracy margin. The 1-PPM divider can be checked with a group of 10 resistors matched to ±1 PPM of the mean. Trimmers can be included in each resistor to provide a means for matching. The 10 resistors are then connected in series to provide a decimal divider having a linearity of ±0.1 PPM. This divider is used to check the linearity of the first dial of the ±1 PPM divider. One of the 10 resistors is then removed and the 9 remaining resistors form a divider with settings of .111111, .222222, etc. that is used to provide a check on all decade positions of the ±1 PPM divider. At present, at least three vendors are marketing a ±1 PPM Kelvin-Varley divider.

Another method for obtaining accurate R_A/R_B ratios involves a series-to-parallel technique. It has been shown that a group of N resistors matched to ±0.1% of the mean has a series resistance that is N^2 times its parallel resistance to an accuracy of 1 PPM.[3] This technique can be extended to provide an accurate input-to-output ratio of $N^2 + 1:1$ by the following method. A group of N resistors is connected in series and another group of N resistors in parallel to form a divider network. The matching conditions are that the resistors within a group are matched to ±0.1% of the mean of the group, and that the means of the two groups are matched to ±0.1%. The positions of the two groups are then interchanged with the series resistors now connected in parallel and the parallel resistors now in series. The average of the two input-to-output ratio settings will be $N^2 + 1:1$ to an accuracy of better than ±2 PPM. For N = 3 a 10:1 ratio is obtained. A 101:1 ratio (N = 10) is the closest to 100:1.

The 10:1 ratio can be used to set the 10-volt reference by comparison to the reference standard cell (V_R). This is done by placing an accurate potentiometer, set to the value of V_R-1 volt, between the output of the ratio network and the reference cell. The polarity of the potentiometer must oppose that of the reference voltage. The nominal value of V_R-1 is 18 millivolts, this can be measured to an accuracy of ±7 microvolts or ±7 PPM referred to V_R. If a conservative ratio accuracy of ±10 PPM and a standard-cell reference accuracy of ±5 PPM are assumed, then the 10-volt reference would be accurate to ±22 PPM.

The 10:1 ratio also would be used to set the 100-volt level by comparison with the 10-volt reference. The accuracy of the 100-volt level would be ±32 PPM. The 101:1 ratio would be used to set the 1000-volt level by comparison to the 10-volt reference. Two methods can be used. In the first, an accurate potentiometer set to 10-(1000/101) or 0.09901 volts is placed betwen the output of the ratio network and the 10-volt reference, with its polarity opposing that of the 10-volt reference. The voltage across the ratio is adjusted for a null at the potentiometer. The potentiometer reading of 0.09901 volts can be measured to an accuracy of ±27 microvolts, or ±2.7 PPM, referred to the 10-volt reference. The accuracy of the 1000-volt level is then ±35 PPM.

A second method is to establish a reference voltage of 10/1.01 or 9.90099 volts in conjunction with the 10-volt reference. The 1000-volt level is then set by obtaining a null between the junction of the 101:1 ratio and the 9.90099-volt reference. The latter reference is set by means of a divider placed across the 10-volt reference. An accurate potentiometer set to 0.09901 volts (10-9.90099) is placed between the slider of the divider and the high side of the 10-volt reference. The slider is then adjusted for a null at the potentiometer. The accuracy is the same, but this method is more convenient in that the 9.90099-volt reference is set at the same time as the 10-volt reference, and all three ranges, 10, 100, and 1000 volts, are set by direct comparison to their appropriate reference.

Dividing Network to Check Linearity

A commercial Kelvin-Varley divider for R_c having a rated linearity of ±10 PPM or better will be satisfactory for checking the linearity of four-digit instruments. The requirement for checking a five-digit instrument is a standard divider linearity of ±3.3 PPM. The commercial voltage dividers are usually limited in range to about 300 volts, so that they cannot be used directly on the 1000-volt range. To provide a divider compatible to all three ranges, a matched-resistance string will be considered. Linearity is often checked by taking readings at the nominal settings of 0.11111, 0.22222, 0.33333, etc. This provides a rapid linearity check of the instrument under test. It is understood that such a technique could give incorrect results because of the possibility of an unusual combination of errors. This probability can be reduced by checking also the linearity of the first decade of the digital voltmeter. The first decade in a precision divider is the most critical, with the second decade having only 1/10 the weight, the third decade 1/100, and so on. With this in mind, a divider consisting of 10 resistors matched to within 30 PPM of the mean could be used to check the linearity of the first decade of the digital voltmeter. This 10-point divider would have a linearity rating of ±3 PPM. Then one resistor would be removed and the remaining resistors would form a divider with settings of 0.88889, 0.7778, 0.66667, 0.55556, 0.44444, 0.33333, 0.22222, and 0.11111. This 9-point divider would have a linearity of ±3.3 PPM. The resistors forming this 9- and 10-point divider must have excellent short-term stability relative to the mean. The resistors could be assembled with individual trimmers to provide a means for matching.

The full-scale levels developed for the calibration system are based on 1.00000, but the digital voltmeter full-scale is based on 0.99990. To provide a means of checking the 0.9999 setting on the digital when using the matched-resistor string, a resistor equal to 0.01% (0.001% for checking the 0.99999 setting on five-place instruments) of the total resistance of the string is inserted between the high resistor in the string and the high side of the voltage level used. The voltage at the junction of the inserted resistor and the top resistor in the string will be the desired 0.99990. The 0.01% value is not exact, but it is close enough to produce negligible error, and it will generally permit the use of a common-valued resistor. The accuracy requirements for this resistor are not stringent; 1% would be adequate. After the full-scale setting is checked, the additional resistor is switched out and the linearity check is made as described.

The schematic of the calibration circuit (Fig. 2) shows the digital voltmeter under test connected directly across the output of the standard divider. In practice, the loading of the digital would alter the linearity of the standard divider. For example, consider an ideal divider set at midscale and having an input resistance of 100,000 ohms. If a load of 10 megohms, which is the typical input resistance of a d-c digital voltmeter, is placed between the slider and the low side, it will introduce an error of approximately 0.25% of reading or 0.125% linearity. This problem is eliminated by the use of a double-divider technique, as shown in Fig. 4. An auxiliary divider, R_C', is connected in parallel with the standard divider, R_C, and the d-c digital voltmeter is connected to the output of R_C'. The slider of the standard divider is set to the desired check point and the auxiliary divider is adjusted for a null at the detector. The difference between the known standard divider setting and the digital voltmeter reading, taken as a per cent of the full-scale range, is the linearity error of the instrument at that point.

Dividers can be designed to provide compensation for a specific load resistance, but the procedure is cumbersome and expensive for the high accuracies required.

Power Supply

The power supply must supply up to 1000 volts at a current equal to that required by the d-c calibration circuit plus an additional 5 to 10 milliamperes for the a-c calibration circuit. The short-term stability (1 minute or longer) should be ±5 PPM or better to permit the operating levels to be set precisely. The resolution of the output-voltage control (internal and/or external) should be 1 PPM.

Detector

The detector should have a sensitivity of 1 microvolt and be capable of operating with a source impedance as large as 100,000 ohms. Electronic galvanometers are recommended because their ruggedness and ease of operation.

A.C.-to-D.C. Converters

To enable the d-c digital voltmeter to give readings of a-c voltage, some accurate means of converting A.C. to D.C. is necessary. Most manufacturers supply such converters as accessories to their basic d-c digital voltmeters. The majority of converters are average-responding instruments, that is, their output is proportional to the average value of the a-c voltage under measurement. Recently, true RMS converters with sufficient accuracy for digital-voltmeter use have been perfected. This type of converter is finding increasing use in the electronic industry. A calibration-system-accuracy margin of 3:1 is desired for both types of converters.

Average-Reading Converters

A block diagram of a typical average-responding A.C.-D.C. converter is shown in Fig. 5. The RC-compensated attenuator may be either manually operated or electrically tied to the d-c digital voltmeter logic circuit to provide automatic ranging. Since the majority of A.C.-D.C. converters operate with 10 volts as the full scale output, the attenuator is chosen to give 10:1 and 100:1 ratios so that full-scale voltages of 100 and 1000 volts are also measurable. The isolation amplifier is a high-input-impedance unit, generally having a gain of 1. It isolates the a-c rectifier from the attenuator, allowing the attenuator to work into a high input impedance to reduce loading errors. The most common form of A.C.-D.C. converter is the operational rectifier. A typical operational rectifier is shown in Fig. 6. The diode rectifiers D_1 and D_2, generally silicon or germanium, are placed in the feedback loop to improve the linearity and stability of the rectification. The output is a half-wave rectified a-c voltage. This waveform is integrated by a filter or operational integrator that may also serve as the output amplifier. The time constant of the integrator is chosen to be much greater than the period of the lowest frequency voltage to be measured. The d-c output voltage, E_{av}, is therefore propertional to the average value of the half-wave rectified input voltage. The rms-to-D.C. transfer function of the rectifier for a sine wave is:

$$E_{av} = \frac{1}{2\pi}\int_{0}^{\pi} A \sin \omega t \quad d\omega t \tag{2}$$

where A = peak amplitude of a-c input voltage

ω = angular frequency of the applied voltage

E_{av} = average value of half-wave rectified sine wave

Integrating we get:

$$E_{av} = \frac{1}{2\pi}\left[-A \cos \omega t \right]_{0}^{\pi} = \frac{A}{\pi} \tag{3}$$

The rms value of a pure sin wave is:

$$E_{rms} = \frac{A}{\sqrt{2}} \qquad (4)$$

Therefore, the ratio of average D.C. from the rectifier, to the true rms of the voltage being measured is:

$$\frac{E_{av}}{E_{rms}} = \frac{\sqrt{2}}{\pi} \qquad (5)$$

In order for the digital-voltmeter converter to produce an output d-c voltage equal to the rms value of the a-c sine wave, a gain of $\pi/\sqrt{2}$ must be provided. This gain is usually obtained by adjustment of the feedback ratios in the operational rectifiers.

Most manufacturers of average responding converters rate these instruments as ±0.1% of full scale. This does not include the d-c digital voltmeter error, nor any error caused by harmonics in the waveform under measurement.

Effect of Harmonics On Accuracy of Average-Reading Converters. Since many applications of a-c digital voltmeters require the measurement of sources with high or unknown harmonic content, it is important to analyze the effects of distortion on the converter accuracy. If we represent the voltage under measurement and its second and third harmonics as:

$$V_{ac} = A \sin \omega t + B \sin (2\omega t + \phi 2)$$

$$+ C \sin (3\omega t + \phi 3) \qquad (6)$$

where V_{ac} = instantaneous amplitude of the voltage under measurement

A = peak amplitude of fundamental

B = peak amplitude of second harmonic

C = peak amplitude of third harmonic

ϕ_2 = phase angle of second harmonic

ϕ_3 = phase angle of third harmonic

ω = angular frequency

Then:

$$E_{av} = \frac{A}{2\pi} \int_0^{\pi} \sin \omega t \ d(\omega t)$$

$$+ \frac{B}{2\pi} \int_0^{\pi} \sin (2\omega t + \phi_2) d (\omega t) \qquad (7)$$

$$+ \frac{C}{2\pi} \int_0^{\pi} \sin (3\omega t + \phi_3) d (\omega t)$$

$$= \frac{A}{\pi} + \frac{C}{3\pi} \cos \phi_3 \qquad (8)$$

In a like manner it can be shown that all even-harmonic terms have no effect on the average value of the half-wave rectified sine wave. It may also be shown that for odd harmonics:

$$E_{av} = \frac{1}{\pi} (A + \frac{C}{3} \cos \phi_3 + \ldots\ldots \frac{N}{n} \cos \phi_n) \qquad (9)$$

where n = n^{th} odd harmonic

N = peak amplitude of n^{th} odd harmonic

If we assumed that the constant of proportionality of the average-responding converter to which the distorted wave has been applied is set to $\pi/\sqrt{2}$, the per cent error between the converter output and the true rms value of the wave is:

$$\% \text{ error} = \frac{\sqrt{1/2 (A^2 + B^2 + C2\ldots. + Q^2)} - \left[\frac{\pi}{\sqrt{2}}\right]\left[\frac{1}{\pi}\right](A + \frac{C}{3} \cos \phi_3 \ldots + \frac{N}{n} \cos \phi_n)}{\sqrt{1/2 (A^2 + B^2 + C^2 \ldots + Q^2)}} \times 100$$

$$= \left[1 - \frac{A + \frac{C}{3} \cos \phi_3 \ldots + \frac{N}{n} \cos \phi_n}{\sqrt{A^2 + B^2 + C^2 \ldots + Q^2}} \right] \times 100 \qquad (10)$$

where Q = last harmonic, odd or even.

The equation shows that, for a small amount of in-phase third-harmonic distortion, say 0.3%, an error of almost ±0.1% is made by the digital-voltmeter converter in addition it its rated ±0.1%. If the same percentage of third harmonic were shifted in phase by 90 degrees with respect to the fundamental, the error contributed would be greatly reduced. The calibration agency and the user of the voltmeter should be certain that the voltage source is free from harmonics to avoid additional errors when using this type of converter.

The effect of odd-harmonic distortion as a function of magnitude and phase angle is shown graphically in Fig. 7. The plot is the result of actual measurements made on a typical average-reading converter; note how it follows the characteristic cosine function.

True RMS Converters

The second type of A.C.-D.C. converter presently in use is the true rms device. Such converters generally employ a temperature-sensitive element in a bolometer bridge configuration. (The term bolometer is used to identify the general class of circuit elements that undergo a change in resistance caused by the current being measured). A typical circuit of a true rms converter is shown in Fig. 8. The resistors R_1 and R_2 form the ratio arms of a Wheatstone bridge. The remaining arms are made up of temperature-sensistors R_3 and R_4. The bridge is adjusted so that $R_1/R_2 = R_3/R_4$, and the input to high-gain amplifier A is zero with no a-c input. The zero-adjust voltage is made equal in magnitude and opposite in polarity to the d-c drop across R_4. When an a-c signal is applied, the bridge is unbalanced and an input is obtained at amplifier A. The amplifier controls the variable d-c supply and adjusts its voltage to return its bridge to balance. If both thermosensitive resistors have the same temperature resistance characteristics and have low residual reactances, the d-c voltage required to rebalance the bridge is equal to the rms a-c input.

The typical accuracy ratings of this form of converter are ±0.2% of reading above 20% of full scale. Where high harmonic distortion is present, this form of converter is more accurate than the average responding type.

A-C Calibration Circuit

An A.C-D.C. digital voltmeter is calibrated in much the **same** way as a d.c. digital voltmeter. Known a-c full-scale voltages are applied to the instrument, and the instrument reading is compared to the known voltage. The linearity of the instrument is then checked at each range. A circuit of the basic a-c calibration system is shown in Fig. 9.

D-C Standard

The basic d-c calibration system may be used to calibrate the a-c digital voltmeter with the addition of an A.C.-D.C. transfer standard, an a-c supply, and a variable-ratio transformer. The d-c supply, as described in the first section of this report, may be used to establish the 10, 100, and 1000 volt full-scale a-c voltages by means of the A.C.-D.C. transfer element. For this reason the d-c supply should be capable of supplying the 5 to 10 milliamperes required for the transfer standard in addition to sufficient current to operate all divider and resistor networks.

A.C.-To-D.C. Transfer Standards

Thermocouples.[4] Among the most convenient and accurate A.C.-D.C. transfer standards is the vacuum thermocouple. The couple may be certified by NBS to a transfer accuracy of ±0.01%. Most couples require a series resistor so that the current through them will be limited to the rated value for any applied voltage. The resistor must have low a-c reactances, thus presenting the same impedance to the a-c source as to the d-c source. NBS will certify a package composed of a resistor and vacuum thermocuple as a transfer device at 10, 100, and 10000 volts to an accuracy of ±0.01%. A schematic diagram of a thermocouple transfer standard and its associated equipment is shown in Fig. 10. A null on galvanometer G is obtained with the thermocouple circuit connected across the d-c standard by adjustment of the potentiometer (Lindeck type). The transfer switch is then thrown into the a-c position, and the a-c supply is adjusted for the same null. The reversing switch is then thrown to the reverse position and the procedure repeated. The a-c level will be the average of the readings in the normal and the reverse positions. Typical 5 to 10 milliampere thermocouples produce a nominal output voltage of between 5 and 15 millivolts. The galvanometer must have sufficient sensitivity to resolve at least 1 microvolt if the full transfer accuracy of the couple is to be realized. The thermocouple is a true rms transfer device.

Peak Comparators.[5] A recent development in a-c measurements has been an improved method for measuring peak voltage. The voltage to be measured is compared to a known d-c reference using a chopper technique. The use of clipping levels allows the peak region of the a-c voltage to be effectively expanded to a point where comparison accuracies of 0.005% are claimed using an oscilloscope as the comparator. This type of transfer device is to a certain extent self-calibrating in that two known d-c voltages may be fed to the comparator and the comparator error determined. When the peak comparator is used for rms measurements it is important that the a-c signal be free from distortion, or else, serious errors may occur.

Dynamometers.[6] Dynamometers are capable of accuracies on the order of ±0.005% as A.C.-D.C. transfer devices. The dynamometer provides a true rms transfer. These instruments are generally exceptionally fragile and therefore not portable; they are also extremely delicate; and they require the use of skilled personnel and the application of correction factors if their full accuracy is to be realized. For these reasons they are not recommeded for general use.

Bolometers.[7] Bridge circuits employing thermal-sensitive elements provide a true rms transfer. These units are less susceptible to burn-out than the thermocouple, but at this time their accuracy is limited to about ±0.05%.

Ratio Transformer

A wide selection of ratio transformers are available commercially with linearities of ±10 PPM or better at voltages up to 350 volts. At this time, at least three vendors produce units that will operate to 1000 volts with the same linearity rating. The low output impedance of ratio transformers, usually less than 10 ohms, eliminates the need for a double-divider technique to reduce the loading effects of the digital voltmeter. For example, the loading of a one-megohm a-c digital voltmeter will cause an error of less than ±10 PPM. The frequency response over which the ratio transformer must maintain its accuracy is dependent on the user's needs and the frequency response of the digital voltmeter being calibrated. If the converter is to be calibrated over its entire frequency and voltage range, two ratio transformers may be required to cover the extreme limits of voltage and frequency.

A-C Voltage Supply

The a-c supply used in the calibration of digital voltmeters must have a voltage range at least equal to the voltage range of the instrument under calibration. In most cases this would be 1000 volts rms. The supply must be capable of supplying the rated thermocouple current plus that of the associated circuitry. The stability of the supply must be on the order of ±0.01% short term. Even though the best a-c digital voltmeter rating is ±0.1%, the resolution at full scale remains 0.01% thus any instability greater than the resolution will cause the last digit of the voltmeter to roll, making readings difficult if not altogether impossible.

As noted previously, harmonics present in the source can cause severe error in the reading of the average responding converter. For this reason the harmonic content of the a-c supply should be lower than 0.03% to produce negligible calibration error. The resolution of the supply should be better than ±0.01% so that the digital voltmeter and termal-transfer element may be set to full accuracy.

Complete Calibration System

A complete d-c and a-c calibration system is shown in Fig. 11. It is essentially a combination of the sections discussed previously. The traceability to NBS is established by (1) the saturated standard cell for the d-c portion and (2) the vacuum thermocouple plus its associated series range resistors for the a-c portion. All other critical components in the system are established on a relative basis by matching techniques. A typical operating procedure follows:

D-C Section

1. Supply V_2 is set to provide proper current level for the oven-controlled, reference-type Zener diode, ZD, whose output is a nominal 11 volts.

2. Potentiometer P_1 is set to the value of the reference standard cell (V_R) minus 1 volt. P_1 is then connected between the junction of the 10:1-ratio units, R_1, R_2, with its polarity opposing that of V_R.

3. Trimmer resistor R_s is adjusted for a null at the potentiometer. This establishes the 10-volt reference.

4. The potentiometer P_1 is then set to 10/1.01, or 0.09901, volts and placed between the slider of R_s' and the high side of the 10-volt reference. The slider of R_s' is then adjusted for a null at the potentiometer, establishing a reference level of 9.90099 volts at the slider.

5. The range switch is set to the 10-volt position and supply V_1 adjusted for a null at detector D_1. This establishes the 10-volt operating level.

6. Switch S_2 is set to the 0.99990 position, and switch S_3 is adjusted to the 1.0 setting. This allows the full-scale digital voltmeter reading of 0.9999 to be checked by adjusting the auxiliary divider R_C' for a null at detector D_2 and noting the digital voltmeter reading. It is a general practice to adjust the full-scale settings during calibration if they are not correct. The vendor's directions for making the adjustment should be followed.

7. Switch S_2 is set to the No. 10 position, and the linearity of the first decade of the digital is

checked by making readings at positions 0.9 through 0.1. Switch S_2 is then set to the No. 9 position, and readings are made at the 0.8 through 0.1 positions for a check of all decades. The difference between the digital reading and the setting of the standard divider as a per cent of full scale is the linearity error at this point.

8. The range switch S_1 is set to the 100-volt position, and supply V_1 is adjusted for a null at detector D_1. This sets the 100-volt operating level in comparison to the 10-volt level using the 10:1-ratio R_4R_5 combination as the comparator.

9. The full-scale and linearity measurements are similar to those for the 10-volt level.

10. The range switch S_1 is set to the 1000-volt position, and supply.V_1 is adjusted for a null at detector D_1. This sets the 1000-volt operating level by comparison to the 9.90099-volt level using the 101:1 ratio R6R7 as the comparator.

11. The full-scale and linearity measurements are similar to those for the 10-volt level.

A-C Section

1. The A.C.-D.C. switch, S_4, is set to the d-c position, and the range switches S_1 and S_5 are set to the 10-volt position.

2. The 10-volt d-c level is set as described in the d-c section.

3. With the thermocouple switch in the normal (N) position, the output of the thermocouple is balanced by the Lindeck potentiometer, as shown by detector D_3.

4. Swtich S_4 is then set to the a-c position, and the a-c supply is adjusted for a null at detector D_3. This sets the 10-volt level across the standard ratio transformer.

5. The slider of the ratio transformer is set for a reading of 0.99990, corresponding to 9.9990 volts. The difference between this value and the digital reading is the full-scale error.

6. The linearity is checked in a manner similar to that used in the d-c section. The first decade is checked from 0.1 through 0.9, and then all decades are checked at 0.1111, 0.2222, etc. The low source impedance of the ratio transformer eliminates the need for an auxiliary divider.

7. The 100- and 1000-volt levels are checked in a similar manner.

8. Measurements are repeated with the thermocouple switch in the reverse (R) position. The average of the two sets of readings is used to determine the a-c accuracy of the instrument under test.

Conclusions

Calibration of high-accuracy devices such as the digital voltmeter require techniques that approach the present state of the art. Certain basic equipment should be available; the following equipment is recommended for the calibration of d-c digital voltmeters:

1. A group of at least three saturated standard cells in an air oven. The cells should be taken to NBS, in their oven, at maximum intervals of two years for calibration and certification.

2. A Zener-diode reference package that can be set to an exact 10-volt level by comparison to a saturated standard cell. This unit will then provide a 10-volt working reference.

3. A group of 10 resistors that can be matched to ±1 PPM of the mean. The resistors should be matched to within ±100 PPM, and should be provided with trimmers for the exact matching. The matching stability of this string should hold for at least 10 minutes and preferably 1 hour or more. The resistors should be temperature controlled.

4. Two groups of six and twenty resistors to form 10:1 and 101:1 respectively, using the series-to-parallel technique. The resistors should be matched to within ±0.1% (1000 PPM) of the mean. A two-position switching arrangement can be devised to provide the series-to-parallel interchange. The 6-resistor group could be a part of the 20-resistor group; this requires an additional switching circuit. These resistors should not require trimmers and could probably be used without temperature control. It would be advisable, however, to place them in the same regulator as the ±1 PPM matched resistors.

For the calibration of A.C.-D-C. converters, the following is recommended:

1. An A.C.-D.C. thermal-transfer reference unit consisting of a 5-ma vacuum thermocouple with noninductive current-limiting resistors to provide full-scale ranges of 10, 100, and 1000 volts. The components should be enclosed in a shielded container that will also provide a thermal lag. The unit should be sent to NBS for calibration and certification at one-year intervals. The frequencies desired should be specified.

The key to an accurate calibration is the development of exact ratios. The calibrator should cross check his ratios by a second method until he has acquired confidence in his technique. For example, the 10:1 ratio established by the series-to-parallel arrangement can be checked by matching 10 resistors, connecting them in series and calculating the ratio at a 10:1 setting. Reference 8 describes an interesting variation of the series-to-parallel method.

In a general paper of this type, specific details are not included. The problems of change in contact resistance, thermal voltages, temperature and load coefficient, and leakage are passed over with little or no mention. The references cited, particularly Wenner's paper, will help fill these voids. Reference 9 describes a method for determining the load coefficient of resistors, a most important consideration in developing ratios for the 1000-volt range.

In the final analysis, the part played by the person performing the calibration should be emphasized. As Wenner expressed it, "Whether or not it (the expected precision) is actually obtained depends to a large extent on the temperment and skill of the observer. To be properly qualified he should have a desire to do the job well, be neither easily fatigued nor perturbed, recognize slight disturbances promptly and be able to locate their source and correct the difficulty, and above all else, be able to differentiate between that which is essential and that which is not essential...."

References

1. F.B. Silsbee, "Establishment and Maintenance of the Electrical Unit.". National Bureau of Standards Circular 475, 1949.

2. R.P. Baker and J. Nagy, "An Investigation of Long-Term Stability of Zener Voltage References," Sandia Corporation Report No. 194, June 1960.

3. F. Wenner, "Methods, Apparatus and Procedures for the Comparison of Precision Standard Resistors," Journal of Research of the National Bureau of Standards, Vol. 25, August 1940.

4. F.L. Hermach, "Thermal Converters as A.C.-D.C. Transfer Standards for Current and Voltage Measurements at Audio Frequencies," Journal of Research of the National Bureau of Standards, Vol. 48, 1952.

5. "A.C.-D.C. to D.C. Comparator," Rotek Instrument Corp. Application Note No. 2, August 1960.

6. J.H. Park and A.B. Lewis, "Standard Electrodynamic Wattmeter and A.C.-D.C. Transfer Instrument, " Journal of Research of the National Bureau of Standards, Vol. 25, 1940.

7. A. Cooper, "The D-930-A Muirhead - Wigan Precision RMS Decade Voltmeter," Technique, Vol. 15, No. 1, January 1961. (This is a publication of Muirhead and Company.)

8. N.J. Harris, "A Precision Megohm Ratio Unit for High Voltage Measurements," The Review of Scientific Instruments, Vol. 23, No. 8, August 1952.

9. Curtis, Electrical Measurements, McGraw-Hill Book Company, Inc., New York 1937.

Note: National Bureau of Standards Handbook 77, Vol. 1, issued February 1, 1961, and entitled Precision Measurement and Calibration - Electricity and Electronics contains four of the references cited above us well as other papers which are standards in the measurement field. It is available from the Superintendent of Documents, U.S. Government Printing Office, Washington 25, D.C.

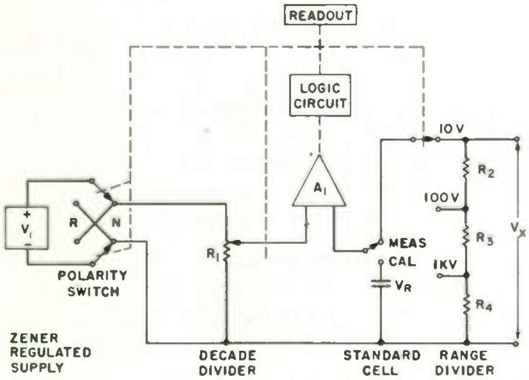


FIG. 1 BASIC CIRCUIT OF A D-C DIGITAL VOLTMETER

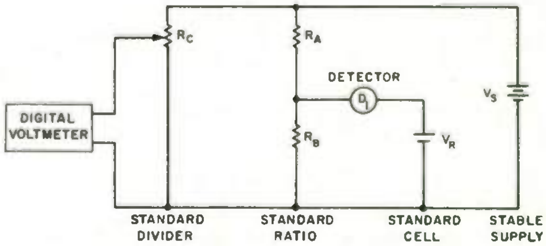


FIG. 2 D-C CALIBRATION CIRCUIT

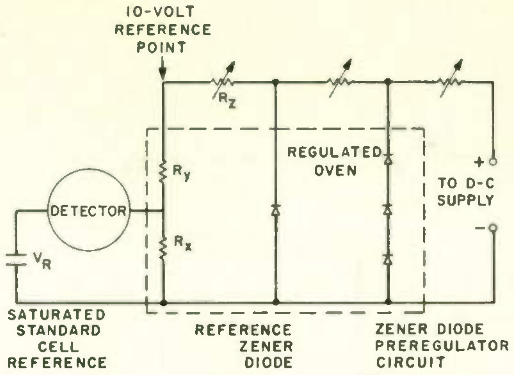


FIG. 3 10-VOLT ZENER DIODE REFERENCE PACKAGE

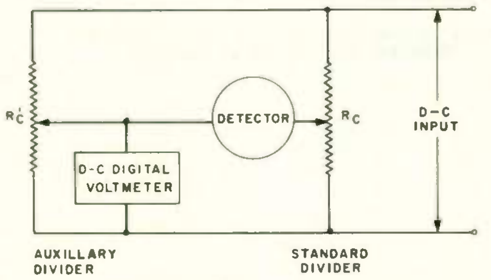


FIG. 4 DOUBLE-DIVIDER TECHNIQUE TO ELIMINATE DIGITAL-VOLTMETER LOADING ERRORS

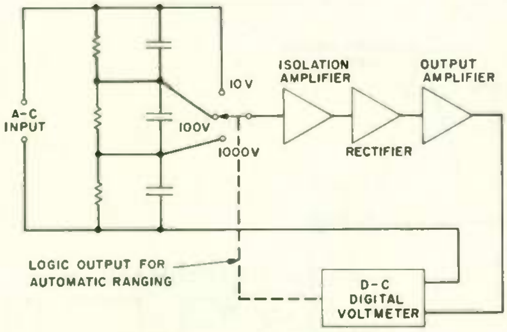


FIG. 5 A.C.-TO-D.C. CONVERTER

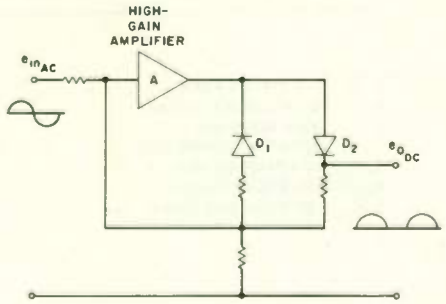


FIG. 6 OPERATIONAL RECTIFIER

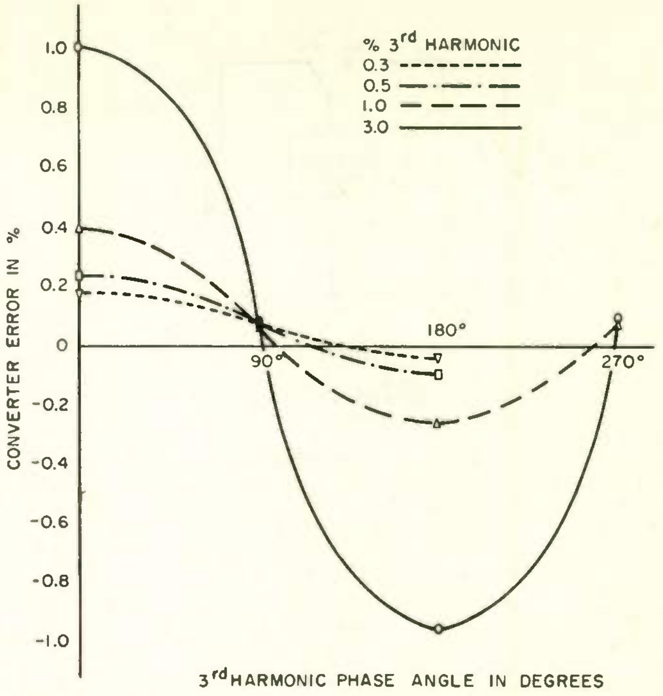


FIG. 7 EFFECT OF 3rd HARMONIC DISTORTION ON A TYPICAL AVERAGE-READING A.C.-TO-D.C. CONVERTER (ACTUAL MEASUREMENTS)

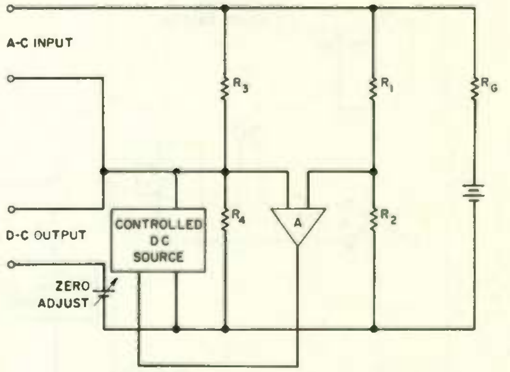


FIG 8. TRUE RMS CONVERTER

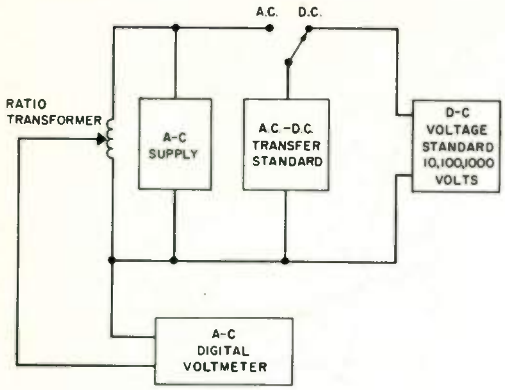

A.C.
D.C.
RATIO TRANSFORMER
A-C SUPPLY
A.C.-D.C. TRANSFER STANDARD
D-C VOLTAGE STANDARD 10,100,1000 VOLTS
A-C DIGITAL VOLTMETER

FIG. 9. A-C CALIBRATION CIRCUIT

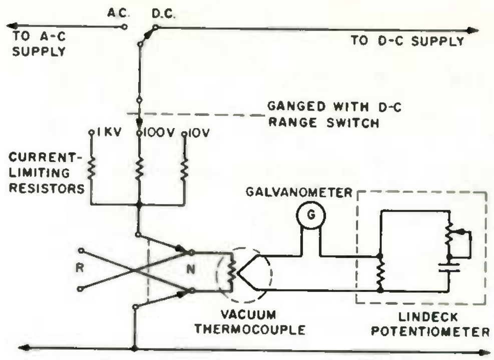
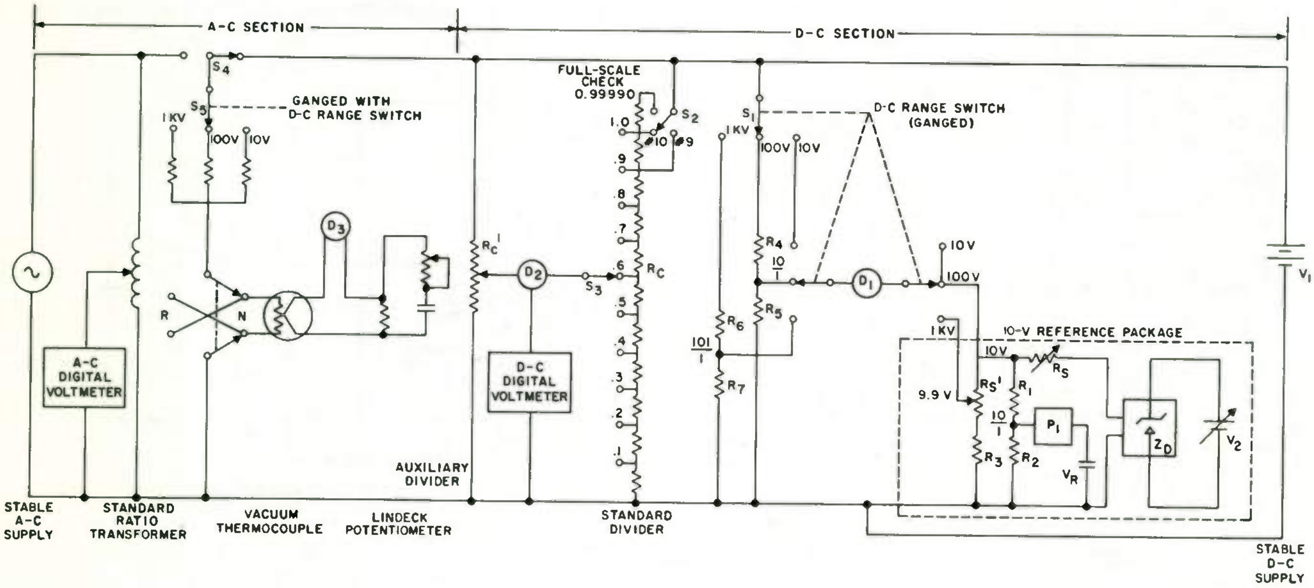


FIG. 10 VACUUM-THERMOCOUPLE A.C.-D.C. TRANSFER STANDARD AND ASSOCIATED CIRCUITRY



R_1/R_2	10-V RANGE DIVIDER
P_1	POTENTIOMETER "K" TYPE
Z_D	ZENER REFERENCE
V_R	SATURATED STANDARD CELL
R_4/R_5	100-V RANGE DIVIDER
R_6/R_7	1000-V RANGE DIVIDER
D_1, D_2, D_3	DETECTOR (SENSITIVITY ONE MICROVOLT OR BETTER

FIG. 11 COMPLETE A-C AND D-C DIGITAL-VOLTMETER CALIBRATION CIRCUIT

TRANSISTOR OPERATIONAL AMPLIFIERS

Paul J. Bénéteau, Larry Blaser and Richard Q. Lane
Fairchild Semiconductor
a division of Fairchild Camera & Instrument Corporation
Mountain View, Calif.

Transistor Operational Amplifier methods are analyzed. Criteria are established for frequency stability and drift reduction. Two examples of amplifiers using silicon planar transistors are given.

1. Introduction

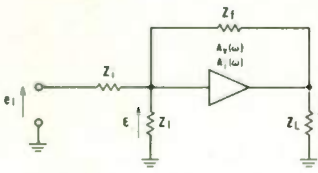
There are a great many uses for operational amplifiers in computing equipment, in instrumentation and elsewhere. Transistors have not always been considered readily adaptable to these requirements due primarily to their being essentially low impedance devices. To overcome this difficulty, Blecher[9] and Okada[1] have proposed a method in which the low-impedance property of transistors is used to advantage. For an amplifier having a constant load and gain, this method is perfectly suitable, but it unfortunately becomes rather cumbersome in the more practical case of variable gain due to the variable feedback fraction which may render the amplifier unstable.

The purpose of this paper is first to analyze in a general manner some of the operational amplifier methods suitable for both vacuum tubes and transistors; the alternatives facing the transistor circuit designer will then be discussed. Secondly, an analysis of the frequency response of these amplifiers will be made; this will include the design requirements to ensure no pulse overshoot. Some of the various drift-reduction methods will then be considered. Finally, two examples of highly stable operational amplifiers including chopper stabilization and using fast silicon planar transistors will be given.

2. Analysis of Operational Amplifiers

2.1 Operational Amplifier Methods.

A perfectly general operational amplifier is shown in Figure 1; Z_1 is the input impedance of the amplifying device (normally but not necessarily high for vacuum tubes and low for transistors), Z_o is the output impedance of the device, $A_v(\omega)$ and $A_i(\omega)$ are the voltage and current gains respectively of the device, Z_L is the load and Z_i and Z_f are the gain-determining operational resistors. The equations representing the amplifier are:



$$\frac{e_L}{e_1} = \frac{-Z_f}{Z_i}\left[\frac{1}{1-\frac{Z_fZ_1+Z_1Z_i+Z_iZ_f}{A_v(\omega)\ Z_1\ Z_i}}\right]$$

Figure 1. A General Operational Amplifier

$$e_L = \epsilon\ A_v(\omega) \qquad (1)$$

$$\frac{e_1-\epsilon}{Z_i} = \frac{\epsilon-e_L}{Z_f} + \frac{\epsilon}{Z_1} \qquad (2)$$

Define B_v as $\frac{\epsilon}{e_L}$ with $e_1 = 0$

$$B_v = \frac{(Z_i\ Z_1)/(Z_i+Z_1)}{\frac{(Z_iZ_1)}{(Z_i+Z_1)} + Z_f}$$

$$= \frac{Z_iZ_1}{Z_fZ_1 + Z_1Z_i + Z_iZ_f} \qquad (3)$$

When equations (1), (2) and (3) are solved, the operational gain A of the amplifier is:

$$A = \frac{e_L}{e_1} = \frac{-Z_f}{Z_i}\left[\frac{1}{1-\frac{1}{A_v(\omega)\ B_v}}\right] \qquad (4)$$

For large A_vB_v products, equation (4) reduces to:

$$A = \frac{e_L}{e_1} \simeq \frac{-Z_f}{Z_i}, \qquad \left(A_v(\omega)\ B_v \gg 1\right) \qquad (5)$$

For example, for an accuracy of 0.1%, A_vB_v should be larger than 1000.

For transistor amplifiers, $A_i(\omega)$ is usually of more interest than $A_v(\omega)$.

$$A_v(\omega) = A_i(\omega)\ \frac{(Z_o\ Z_L)/(Z_o+Z_L)}{Z_1}$$

Assuming $Z_f \gg (Z_o\ Z_L)/(Z_o + Z_L)$, equation (4) becomes:

$$\frac{e_L}{e_1} = \frac{-Z_f}{Z_i}\left[\frac{1}{1-\frac{Z_fZ_1+Z_1Z_i+Z_iZ_f}{A_v(\omega)\ Z_1\ Z_i}}\right] \qquad (6)$$

Now, the maximum power gain in transistor amplifiers usually occurs in the common-emitter configuration and for this type of amplifier, the voltage gain times the input impedance is approximately constant^2. Therefore, the term $A_v(\omega)\, Z_1$ in the denominator of equation (6) is approximately constant, from which it follows that the most efficient operational amplifier will be the one with the lowest input impedance possible. With this point in mind, it is now worthwhile to develop equations for this low-impedance case and to consider its drawbacks as compared to the high-impedance case.

From equation (3), for $Z_1 \ll Z_i$, Z_f ,

$$B_v \Big|_{Z_1 \text{ small}} \simeq \frac{Z_1}{Z_f}$$

Using

$$A_v(\omega) = A_i(\omega)\, \frac{(Z_o\, Z_L)/(Z_o + Z_L)}{Z_1}$$

$$= A_i(\omega)\, \frac{Z'_L}{Z_1} \;,$$

and defining

$$B_i = \frac{I_f}{I_L} \simeq \frac{-Z'_L}{Z_f}$$, the expression for the

gain then becomes, using equation (4),

$$\frac{e_L}{e_1} \simeq -\frac{Z_f}{Z_i}\left[\frac{1}{1 - \frac{1}{A_i\, B_i}}\right] \qquad (7)$$

For large $A_i\, B_i$ products, equation (7) reduces to:

$$\frac{e_L}{e_1} \simeq -\frac{Z_f}{Z_i} \qquad (8)$$

as in the high-impedance case of equation (5). A low-input-impedance amplifier having a fixed $A_i\, B_i$ product will thus be indistinguishable from a high-input-impedance amplifier having the same value of $A_v\, B_v$.

A comparison of equations (5) and (7) shows that the quantity to consider for the high-impedance amplifier is $A_v\, B_v$ while that for the low-impedance amplifier is $A_i\, B_i$. Remembering that the operational gain $A = -Z_f/Z_i$, these feedback factors B_v and B_i can be expressed approximately as:

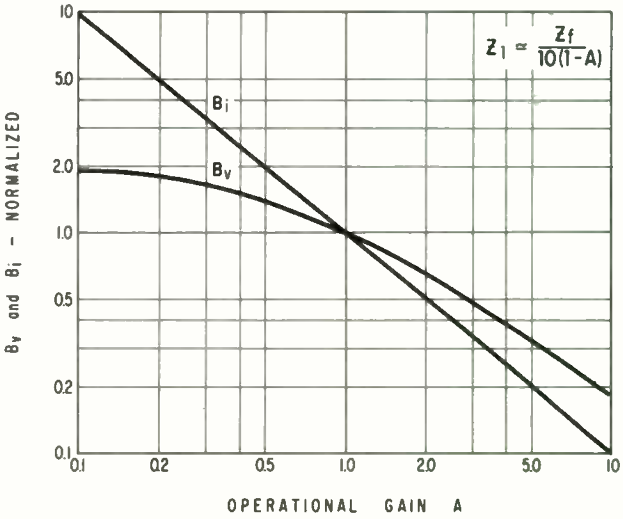
$$B_v \simeq \frac{1}{1 - A}$$

$$B_i \simeq \left(\frac{1}{A}\right)\left(\frac{Z_L}{Z_i}\right) \qquad (9)$$

where the assumption is made that the amplifier input impedance is negligibly high or low, respectively, compared with Z_i. For stability of the amplifier, both $A_v\, B_v$ and $A_i\, B_i$ will be required to have less than 360 degrees phase shift at the frequency where they become less than 1 or 0 db. When the usual 180° phase shift of the inverting amplifier is taken into account, the added phase shift is usually kept less than 135° for an adequate phase margin of 45°.

The designer of operational amplifiers has the choice of input impedance of the amplifier. Equation (6) shows that, for a given number of common-emitter transistors, optimum utilization will result when the input impedance is as low as possible. However, even though this case is the optimum one, it will not always be used for several reasons. Firstly, if the operational gain is to be changed, Z_f is usually the controlling impedance and both B_v and B_i change as is shown in equation (9). Unfortunately, the phase margin changes as the operational gain changes; this holds to a different extent for high-impedance and for low-impedance amplifiers as is shown in figure 2. It is seen that the variation of B_i, and thus of phase margin, with operational gain is far more serious than that of B_v over the gain range of 0.1 - 10. If such a large gain range is not required, then the current amplifier may be suitable, but, in general the A_i will have to be programmed along with the variable Z_f to keep the $A_i\, B_i$ product more or less constant. The same problem arises, of course, with the high-impedance amplifier but the permissible operational gain range is very much wider. For example, a gain range of 0.33 - 3 changes B_v by 3:1, which is perhaps tolerable, compared to a change in B_i of 9:1, which is probably not.

NORMALIZED FEEDBACK FRACTION VARIATIONS



To summarize, the designer of transistor operational amplifiers has an infinite choice of input impedance. The lowest possible input impedance one is the most efficient,while the highest possible input impedance one can have its gain varied most safely. There are, obviously, intermediate values which form the best compromise between these two factors. By considering equation (6), it can be shown that there is little advantage in making Z_1 less than $(0.1\ Z_f)/(1 - A)$, where A is the operational gain. If a large range of operational gains is contemplated, it may well be that a value of Z_1 greater than the above would be chosen to ensure adequate phase margins.

2.2 Frequency and Pulse Response of Operational Amplifiers.

2.2.1 Frequency Response. For the purpose of this analysis, it will be assumed that the amplifier will have a forced roll-off at a frequency ω_1 due to an RC insertion (usually at a frequency below 100 cps) plus an undesirable roll-off at a frequency ω_2 due to the open-loop frequency response. It is, of course, desirable for phase margin (stability) reasons to have ω_2 as high as possible and thus control the roll-off solely by the forced ω_1. See Figure 3.

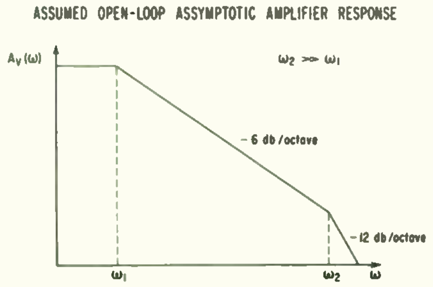


Figure 3. Assumed Open-Loop Assymptotic Amplifier Response.

The operational gain A (ω) can be written as:

$$A\ (\omega) = \frac{A_v\ (\omega)}{1 - A_v(\omega)\ B_v} \qquad (10)$$

Letting the dc open-loop gain be A_v (0), letting $p = j$, and using:

$$A_v\ (p) = \frac{-A_v\ (0)}{(1 + p/\omega_1)(1 + p/\omega_2)} ,$$

it can easily be shown that for $\omega_1 \ll \omega_2$ and $A_v B_v \gg 1$, equation (10) reduces to;

$$A\ (p) = \frac{A_v\ (0)\ \omega_1\ \omega_2}{p^2 + \omega_2 p + A_v(0)\ B_v \omega_1 \omega_2} \qquad (11)$$

The exact closed-loop-3db frequency ω_{2c} is given by a rather complicated expression. It can, however, be simplified by assuming the influence of ω_2 on frequency response is small compared to that of ω_1. If this is done, then ω_{2c} is easily obtained from equation (11) and

$$\omega_{2c} \approx A_v(0)\ B_v\ \omega_1 \qquad (12)$$

This means, of course, that only for a 6db/octave slope of open-loop response is there direct gain-bandwidth trading. If the influence of ω_2 is not negligible, then equation (12) is not valid, and the more exact expression given in the appendix will have to be used (equation 23A).

2.2.2 Pulse Response. A sufficient but not necessary condition for no pulse overshoot can be very easily obtained from equation (11) by setting the discriminant of the denominator to a quantity greater than or equal to zero. This ensures that the poles of the gain function do not leave the real axis[3].

i.e. $b^2 - 4\ ac \geq 0$

$$\omega_2{}^2 - 4\ A_v(0)\ B_v \omega_1 \omega_2 \geq 0$$

or $$\omega_2 \geq 4\ A_v(0)\ B_v\ \omega_1 \qquad (13)$$

This is usually a rather stringent requirement since, if fast pulse response is required, either ω_1 or A_v (0) B_v will be high.

2.3 Drift Stabilization

One of the principal requirements of an operational amplifier is that output drift be small. The early stages of the amplifier are the main sources of drift since they are followed by the remaining gain of the amplifier. In this respect, drift is analogous to noise and if it is possible to precede the amplifier with another high-gain low drift amplifier, the drift at the output for a given operational gain may be greatly reduced. This is analogous in a receiver to using a high gain, low noise preamplifier before a stage of noisy mixing in order to increase the signal to noise ratio at the receiver output.

2.3.1 Methods of Drift Stabilization. In transistor direct-coupled amplifiers the main sources of drift are as follows:

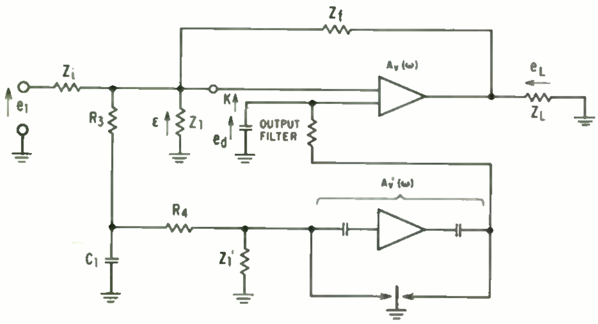
1. The positive temperature coefficient of the emitter current gain h_{FE}.
2. The positive temperature coefficient of the emitter current gain h_{fe}.
3. The positive temperature coefficient of collector to base leakage current I_{CBO}.

By using silicon transistors having surface passivated or "planar" construction the contribution of I_{CBO} to input drift may generally be neglected.

The use of differential input stages having common heat sinks or particularly two transistor die on a common header greatly reduces drift sources due to the effect of the differential temperature coefficient of V_{BE} and h_{FE}. Planar transistors have a much smaller spread in these temperature coefficients than had earlier transistors, thus rendering a better match possible thereby further reducing differential changes.

In computing applications, however, the small drifts obtainable by the above methods [5 μV/°C at the summing junction in the best case[4] and 10μV/°C in a typical high-performance amplifier(s)] may still be not acceptable and the drift must be further reduced by inserting a high gain low drift amplifier between the summing junction and the direct coupled amplifier.

Figure 4 shows a drift reduction scheme suggested by Goldberg[6], in which the high gain low drift amplifier is formed by an ac amplifier with half wave modulation and demodulation. This scheme is analyzed in the appendix and only the results will be quoted here.



$$e_L = -\frac{e_1 Z_f}{Z_i} + \frac{K z^3}{Z_1 Z_i Z_3 [1 - A_v'(\omega)]}$$

Where $Z_3 = R_3 + R_4 +. Z_1'$

and $z^3 = Z_1 Z_3 Z_i + Z_f Z_1 Z_i + Z_f Z_1 Z_3$

Note: Primed values always refer to chopper circuit.

Figure 4. Goldberg Drift Reduction Scheme

e_L = output voltage at load

e_1 = input signal

ϵ = voltage at summing junction

k = drift potential

e_d = correction voltage from chopper amplifier

C_1R_4 forms a filter to reduce the transmission of the chopped drift voltage into the summing junction. R_3 isolates C_1 from Z_1 in order to avoid loading the amplifier input at frequencies where C_1 has little reactance.

$$e_L \simeq -\frac{e_1 Z_f}{Z_i} + \frac{k\,z^3}{Z_1 Z_i Z_3 \left[1 - A_v'(\omega)\right]} \qquad (14)$$

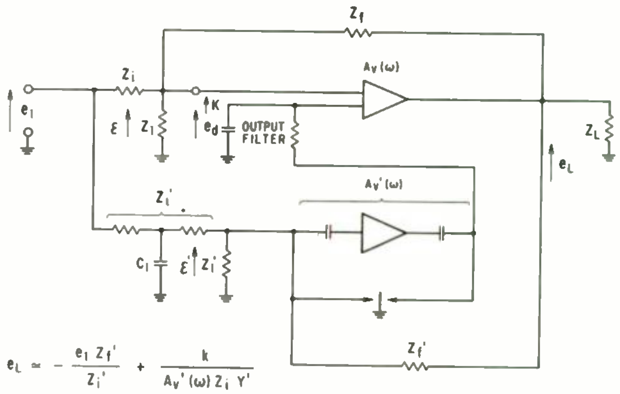
where $Z_3 = R_3 + R_4 + Z_1'$

and $z^3 = Z_1 Z_3 Z_i + Z_f Z_3 Z_i + Z_f Z_1 Z_i + Z_f Z_1 Z_3$

Equation (14) shows that the drift is reduced directly as A_v' (ω), the gain of the chopper amplifier, provided A_v' (ω) >> 1.

The scheme shown in Figure 4 suffers one serious disadvantage, that is the inability of the chopper amplifier to correct for errors due to base current flowing in Z_i and Z_f. A common remedy is to insert a low leakage capacito at point S. This eliminates the problem, but if the amplifier is driven to saturation the capacitor can acquire a charge, thus blocking A_V and requiring that a shorting or reset switch be placed across the capacitor.

Blecher[9] and Okada[1] suggest an alternative scheme, Figure 5, in which separate summing impedances are provided for the low drift amplifier in order to avoid the capacitor.



$$e_L \simeq -\frac{e_1 Z_f'}{Z_i'} + \frac{k}{A_v'(\omega) Z_i Y'}$$

Where $Y' = \dfrac{Z_1'}{Z_f' Z_i' + Z_1' Z_i' + Z_1' Z_f'}$

Figure 5. Alternative Drift Reduction Scheme

These impedances Z_i' and Z_f should have the same ratio as Z_i and Z_f.

i.e. $$\frac{Z_f'}{Z_i'} = \frac{Z_f}{Z_i}$$

e_1 = input signal

e_L = output voltage at load

ϵ = voltage at direct coupled amplifier summing junction

ϵ' = voltage at chopper amplifier summing junction

k = drift potential

e_d = correction voltage from chopper amplifier.

$$e_L \simeq -\frac{e_1 Z_f'}{Z_i'} + \frac{k}{A_v'(\omega)\, Z_i'\, Y'} \qquad (15)$$

$$\text{where } Y' = \frac{Z_1'}{Z_f' Z_i' + Z_1' Z_i' + Z_1' Z_f'}$$

Equation (15) shows that the drift is again reduced directly as $A_v'(\omega)$.

The scheme of figure 5 suffers the disadvantage of more complicated switching if a range of operational functions is to be performed, but is, however, free of the blocking problem.

2.3.2 The Effect of the Chopper Amplifier and its Associated Filters on Overall Stability. The Goldberg scheme of Figure 4 requires two filters, the chopper amplifier output filter which smooths the half wave demodulated output from the chopper amplifier and $R_3R_4C_1$ which isolates the summing junction from the modulating side of the chopper.

The alternative scheme of Fig.5 requires the same output filter but may be able to dispense with the input filter if Z_i' is much larger than the impedance of the signal source and the input contains no components that will beat with the chopper frequency. This is a considerable advantage of the scheme because the condition given by equation 21A, in the Appendix,

$$\omega_1' \geq \frac{4\,\omega_2'\, A_v'(0)\, Z_i'\, Y'}{Z_i\, Y},$$

might other wise be difficult to meet without making ω_2' very small thus reducing the speed of drift correction. A restricted chopper amplifier bandwidth will have the same effect as a low ω_1'. The chopper amplifier output impedance Z_o' should be made low2 in order that the phase shift of the modulation due to the phase of Z_o' occurs well beyond the radian frequency.

$$\frac{4\,\omega_2'\, A_v'(0)\, Z_i'\, Y'}{Z_i\, Y}.$$

It should be realized however that the quantity

$$\frac{4\, A_v'(0)\, Z_i'\, Y'}{Z_i\, Y}$$

dictates the separation between ω_1' and ω_2'. In other words both

$$\omega_1' \geq \frac{4\, A_v'(0)\, Z_i' Y' \omega_2'}{Z_i Y}$$

$$\text{and} \qquad \omega_2' \leq \frac{4\, A_v'(0)\, Z_i' Y' \omega_1'}{Z_i Y}$$

are stable conditions provided of course that they do not conflict with the inequalities required with respect to ω_1 and ω_2 of the wideband amplifier.

3. Amplifier Design Considerations

3.1 Wide-Band Amplifier

The main design requirements of the wide-band amplifier are generally as stated below in order of relative importance:

(1) The amplifier should have one dominant roll off only, any other corner frequencies must be higher than $A_v(0)\, B_v \omega_1$. If this is not practicable then the unwanted negative corners will have to be cancelled with positive corners by using lead networks or by placing a positive corner in the dominant roll-off network.

There are two principal causes of negative corner frequencies. These are (i) the frequency dependence of common emitter current gain which produces a corner at $f_\beta = f_t/\beta_o$; and (ii) the collector to base capacitance of the transistor which when operating in the common emitter configuration is multiplied by $(1 - A_v)$ where A_v is the stage voltage gain*. This Miller capacitance appears in shunt with the input impedance of the stage causing a corner at approximately

$$\frac{1}{2\pi(1 - A_v)\, C_{ob}\, h_{ie}}.$$

When using "fast" transistors (i.e.,f >1mc/s) the corners due to Miller capacitance are usually more troublesome than those due to f_β . However, by ensuring that each common emitter stage drives into a low load impedance the voltage gain of each stage may be restricted thereby reducing the factor by which C_{ob} is multiplied. This occurs naturally in a cascade of common emitter stages if each following stage has a higher collector current than the preceding stage. A large voltage gain may then be obtained in a later or in the final stage by using the common base configuration which is free of the Miller capacitance problem due to having no phase inversion.

(2) This amplifier should have minimum drift before stabilization since this reduces the gain of the chopper amplifier for a given overall drift. A further advantage of the reduced chopper amplifier gain is the reduced separation required between ω_1' and ω_2' as discussed in Section 2.3.1. Differential input stages minimize the drift and also have the advantage of much better common-mode rejection.

*Generally, the low-frequency value of A_v is used. However, under certain conditions of cascaded transistor amplifiers, the fact that A_v is generally complex will lead to a more complicated input network consisting of a capacitor in series with a parallel RC.

(3) The amplifier should have low noise at the output. This can be readily achieved by placing the dominant roll off at the output in order to reduce the high frequency noise components developed in the whole wide-band amplifier. Unfortunately placing the roll off at the output reduces the available output at high frequencies unless a very large quiescent current is maintained. The positioning of the roll off is therefore a matter of compromise, as when it is moved toward the input the noise at the output increases, but so does the swing available at high frequencies.

3.2 Chopper Amplifier.

The chopper Amplifier, being ac coupled, has negligible drift and therefore may have a single-ended input. If the assumption is again made that $A_v'(\omega)\ Z_1'$ is a constant in the common-emitter configuration, inspection of equations (14) and (15) shows that the maximum drift reeuction of the operational amplifier occurs as Z_1' tends to zero. However, precisely the same considerations of Section 2.1 apply in this case, namely that the input impedance will generally be chosen to suit the compromise between low-impedance high-efficiency amplifiers and high-impedance very stable amplifiers. Specifically, there is little advantage in making Z_1' less than $(0.1\ Z_f')/(1-A)$; as previously stated, it may well be that a value of Z_1' greater than the above would be chosen to ensure sufficient phase margin over the contemplated range of operational gains.

The bandpass of the ac amplifier should be sufficiently wide to ensure negligible phase-shift of the modulation, i.e., the upper half-power point should be much greater than that of the output filter (ω_2'). The output impedance should be low as discussed in Section 2.3.2.

Generally, the chopper amplifier is of straightforward design, and few precautions need be taken.

4. Experimental Amplifiers

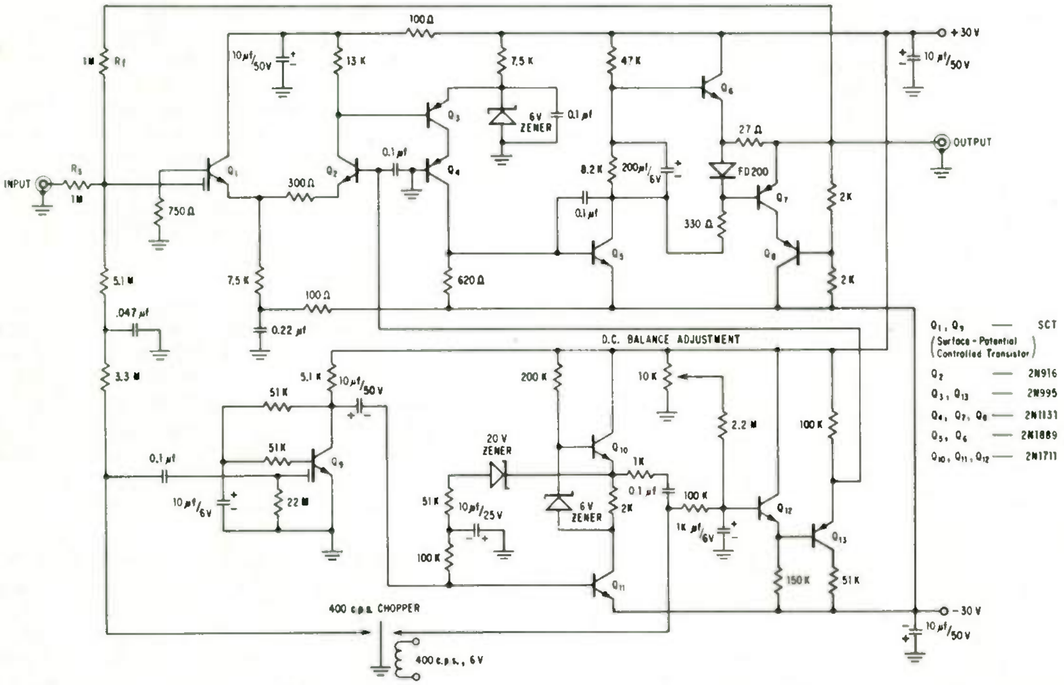
4.1 High Input Impedance Amplifier.

This amplifier, Fig. 6 has a differential input stage using the new Surface Controlled Transistor[7,8] , which equivalent circuit appears in Appendix VI.

The grid impedance of this device is essentially capacitive, approximately 16 pf in the early device used here, and therefore renders the feedback fraction B_v dependent only on the ratio $Z_i/(Z_i + Z_f)$ and independent of the parallel value $Z_i Z_f/(Z_i + Z_f) = Z_p$. Furthermore since no bias current flows in the summing components it is not necessary to provide Goldberg's blocking capacitor or Blecher and Okada's additional summing network for the chopper amplifier. The chopper amplifier has a SCT front end in order to reduce the loading of the summing junction.

The sole effect of increasing Z_i and Z_f is the lowering of the corner, $\omega_{SCT} \approx 1/(C_g Z_p)$ introduced by the SCT grid capacitance. Naturally this reduces the separation between ω_1 and ω_2 and can cause pulse overshoot which may be objectionable in such an application as analog to digital conversion. The performance of this amplifier appears in figures 8, 9, 10 and 11.

HIGH INPUT IMPEDANCE OPERATIONAL AMPLIFIER

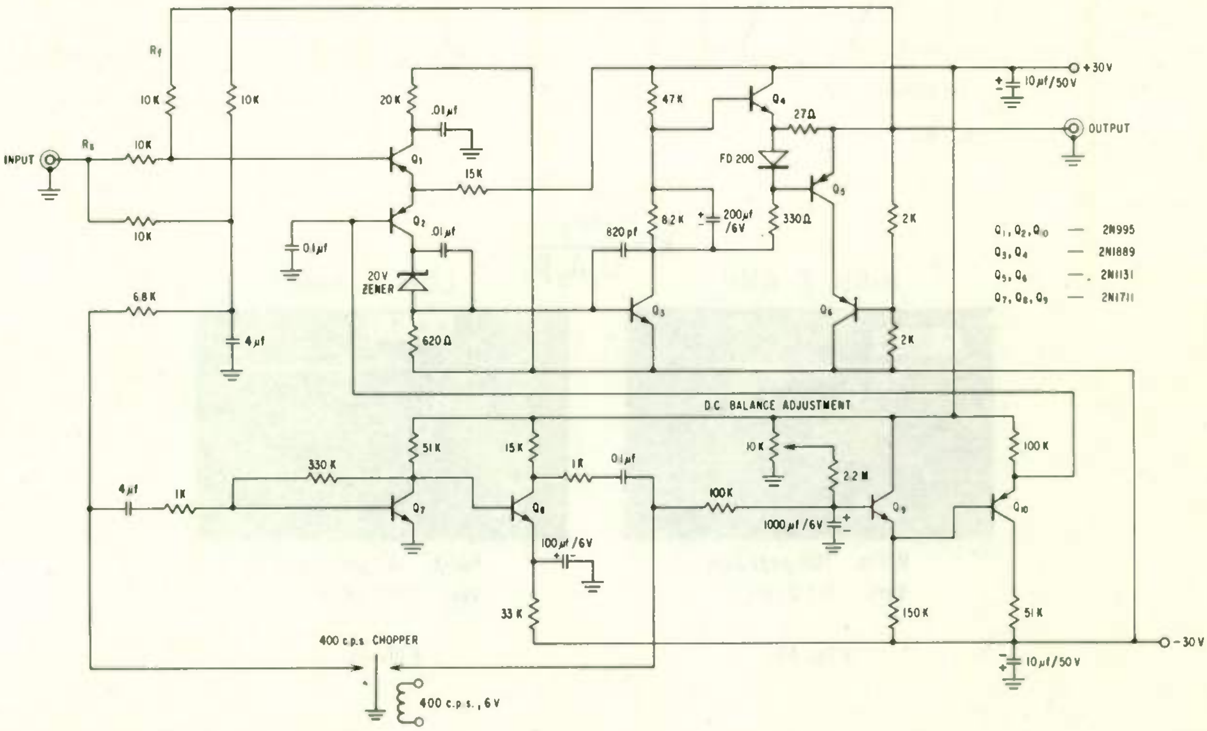


4.2 Low Input Impedance Amplifier.

The amplifier of figure 7 uses a similar "bootstrapped" output amplifier as is used in the high input impedance amplifier of figure

A differential pair of 2N995 PNP transistors loaded essentially by the input impedance of Q_3 forms a wideband input stage having an input impedance of approximately 1 kΩ. It is recalled from the discussion of the constant input impedance voltage gain product of common emitter stages, that $Z_1 \approx Z_f/10$ was shown to be the optimum for gain accuracy. This amplifier was designed to use summing components of less than 10kΩ. In order to avoid an offset caused by base current flowing in the summing resistors, a separate pair of summing resistors are provided for the chopper amplifier. The performance of this amplifier appears in figures 8, 9, 12 and 13.

LOW INPUT IMPEDANCE OPERATIONAL AMPLIFIER



References

1. Stable Transistor Wideband DC Amplifiers. Robert H. Okada, Transactions of the AIEE, March 1960, pp. 26.

2. Design Considerations in a Chopper-Stabilized Transistor Operational Amplifier. R.O.Gregory, Presented at IRE Maecon, Kansas City, Mo., Nov.15-16, 1960.

3. Circuit Theory and Design, John L. Stewart, John Wiley & Sons, Chapter 10.

4. A New D-C Transistor Differential Amplifier. D.F.Hilbiber. Presented at Solid State Circuits Conference, Philadelphia, Pa., Feb.15, 1961. Also available as TP-16, Fairchild Semiconductor, Mountain View, Calif.

5. The Design of High Stability DC Amplifiers. Paul J. Bénéteau, Semiconductor Products, February 1961, pp.27-30. Also available as APP-23, Fairchild Semiconductor, Mountain View, California.

6. Stabilization of Wide-Band Direct-Current Amplifiers for Zero and Gain. Edwin A. Goldberg, RCA Review, June 1950, pp.296-300.

7. A New Semiconductor Tetrode - The Surface-Potential Controlled Transistor. C.T.Sah, IRE Proc., November 1961, pp.1623-1634.

8. Applications of the Surface-Potential Controlled Transistor Tetrode, H.Z.Bogert, D.A.Tremere and C.T.Sah. IRE Solid State Circuits Conference, Philadelphia,Pa.,Feb.1962.

9. Transistor Circuits for Analog and Digital Systems. Franklin H. Blecher, BSTJ, March 1956, pp.295-332.

10. D.C.Amplifiers for Use in Analogue Computers. C.M.Cundall, et al. Proc. IEE, April 1960 pp.1354-1364.

11. A Drift-Compensated D.C.Operational Amplifier Employing a Low-Level Silicon Transistor Chopper. W.Hochwald and F.H.Gerhard,Proc.IRE,NEC, Chicago,Ill.,1958,Computers,Vol.14,pp.798-810.

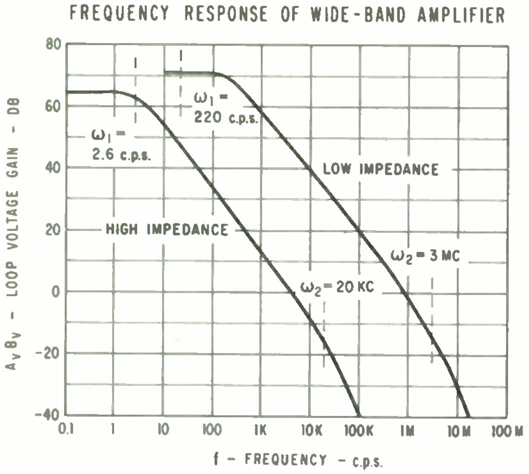


Fig. 8.

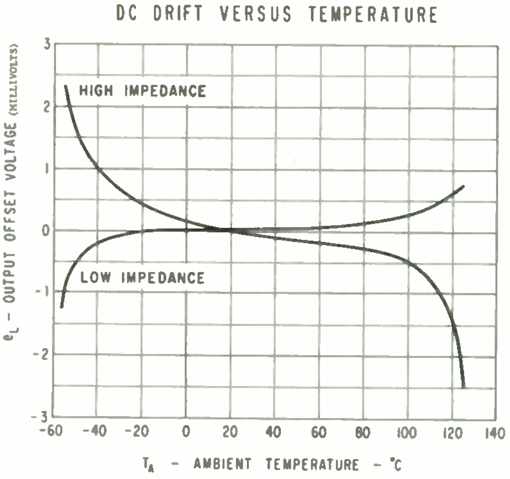
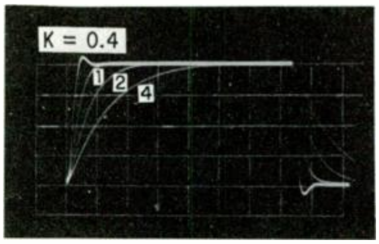


Fig. 9.

$$K = \frac{\omega_2}{\omega_1 A_V \beta_V}$$

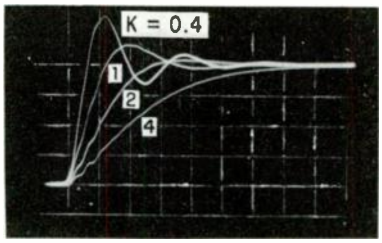
HIGH Z AMP



Horiz. 100 μsec/cm.
Vert. 0.1 V/cm.

Fig. 10.

LOW Z AMP



Horiz. 0.1 μsec/cm.
Vert. 0.1 V/cm.

Fig. 12.

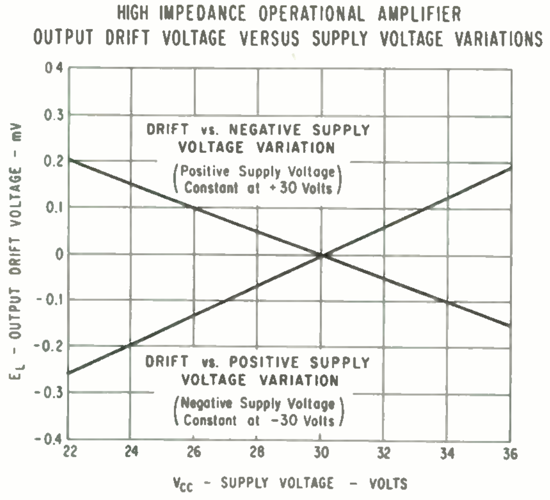


Fig. 11.

From the wideband amplifier closed-loop response given by equation (11), the closed-loop 3 db frequency ω_{2c} can easily be calculated by noting that the numerator is not a function of p. ω_{2c} is then very nearly given by the lower of the two poles and

$$\omega_{2c} \approx \frac{\omega_2}{2}\left(1 - \sqrt{1 - \frac{4\, A_v(0)\, B_v\, \omega_1}{\omega_2}}\right) \tag{23A}$$

For the case where ω_2 is very large, the root can be approximated by

$$1 - \frac{2\, A_v(0)\, B_v \omega_1}{\omega_2} ,$$

and ω_{2c} is then given by

$$\omega_{2c} \approx A_v(0)\, B_v \omega_1 \tag{24A}$$

This is the same result that was found in eq.(12).

A.VI. Equivalent Circuit of Surface-Controlled Transistor.

A small modification of the familiar hybrid-pi transistor equivalent circuit shown in Fig.14 gives a useful representation of the observed characteristics of the S.C.T.

$$\frac{e_L}{e_g} = \frac{g_m\, Z_t\, R_L}{\beta_o\, r_e \left[1 + Z_t\, p\, C_c\, (1+R_L/r_e)\right]} \tag{25A}$$

$$\text{where } Z_t = \frac{(R_b + r_b')\, Z_e}{R_b + r_b' + Z_e} , \quad Z_e = \frac{(\beta_o + 1)\, r_e}{1 + p/\omega_\beta} ,$$

$$\text{and} \quad g_m = \partial I_c \big/ \partial e_g \Big|_{e_o = 0} .$$

The ω_{3db} of the SCT is given by

$$\omega_{3db} = \frac{R_b + r_b' + \beta_o\, r_e}{(R_b+r_b')\left[\dfrac{1}{\omega_\beta} + \beta_o C_c (R_L+r_e)\right]} \tag{26A}$$

$$= \frac{1 + \dfrac{\beta_o\, r_e}{R_b + r_b'}}{\dfrac{1}{\omega_\beta} + \dfrac{1}{\omega_m}} \tag{27A}$$

$$\text{where} \quad \omega_m = \frac{1}{\beta_o\, R_L\, C_c} \quad \text{if} \quad R_L \gg r_e . \tag{28A}$$

Note that $C_1 + C_2 \approx 16$ pf for the early device used. From equation (25A), setting p = 0 and re-arranging,

$$\text{Low-Frequency gain} \approx \frac{g_m R_L}{1 + h_{ie}/R_b} \tag{29A}$$

From equations (27A) and (29A), it is apparent that by manipulation of R_b, the base source resistance, low frequency gain may be traded for bandwidth.

For further details regarding the SCT, see references (7) and (8).

TRUE RMS MEASUREMENTS UTILIZING THE GALVANOMAGNETIC EFFECTS IN SEMICONDUCTORS

Max Epstein and Larry J. Greenstein
Armour Research Foundation of
Illinois Institute of Technology
Chicago, Illinois

Summary

The development of high-mobility intermetallic semiconductors results in a more efficient application of the Hall and magnetoresistive effects in the design of various electronic devices. Although the two effects are related, their utilization in true rms measurements is different. The Hall potential is proportional to the product of the magnetic flux density normal to the Hall element surface and the current through the element. This proportionality holds over a wide range of magnetic fields and currents. Hence, if the Hall element is constructed in such a way that the magnetic field is produced by the same current which passes through the element, the square of the signal over a wide range of amplitude can be obtained. The magnetoresistance provides a change in resistance proportional to the square of the applied magnetic field. The magnetoresistive effect can be made significant when the semiconductor specimen is in the form of a Corbino disk.

Both the Hall effect and magnetoresistance devices require the design of an efficient magnetic circuit. The design involves the use of high permeability materials. The use of ferrites facilitates the construction of such a circuit and the preparation of a very thin Hall element or Corbino disk. Since the magnetic flux density is obtained by means of a current-carrying coil, a high impedance of the signal source is required to obtain wideband operation.

The design of the low-pass weighting circuit at the output of the squaring device requires considerations of the type of signal which is being measured.

Introduction

One of the most important parameters of an electrical signal is its root-mean-square (rms) value. This quantity provides a measure of the energy content of the signal and can be mathematically expressed by

$$V_{rms} \equiv \left[\lim_{T \to \infty} \int_0^T v^2(t)\, dt \right]^{1/2} \qquad (1)$$

where v(t) is the electrical signal for which V_{rms} is the root-mean-square value.

Most present-day meters which provide so-called rms readings measure the full-wave rectified average of the signal and relate it to the rms value by a form factor of $\pi/2\sqrt{2}$. That is, the rms reading presented on the output scale of such meters is approximately

$$V'_{rms} \cong \frac{\pi}{2\sqrt{2}} \lim_{T \to \infty} \int_0^T v(t)\ dt \qquad (2)$$

Whereas this relationship is valid for pure sinusoids, it does not apply in general to complex waveforms. Consequently, accuracy in measuring rms values of non-sinusoidal waveforms necessitates the use of appropriate squaring circuitry.

RMS instruments which incorporate a squaring function in their operation are defined here as true rms meters. The following discussion presents basic considerations in the design of a true rms meter in which the squaring function is performed by utilizing galvanomagnetic effects in intermetallic semi-conductors.

Basic Elements of a True RMS Meter

A simplified block diagram of a true rms meter is shown in Fig. 1. The integrating circuit is generally a low-pass filter which approximates the averaging process indicated by Equation (1). In some instruments the averaging is inherent in the squaring and/or readout devices. The square-root readout circuit is designed so that the values read from the output amplitude scale correspond to the square-root of the input. If a linearly calibrated amplitude scale is desired, then it must be preceded by an appropriate square-rooting circuit.

The squaring devices utilized in true rms meters fall into three basic categories, namely, (1) electromechanical, (2) thermal, and (3) electronic. The electromechanical method of squaring utilizes the interaction between current and magnetic fields, the force being proportional to the product of their magnitudes; e.g., a D'Arsonal movement with an electromagnet. Such devices are applicable to signals of very low frequencies only. The thermal squarers utilize the conversion of electrical energy into thermal energy. If the conversion is obtained in a resistive device such as a thermistor, the change in its resistance indicates the power dissipated. Such instruments suffer from a limited dynamic range of operation. One type of electronic squarer consists of a combination of linear resistors and voltage biased diodes arranged so that a square-law characteristic is approximated by a continuous sequence of linear segments. This method also suffers from limited dynamic range.

Another type of electronic squarer can be obtained by exploiting galvanomagnetic effects in solid state devices. The use of Hall effect in intermetallic semi-conductors results in a squaring device capable of wideband operation over a large dynamic range of amplitudes. Furthermore, such a device can be made quite compact and simple in construction. Another galvanomagnetic phenomenon in semi-conductors is the magnetoresistive effect, which provides the possibility of a squarer having an even simpler configuration than that using the Hall effect. Whereas the magnetoresistive squarer operates over a frequency range comparable to that of the Hall effect squarer, the latter has a larger dynamic range.

Before discussing the construction and performance of these galvanomagnetic squarers, it will be useful to discuss briefly the principle of both the Hall and magnetoresistive effects. These concepts are introduced in the next section.

Galvanomagnetic Effects

Hall Effect

The Hall effect occurs when a transverse magnetic field is applied to a current-carrying conductor. The magnetic field exerts a force F (Lorentz force) on the current-carrying charges in the direction perpendicular to both the current I and the magnetic flux density B , Fig. 2. The deflected charges produce an electric field E_H (Hall field) in the direction of the force.

The Hall potential measured across the width of the specimen, Fig. 2, can be shown to be[1]

$$V_H = \frac{R_H}{t} B \times I \qquad (3)$$

where R_H is the Hall coefficient. The Hall potential in Equation (2) is given in volts when I is in amperes, B is in webers per meter - squared, R_H is in meters - cubed per coulomb, and t (thickness) is in meters.

If the current and the magnetic field are made proportional to two given signals, the Hall output voltage is proportional to the product of the two signals. This product is independent of frequency, resulting in many applications of such a device to broadband multiplication. If I and B are both made proportional to the same signal, then the device functions as a squarer.

Magnetoresistance

The Hall field balances out the Lorentz force for all of the charge carriers which move with velocities near or equal to the average velocity. Those charge carriers will therefore continue to move along the same path as in the absence of the magnetic field (path a in Fig. 2). However, charge carriers which move with velocities considerably larger or smaller than the average velocity will be deflected along either paths b or c in Fig. 2. The deflected charges acquire smaller drift velocities between collisions which result in smaller mobilities and thus in reduced conductivity or increased resistivity. This change in resistance is defined as the physical magnetoresistence.[2] For small magnetic fields the relative increase in resistivity is proportional to the square of the applied magnetic flux density. For high fields the physical magnetoresistance becomes linear and approaches an upper limit asymptotically as B increases indefinitely.[3]

If the geometrical configuration of the specimen is chosen in such a way that the Hall field is eliminated, the Lorentz forces on the charge carriers are not balanced out. Consequently, the charge carriers move in larger trajectories and thus contribute to a larger magnetoresistive effect. This contribution is known as the geometrical magnetoresistance. The geometry for which this phenomenon is most pronounced is that of the Corbino disk,[4] which can provide a change of resistance (in the presence of a transverse magnetic field) considerably greater than that obtained from the physical magnetoresistance.

Figure 3 shows a Corbino disk with one electrode at its center and the other along its circumference. In the absence of a magnetic field B the current flows radially between electrodes, so that a circumferential Lorentz force is established when B is applied. Due to the disk geometry, no charge separation can be sustained in that direction and hence, no Hall field can exist. Instead, a circumferential component of current flows (in addition to the radial component) which thus increases the resistive path between the electrodes. This can be thought of as an increase in the resistivity of the specimen in the radial direction. This resistivity is given by[5]

$$(B) = \frac{1 + [\sigma(B)\, R_H B]^2}{\sigma(B)} \qquad (4)$$

where $\sigma(B)$ is the volume conductivity in the presence of a magnetic field due to the physical magnetoresistance only and is independent of the geometry of the specimen.

Construction of the Squaring Devices

Choice of Material

The application of Hall effect and magnetoresistance to the design of an rms measuring device requires considerations of the output signal magnitudes that can be practically achieved. The galvanomagnetic effects are largest in materials which have high mobilities of one type of carrier. The highest known mobilities are found in n-type indium antimonide, an intermetallic semiconductor compound of group III - V. The electron mobility of single crystal indium antimonide at room temperature is about 60,000 cm^2/volt-sec. In the work reported here the material used was n-type indium antimonide having a carrier concentration of 10^{15} per cm^3 at -196°C (type 7-N made by Ohio Semiconductors, Inc., Columbus, Ohio). It should be noted that indium antimonide has a very narrow energy gap, resulting in large variations of the Hall coefficient and resistivity with temperature. Other compounds, such as indium arsenide, exhibit better temperature characteristics with a somewhat lower mobility.

Construction of the Hall Effect Squarer

The Hall effect squarer consists of a coil wound on a magnetically permeable core and a Hall element placed in the gap of the core. The signal to be measured is applied in the form of a current to the series combination of the coil and Hall element.

In order to obtain an efficient magnetic circuit, i.e., high magnetic flux density in the gap of the core structure for a given current, the device is built in a very small cup core of ferrite material (type H made by General Ceramics). The Hall element is located between the center cores of the two cups, and a coil wound on a nylon bobbin is placed inside, Fig. 4. This construction results in a closed magnetic structure having minimum leakage and vulnerability to external fields.

The detailed construction of this assembly is as follows: A specimen of indium antimonide 3/16 inch x 1/16 inch and .010 inch thick is glued to a ferrite wafer of the same area and about .040 inch thick. The semiconductor is then lapped down to a thickness of about .0005 inch. The contacts are made by soldering thin copper wires using low-temperature, lead-free, stainless steel solder dipped in Eutector flux to prevent oxidation of the tin during the soldering process. Another wafer of ferrite having the same dimensions as the previous one is then glued on top of the semiconductor. This assembly is placed in the gap between the center legs of the cup-core, Fig. 4. The fabrication of a very thin semiconductor waver results not only in an efficient Hall element, as indicated by Equation (3), but also in a very small magnetic circuit air gap, thereby increasing the efficiency of the entire device. Also, the reduction in effective area of the center leg of the cup-core increases the magnetic flux density at the element, thus increasing the Hall output. The magnetic field is obtained by passing the signal current through a 25 turn coil of No. 38 magnet wire, which is located inside the cup-core, Fig. 4. The coil leads are brought outside through openings in the cup-core.

The wires from the Hall element are brought outside and connected to tin contacts previously wetted onto the cup-core by ultrasonic soldering. Figure 5 shows the completed assembly. A one cent coin is also shown to indicate the size of the device.

Construction of the Magnetoresistive Squarer

The method of construction of the magnetoresistive squarer is very similar to that of the Hall effect device. In this case, however, the semiconductor wafer is in the form of a very thin disk 3/16" in diameter (the diameter of the center leg of the cup-core), and only two contacts are made; one to the center of the disk and the other to its circumference. In all other respects the two assemblies are made identically.

Experimental Results

Hall Effect Squarer

Figure 6 shows a schematic diagram of the Hall effect squarer. To determine the dynamic range of the device, a sinusoidal signal at 1000 cps was applied and the input current level was monitored. Table I shows the output voltage measured with an HP 425A d-c voltmeter.

Table I
Input-Output Characteristics of the Hall Squarer

Current Input (milliamperes) (f = 1000 cps)	D-C Voltage Output (millivolts)
.316	0.003
1	0.030
3.16	0.3
10	3
31.6	30
100	300

The lower limit of the input current is in general, determined by the sensitivity of the readout device. The practical upper limit of the current (100 ma for the device constructed) is established by considerations of Hall element heating and Hall coefficient variations. To maximize the dynamic range of the circuit, the number of turns of the magnetic field coil is chosen so that the ferrite just begins to saturate at the maximum allowable current. For the constructed cup-core device the appropriate number of turns is 25, as indicated in the discussion on construction.

The frequency response of the squarer was tested using a current input of 10 ma. The output signal remained constant at 3 mv up to a frequency of 200 kc, dropped to 2.75 mv at 300 kc, and to 2.4 mv at 400 kc. The finite frequency response of the device is due primarily to the core losses and capacitance associated with the field producing coil. These factors contribute to a frequency-dependent disparity, in both amplitude and phase, between the magnetic field and the Hall element current.

Magnetoresistive Squarer

The Corbine disk was connected in a bridge circuit as shown in Fig. 7. A direct current of 100 ma was passed through the disk and the bridge was balanced. As in the case of the Hall effect squarer, the output was measured by an HP425A d-c voltmeter. Table II shows the input-output data for a sinewave test signals at 1000 cps.

Table II
Input-Output Characteristics of the Magnetoresistive Squarer

Current Input (milliamperes) (f = 1000 cps)	D-C Voltage Output (millivolts)
1	0.0026
3.16	0.026
10	0.260
31.6	2.6
100	26

The lower and upper current limits for the magnetoresistive squarer are established by the same considerations as for the Hall effect squarer.

The frequency response of the device was found to be uniform up to at least 100 kc. In the present portion of the study, improved bridge balancing techniques are being devised so that the performance of this squarer can be evaluated at higher frequencies. The core losses and capacitance associated with the field producing coil are expected to limit the operating bandwidth of the magnetoresistive squarer to approximately that of the Hall effect squarer.

Integrating and Readout Devices

Many of the devices which have static square-law transfer characteristics, particularly those of the thermal and electromechanical type, respond sluggishly to time-varying inputs. Consequently, their outputs are effective integrations of the squared inputs. The Hall effect and magnetoresistive squarers, however, respond instantaneously (for all practical purposes) to applied stimuli and must therefore be followed by suitable averaging devices to achieve rms information. In many cases this averaging process is at least partially provided by the readout device, which may be a pen recorder, dc meter, or other such instrument. In general, however, additional signal weighting circuitry at the squarer output is also necessary in order to obtain the degree of averaging desired.

Let us assume that the integrating (or weighting) circuit of Fig. 1 is a linear, passive network incorporating the combined frequency response of the readout device and any additional circuitry inserted at the squarer output. If the impulse response of this circuit is h(t), then the circuit output $V_o(t)$ is given by the convolution integral

$$V_o(t) = k \int_0^t h(t-\lambda)\, v^2(\lambda)\, d\lambda \qquad (5)$$

where k is the proportionality constant of the squaring circuit. It is seen from Equation (5) that the integrator output is a weighted average of the input $v^2(t)$ up to the instant of measurement, the weighting being performed by the impulse response of the circuit.

Ideally, h(t) should be of such a form that, after the integrator input has been applied for a reasonable length of time the output closely approximates

$$k \left[\lim_{T \to \infty} \frac{1}{T} \int_0^T v^2(t)\, dt \right]$$

The degrees of freedom available in synthesizing the weighting circuit should be utilized so as to achieve this objective. This involves two basic steps, namely, (1) establishing some time duration T' such that the rms value of the applied signal within any interval of magnitude T' is approximately equal to its rms value evaluated over all time ; and (2) synthesizing the network so that h(t) is close to 1/T' over a time interval of T'', where T'' $\gg$ T', and close to zero elsewhere.

The first step stated above requires some knowledge of the signal being measured. For example, if v(t) is known to be a periodic signal, then T' should be a multiple of, or several times larger than, the fundamental period. As another example, if v(t) represents a stationary random process then T' should be several times larger than the maximum time separation for which the associated autocorrelation function is significant.

For a specified value of T', the theoretical optimum weighting circuit is one having a transfer characteristic given by

$$F(\omega) = e^{-\frac{1}{2} j\omega t} \frac{\sin(\frac{1}{2}\omega T')}{(\frac{1}{2}\omega T')} \qquad (6)$$

Such a network has a rectangular impulse response of duration T' and height 1/T'. Although such a circuit is not physically realizeable it can be approximated by means of sophisticated synthesis techniques. In practice however, it is often more convenient to utilize simple integrating circuits which compensate for the crudeness of their approximation to the ideal network by providing an effective integration time significantly longer than T'.

An example of the above is the elementary R-C low-pass filter. The impulse response of this network is

$$h(t) = \frac{1}{\tau} e^{-t/\tau} \qquad (7)$$

where $\tau \equiv RC$ is the effective integration time. The experiments reported here utilized an HP 425 d-c meter for the readout device. The input low-pass filter of this instrument has a frequency-amplitude characteristic similar to that of an R-C circuit with an effective integration time of 5 seconds.

Conclusions

The application of galvanomagnetic effects in semiconductors to the design of true rms measurement techniques has been described. The dynamic range of the squaring device depends upon the sensitivity of the readout instrument and with proper design can exceed that of any existing rms instrument. The squarer units reported are very small and simple in construction, and are capable of wideband operation from dc up to at least 300 kc.

Acknowledgement

The authors acknowledge the support of the Bureau of Ships, Department of the Navy. In particular, the encouragement of Mr. G. M. Milligan of the Underwater Sound Laboratory, New London, Connecticut, is greatly appreciated.

References

1. M. Epstein, L. J. Greenstein, and H. Sachs, "Principles and Applications of Hall-Effect Devices," Proc. of the N. E. C., Vol. XV, p. 241-252 (1959).

2. H. Weiss, and H. Welker, "Zur transversalen magnetischen Widerstandsanderung von InSb," Zeitschrift fur Physik, Vol. 138, p. 322-329 (1954).

3. A. H. Wilson, "The Theory of Metals," Cambridge University Press, 2nd ed., p. 235-242 (1954).

4. O. M. Corbino, "Azioni Elettromagnetiche dovute agli Ioni dei Metalli Deviati dalla Traiettoria Normale per Effeto di un Campo. Nuov. Cim. (6), 1, p. 397, 1911; Phys. Zeit. 12, p. 561, 1911.

5. M. Epstein, J. N. Van Scoyoc, and L. J. Greenstein, "Magnetoresistive Magnetic-Field Sensor," Proc. of the N. E. C., Vol. XVII, p. 611-616 (1961).

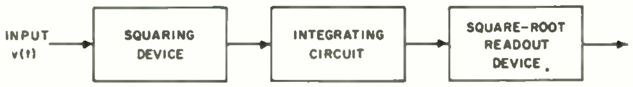


FIG. I BLOCK DIAGRAM OF TRUE RMS METER

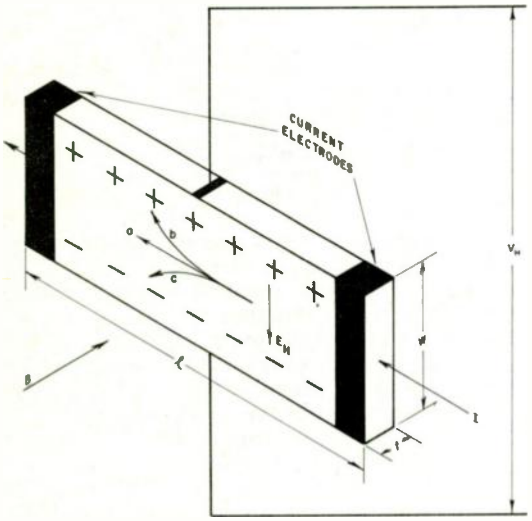


FIG. 2 HALL EFFECT AND PHYSICAL MAGNETORESISTANCE

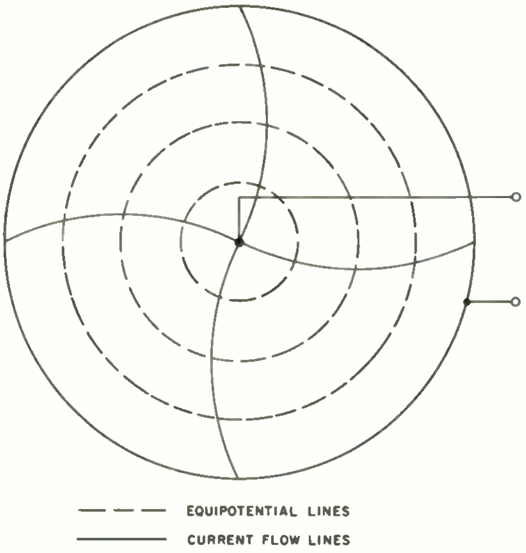


FIG. 3 EQUIPOTENTIAL AND CURRENT-FLOW LINES IN A CORBINO DISK

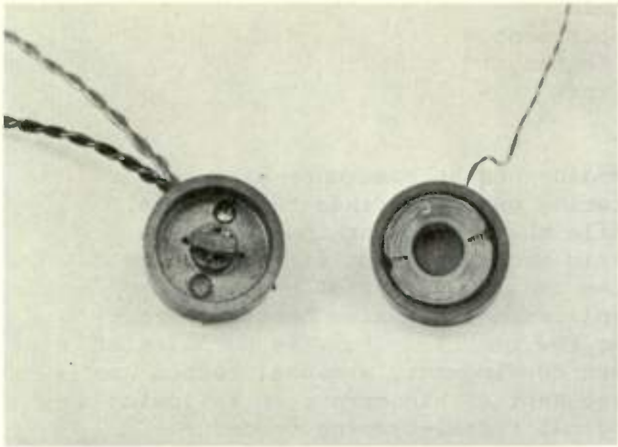


Fig. 4. View of Hall element and field producing coil in ferrite cup-core.

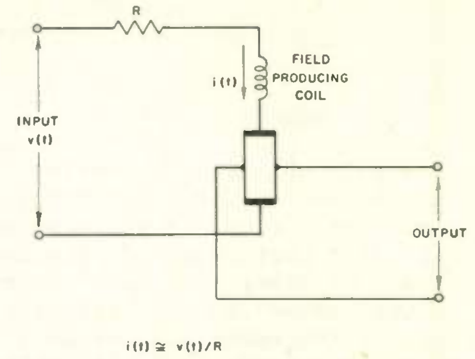


FIG. 6 SCHEMATIC DIAGRAM OF HALL EFFECT SQUARING CIRCUIT

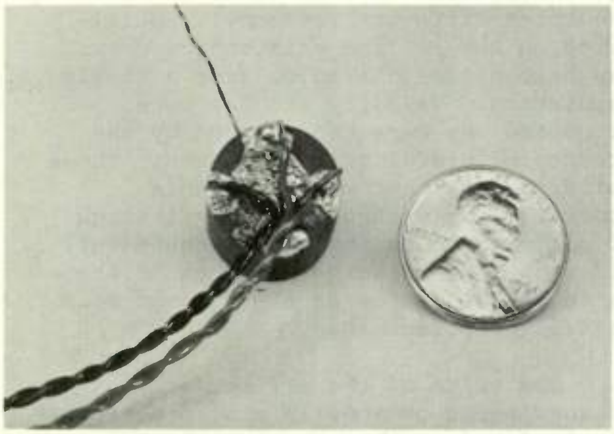


Fig. 5. Assembled Hall effect squarer.

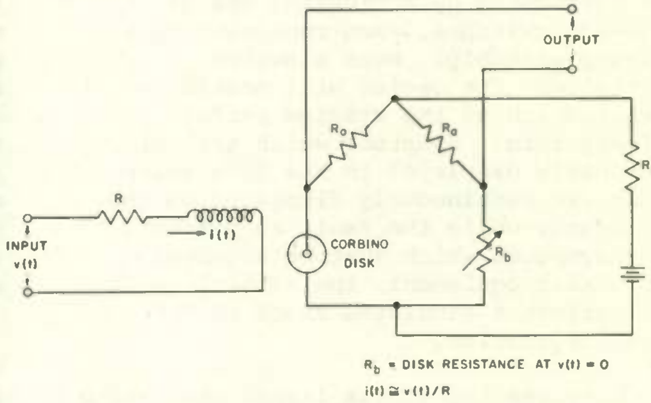


FIG. 7 SCHEMATIC DIAGRAM OF MAGNETORESISTANCE SQUARING CIRCUIT

DEVELOPMENT AND EVALUATION OF A SIMULATOR FOR EVALUATING TROUBLESHOOTING PERFORMANCE ON A COMPLEX ELECTRONICS SYSTEM

A. J. Bernstein
Defense Systems Department
General Electric Company
Syracuse, New York

Summary

As a result of the high availability requirement on the Atlas Radio-Command Guidance System, it is important to maintain peak proficiency of the in-line maintenance crews. However, since faults in the system occur at relatively low frequency, the crews are not able to maintain their troubleshooting skill through the natural conduct of their work. While in some systems, maintenance crew proficiency has been maintained by "bugging" the line equipment, such a procedure on the Atlas system would have very deleterious effects on the weapon's availability. As a consequence, a simulator was developed to enable periodic exercising and evaluation of personnel in system troubleshooting.

This simulator, called the Data Flow Evaluator Set, replicates the system logic in binary form. A "trouble" is introduced by activating one of 250-odd switches, each representing a faulty assembly. When a switch is activated, the device will provide a replication of the symptom pattern in binary form. Symptoms which are automatically displayed in the line equipment are continuously displayed on the simulator while the fault is present. For symptoms which must be detected with test equipment, the subject needs to perform a simulated check to get the information.

By reading status lights and making checks the subject can localize the trouble to a removable assembly. In order to remove an assembly, he pushes the button which represents that assembly on the face of the simulator. If he has removed the proper assembly, the trouble will be automatically cancelled. If he has not removed the proper assembly, the symptom pattern remains and he continues signal-tracing until he finds the trouble. While the man is working, an automatic record is kept of the elapsed time and the number of checks and replacements he makes before correcting the trouble. For the purposes of test development, a manual record was also kept of his errors in following a logical signal-tracing procedure.

In order to test the capability of the simulator, a study was performed on two samples of Air Force personnel. As a result of this research it was learned that the time and check scores are reliable and independent parameters, while the replacements scores and the error scores are too unreliable to be used. Moreover, it was learned that scores on three different tests representing three subsystems of the major system are independent of each other and that the three different personnel specialities on the in-line maintenance crews may be considered samples from a single population. Validity studies were attempted but were handicapped by the absence of useful criterion data. Those validity data that were available (correlations with observer's rankings of performance on the actual equipment) did show some positive evidence of the value of the device as a predictor of actual job performance.

The value of the device as a trainer seems obvious in that practice in simulated troubleshooting should generalize to the job. Moreover, it is possible to use the device as an aid in lecturing, since a plastic overlay, showing the data flow lines, is provided with each machine and the machine has the capability of displaying all symptoms at once whether they appear at displays or check points on the line equipment. The use of the device as a

trainer has not been systematically tested, however, and therefore further research on training applications is recommended.

Introduction

Background

About mid-way in the development of the Atlas ICBM weapon system it became apparent to operational planners that steps should be taken to assure a high state of readiness of the personnel responsible for the in-line operation and maintenance of the system. Such a need seemed particularly great because of the weapon's employment characteristic which would require rapid reaction time of the largest possible force of weapons after extensive waiting periods. As a consequence of this view, plans were formulated to develop a program for on-the-job proficiency evaluation of weapons crews in order to identify field training needs and to exercise the crews in their operational and maintenance tasks.

This program of testing and exercising the crews was finally made a development responsibility of the system contractors and was named the Unit Proficiency System (UPS). As an operating ground-rule, the UPS program was to employ the prime equipment as much as possible but was restricted to activities that would not cause serious degradation of equipment life or reaction time.

In the case of the Radio Tracking System, operational tasks and simple maintenance tasks, like checkout and inspection, could typically be handled within these constraints. However, for system troubleshooting tasks, where faults would have to be introduced into the line equipment, these ground-rules could not be satisfied. As a result, a means for evaluating troubleshooting proficiency on the Radio Tracking System without the use of the line equipment, but nonetheless predictive of line performance, was required. This requirement eventually resulted in the development of a Radio Tracking System simulator set which became known as the Data Flow Evaluator Set. A description of these devices will follow, but first it is necessary to give a brief description of the configuration and maintenance philosophy of the Atlas Radio Tracking System in order to describe the tasks that were to be simulated.

The Atlas Radio Tracking System

The Radio Tracking System is a ground-based electronic system which together with a special-purpose digital computer makes up the Atlas Radio-Command Guidance System. It is the function of the Radio Tracking System to collect data on the missile flight path and feed it to the computer for comparison with a programmed trajectory and then to transmit correction commands (determined by the computer) to the missile's autopilot.

Hardware. The system consists of several antennas, a large number of cabinets containing electronic components, and a console at which system operation (countdown activities) and status monitoring takes place. The console consists of a two-part maintenance wing containing meters, scopes and status lights, which display the static condition of the equipment, and an operations wing which summarizes the static maintenance displays and contains displays and controls for dynamic checks.

The electronic components within the cabinets are, for the most part, packaged in hinged-panel assemblies with quick-disconnect mechanical and electrical connections. The panels contain primary and secondary test points, and the outside of each cabinet contains status lights and meters for most of the functions within the cabinet.

Personnel. The crew assigned to in-line operation and maintenance of the system consists of a Guidance Control Officer (usually a Captain or Major), a Radio Tracking System Technician (usually a Master or Staff Sargent) and a Radio Tracking System Specialist (usually an Airman First or Second Class), all of whom have had some form of maintenance training on the system. Operation of the system consists of the initiation and evaluation of an automatic check during count-down and the monitoring of system performance during the early part of the powered flight of the missile. Maintenance of the system at the line site consists of the performance of periodic checks and inspections and the localization and correction of faults. System malfunctions are usually detected on status displays on the console which also localize the trouble to a subsystem, group or cabinet. Localization to a particular assembly is accomplished by further analysis of the symptom pattern as shown by the cabinet-mounted confidence lights and by interrogation of the test points on the panels. When the faulty panel is identified it is removed as a whole and replaced from an in-line supply of panels. The faulty panel is then sent to a shop for further troubleshooting and repair. In the performance of these in-line maintenance duties, the crew operates as a team or independently - at the option of the Guidance Control Officer.

Need for a Simulator

In order to maintain the high level of operational readiness or availability which is required of the Atlas Radio Tracking System, it was decided that special emphasis should be placed on exercising and evaluating the crews in the localization of faults which was potentially the largest consumer of down-time. It was further decided to go to a special device for the following reasons:

(1) evaluation could be accomplished without occupying and possibly degrading the operational equipment;
(2) evaluation on specific problems would not be dependent on their occurrence in the system;
(3) a large sample of problems could be administered within a relatively short time;
(4) the evaluation could be administered by non-specialized personnel;
(5) results could be automatically scored, resulting in high reliability of scores.

A major assumption underlying the use of such a simulator, of course, is that performance on the simulator is generalizable to performance on the job. For this reason it was decided to attempt to replicate the logic of the line system as faithfully as possible but to reduce the complexity of the device by treating the system logic in binary (go - no go) form. To a large extent, this approach was inspired by the MAC Trainers developed by the Air Force Personnel and Training Research Center for the B-47 Bomb-Nav System.

The Data Flow Evaluator Set

The resultant simulator became known as the Data Flow Evaluator Set or DFE. This device is actually three physically separate units corresponding to a division of the line system into the rate measuring function, the track data processing function and the track rf and servo function. The three units are called the Rate Simulator, the Track Data Processing Simulator, and the Track RF/Servo Simulator which will be referred to as Rate, TDP and TRF respectively in this report.

The front panel of each unit of the Data Flow Evaluator set symbolically shows the cabinets and panels of the function it represents. In its dynamic operation, it treats system data in binary form. Displayed data are not presented in terms of values on a continuous scale, but in terms of whether the voltage, waveform, etc. is in or out of tolerance at the particular point being monitored. The major functions of the system are made "operative" so that most of the critical measurements associated with trouble localization in the actual equipment may be performed in a simulated way. Figure 1 shows the Rate simulator of the Data Flow Evaluator Set.

Each simulator unit is designed to provide for the quick insertion of a large number (approximately eighty per unit) of realistic malfunctions. Malfunctions are inserted at the tester's panel by pushing a button representing a particular fault. Pushing the button causes the relays within the cabinet to make or break contact and establish the symptom pattern (in binary form) on the front panel. Data displayed on the line equipment are also displayed on the simulator, but in the case of test points, the subject must address them to determine if the indication is go or no-go. Counters on the tester's panel automatically record the number of checks he makes, the number of corrective actions (replacements) he takes, and the total time elapsed in localizing the fault. Figure 2 shows the instructor's panel of the Rate Simulator.

The components of the simulators are of standard design throughout. The relays are hermetically sealed and are the plug-in type. The push-button switches are all of the same specific design and are illuminated. Indicator lights are marked to show what they represent (e.g. meter indications are symbolically marked with the face of a meter). The entire unit may be illuminated to check for burned-out lamps. A relay tester is built into the back of the evaluator to check relay operation and contact continuity. Electrical power is obtained from a standard 110 volt, 60 cycle wall outlet. The units are mounted on retractable casters for mobility.

The face of each unit displays the remove-and-replace panels ("black boxes"), meters and confidence lights of each relevant cabinet as well as the console displays. After a malfunction is set into a unit by the tester, a "Start" switch is pushed, illuminating the console section of the unit and simultaneously activating the three counters.

The console section of the front panel reveals the system status summary and the results of the dynamic system test. The console-mounted scope displays and meter indications are also simulated to give go - no go indications for the functions they represent. From these data, the subject must determine which cabinet or group of cabinets the fault is in. After determining the faulty cabinet or group, the subject must press the switch representing the cabinet he wants to address. As each cabinet is addressed, additional indicators representing meters and confidence lights on the cabinet are displayed to him replicating the symptom pattern. Symbolic test points as found within the cabinets are also available. To check a test point with external test equipment, a selector switch must be set to the test point and a push-button representing the appropriate test equipment (meter or scope) must be depressed. If the push-button lights, then the signal at the test point is good; otherwise it is bad. By following the data flow within the cabinet (checking test points, etc.) the subject localizes the malfunction to the faulty panel. To "replace" a panel, he must push its "Replace" switch which is appropriately labeled on the DFE unit. If he is correct the Data Flow Evaluator automatically removes the problem and stops the counters. The tester would then record the results, reset the counters to zero and go on to the next trouble.

Although present use of the simulator does not include the insertion of more than one malfunction at a time, any combination of the malfunction options in each unit may be set into a unit at the same time. Malfunctions generally require 15 seconds to 3 minutes for solution, depending upon the complexity of the malfunction, and the skill of the subject.

The Data Flow Evaluator Set also has applications for training. For example, the entire panel of a unit may be illuminated and a malfunction switch pushed. With this procedure, the resulting overall indications (or profile) for a given malfunction may be observed. This training procedure may be enhanced by placing a transparent overlay with data flow lines (which is supplied with the unit) on the front panel of the unit to actually show the relationships among the cabinets and panels. This application, however, has not been exploited and should be subjected to further field research to assure that negative transfer or an inhibition of performance on the actual equipment is not taking place.

Field Evaluation

Pilot Study

Upon completion of the first prototype sets of the simulators, a pilot study was conducted at Keesler Air Force Base on 59 trainees to determine the scoring parameters that could be used with the simulators. The four scoring parameters studied were: (1) the time taken to correct a trouble (time score), (2) the total number of checks made in correcting a trouble (checks score), (3) the number of unsuccessful replacements made before making the correct replacement (replacements score), and (4) the number of erroneous checks made in localizing the trouble (error score). In the first three cases, the parameters were automatically recorded, but in the last, the subjects' checks were compared with an "ideal" set of checks as formulated by experienced technicians and his score was the number of "extra" or "irrelevant" checks made. The specific research questions asked were:

(1) Is there a proper spread of scores for each parameter?
(2) How reliable are the scores?
(3) Are the different parameters within a unit independent?
(4) Are the different units independent within each parameter?
(5) Do the three personnel specialities represent different norms groups?

The results of this study, which use a random sample of 50 items from each DFE unit, revealed that there was a satisfactory spread of test scores on all parameters. The ranges of scores are as follows:

Rate Simulator

Time (min.)	23 to 105
Checks	97 to 236
Replacements	0 to 39
Errors	42 to 187

TDP Simulator

Time (min.)	16 to 98
Checks	93 to 220
Replacements	4 to 27
Errors	61 to 169

TRF Simulator

Time (min.)	28 to 109
Checks	106 to 275
Replacements	4 to 41
Errors	18 to 123

All distributions were skewed to the positive (poor performance) end with the mode lying at a point on the range that is about 20 percent above the lower end.

The assessment of parameter reliability was accomplished by computing split-half correlations for each of the four scoring factors on each of the three Data Flow Evaluator units. The means of the reliability coefficients (computed by Z transformations) for each of the scoring factors (50-item length) were found to be as follows: Time $\bar{r}$ = .94, Checks $\bar{r}$ = .89, Replacements $\bar{r}$ = .66, and Errors $\bar{r}$ = .74. On the basis of these findings, the Replacement and Error scores were deemed to be of little use in testing and consequently they were eliminated from further analysis.

The independence of the scoring parameters and subtests (or units) was tested by computing the intercorrelations among the six remaining variables. It was found that the mean corrected correlation between time and checks scores across the three units was .41. The corrected correlations among the three subtests (or units) for each parameter were as follows:

Time Scores

Rate vs. TDP	.43
Rate vs. TRF	.58
TDP vs. TRF	.48

Check Scores

Rate vs. TDP	.49
Rate vs. TRF	.38
TDP vs. TRF	.60

From this analysis it was concluded that the scoring parameters and subtests should be considered independent measures and scored separately and that the profile of scores on subtests (or units) could be considered somewhat diagnostic.

In order to determine whether the three different personnel specialties (Guidance Control Officer, Radio Tracking System Technician, Radio Tracking System Specialist) constituted different populations for norms purposes, the scores of the three groups were compared on the six scoring factors. The null hypothesis was tested using a non-parametric median test with a confidence level of .05. In no case could the null hypothesis be rejected, and it was therefore concluded that all specialists should be considered as belonging to a single norms group for proficiency testing.

Final Study

As a result of the apparent success of the pilot study, it was decided to continue work on the tests to see if their predictive validity could be assessed and to shorten administration time and provide alternate forms.

Validity. The validity study was performed on all five operating crews (fifteen men) that had had any experience with the system. In the traditionally difficult search for criteria, it was finally decided to use rankings of the men's field performance as provided by six contractor field service personnel who had been working with them during the initial installation and checkout of the equipment. Since it was impossible to take the time to develop more sophisticated metrics, the judges were simply asked to rank the men on the following three variables:

1. Ability to troubleshoot the Rate Measuring Subsystem.
2. Ability to troubleshoot the Data Processing loop of the Tracking Subsystem.
3. Ability to troubleshoot the RF and Servo loop of the Tracking Subsystem.

Several additional variables were also included to provide the judges with some practice in ranking the men. Rankings were accomplished by giving each judge a deck of fifteen cards (one for each man) set up in random order and asking him to stack them from best to worst. The order in which the variables were presented to the six judges was varied in each of the six possible ways.

The reliability of the rankings was assesed by measuring the agreement among the judges. A coefficient of concordance for small samples (W') was computed for each variable and transformed to an average rank-order correlation. These $\bar{R}$ figures were found to be .52, .70 and .77 respectively, too low to allow much confidence in their usefulness as criterion measures. Furthermore, there appeared to be a substantial "halo" affect in the rankings since the intercorrelations among them were as high or higher than the reliability measurements, and their mean correlation with rankings on "overall knowledge of his responsibilities", a purposely ambiguous variable was found to be .80.

Test validity was estimated by computing the correlation between each scoring parameter, as measured on a 30-item sample of the three tests, and the judges' rankings on the relevant criterion variable. These correlations corrected for attenuation were as follows:

Rate time	=	.35
Rate checks	=	.75
TDP time	=	.49
TDP checks	=	.24
TRF time	=	.21
TRF checks	=	.21

The multiple correlation (using time and checks scores) for each criterion variable was found to be .86, .53, and .28 respectively.

Reliability and Test Length. To assess and improve the reliability of the test scales, the whole tests were administered to a sample of 44 men undergoing crew training at Vandenberg AFB. Item - total correlations were computed for all the items and the reliability of the six test variables were estimated from these data. These correlations ranged from .88 to .93.

Since the mean test time for each of the three units ranged from one and a half to two hours, it was considered very desirable to reduce the overall time of the tests. To do this, the item analysis data were reviewed and all items with an item validity index below .10 on both scoring parameters were eliminated. The net effect of this procedure was to lower the test time by an average of fifteen minutes and raise the estimated reliability by an average of .04. To further reduce test time, each test was split into two parts by assigning alternate items to either form. This procedure provided two equivalent forms for each test, each of which would take an average inexperienced technician about forty minutes to complete. The reliability of these alternate forms is estimated to be something over .85.

Conclusion

Although the validity data do not fully support the contention that the Data Flow Evaluator is highly predictive of job performance, it seems obvious that the empirical findings are of limited value because of the unreliability of the criterion measures and the smallness of the sample (15 cases). The fact that sample size is a major factor in these findings seems apparent in that the corrected validity correlations ranged from .21 to .75 on a battery of tests and criteria that are very similar in nature. Furthermore the high reliability of the tests and their high face validity suggest that they should be very useful in fulfilling their original purpose of diagnostic measurement.

During the development and evaluation of the devices, it was often suggested that although the units were designed for evaluation, they seemed to have very high promise as training devices and in fact the prototype models at Keesler Air Force Base were later used for that purpose. The fact that the devices can be used with groups of about 15 men to show the system data flow and how a large variety of troubles affect this data flow, has a lot of intuitive appeal. Moreover, the men thoroughly enjoyed working the devices and very often returned after duty hours to have their instructors or buddies insert troubles for them to find. While a claim of the training value of the machines would have to be verified by experimental findings to assure that in some way the simulation is not introducing proactive interference and thereby inhibiting the learning of actual job skills, it does seem that the devices wouldbe very useful in training.

In view of the fact that the devices do provide reliable measurement, are inexpensive to build and operate, are easy to use, do not interfere with in-line equipment, and are very appealing to the men, it is felt that they are a sound investment.

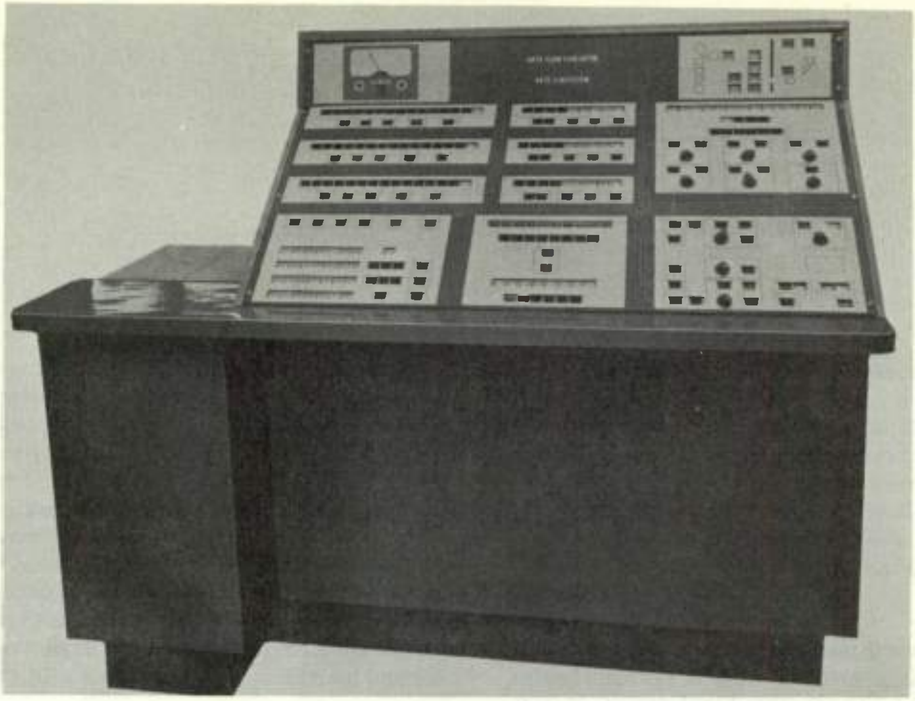


Fig. 1.

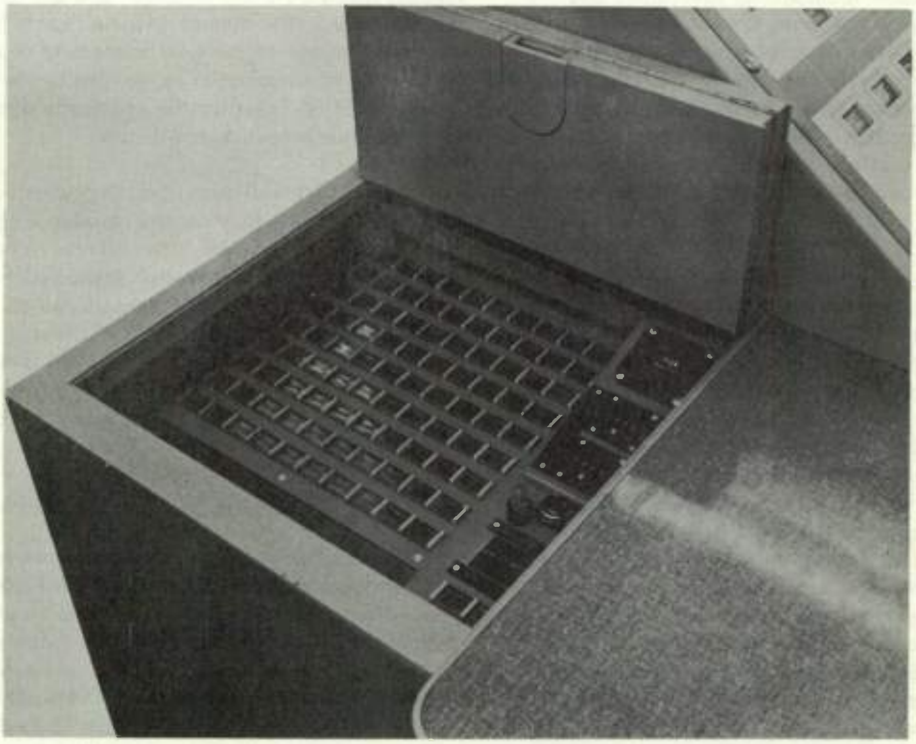


Fig. 2.

PERSONNEL SUBSYSTEM DEVELOPMENT AND INERTIAL GUIDANCE SYSTEMS

Gerald J. Ferwerda
AC Spark Plug Division
General Motors Corporation
Milwaukee, Wisconsin

Summary

This paper describes the role of personnel subsystem development during research and development at AC Spark Plug Division of General Motors Corporation, Milwaukee. The products and processes described provide the human performance required to assure effective accomplishment of the assigned function and to assure that the reliability built into the system is not degraded to the extent that the system looses its usefulness.

The development of qualified personnel to operate and maintain complex man-machine weapon systems begins with the engineering data generated in design trade-off studies and system functional analyses. These studies and analyses are made by an integrated group of engineering and human factors personnel whose main objective is to develop a design which effectively utilizes human capabilities in combination with available or projected equipment techniques, functions, and capabilities to achieve optimum system performance. These efforts also result in a delineation of the functions which can be most reliably and economically performed by humans and those which can be more appropriately mechanized or automated. The human functions are then analyzed to derive the specific duties and tasks which must be performed to operate, maintain, and control the weapon system.

The set of "basic data" resulting from the system analyses forms the basis for development of subsequent "personnel products." The kinds (types) and numbers of military personnel required to operate, maintain, and control the system are defined via a Qualitative Personnel Requirements Information (QPRI) program. Technical instructional courses and aids needed to qualify the military personnel to perform all elements of their jobs in a realistic situation are designed via a Training and Training Equipment program. Publications and job guides containing descriptions of and instructions for the effective use and maintenance of equipment are generated via a Technical Manual program. A man-machine system is designed via the Human Engineering program, and the requirements for the maintenance subsystem are delineated via a Maintenance Analysis program. The total personnel subsystem is then tested and evaluated to determine whether or not the criteria used for the development of the personnel subsystem products have been met.

Introduction

The development of qualified personnel to operate and maintain complex man-machine weapon systems is an integral part of the system development sequence. Recognition of this concept by such military agencies as the Air Force Ballistics Systems Division has resulted in various requirements for weapon system contractors to provide personnel subsystem development efforts and products.[1] Included within the term "Personnel Subsystem" are the processes and products essential for personnel and hardware integration; personnel identification, development, and employment in the system; and maintenance of personnel effectiveness.

The development of personnel subsystem closely parallels that of its system counterpart, the hardware subsystem. Both subsystems have specified performance required of them and both contribute to total system reliability and effectiveness. In addition, both have measurable inputs and outputs, and are needful of maintenance. The degree to which each "receives" from or "contributes" to the total system is dependent upon the degree of automation expressed by the system design, that is, whether the system is a machine-favored or a man machine-favored design.

This paper will describe the personnel subsystems developed for military inertial guidance systems by AC Spark Plug Division. The efforts to be described are consistent with general philosophies set forth by the military in the form of specifications, standards, and exhibits. Although certain of these documents and philosophies are specific to a particular weapon system, the concept is applicable to many types of systems, that is, weapon systems, computer systems, radar systems, and so forth.

Personnel Subsystem Basic Data

Because most of the personnel subsystem processes and products, as well as much of the logistical and operational planning data, are based on essentially the same fundamental information, the personnel subsystem should be developed based on a single source of closely integrated technical data. In the past, this "fundamental information" took the form of functions and task analyses: functions analyses, which delineate system functions and allocate these functions, on the basis of the capabilities and limitations of men and machines, to men and/or machines; and, task analyses, which present detailed information about the tasks assigned to men based on analyses of the functions allocated to men and men-machine combinations.

Basic data for Titan II personnel subsystem development, for instance, took a more sophisticated approach than the functions and task analyses programs used on systems of the past. A specification which defined the data to be furnished by all Titan II contractors to serve as the basis for the development of the Titan II personnel subsystem was placed on contract for all Titan II contractors.[2] The data called for by the exhibit were of the following types:

Type I. System Functional Flow Diagrams

> The flow diagrams identify and show the sequence of all system functions and activities programmed for the operational system.

Type II. Operation/Maintenance Activities Analysis

> This analysis translates system functions in terms of equipment and personnel required in the performance of each system activity identified in Type I data.

Type III. Performance Standards Analysis

> These data are a further reduction and detailed specification at the input-output level of the personnel performance described in the Type II analysis.

The Type I functional flow diagrams graphically describe the sequence of system functions and activities. The diagrams reflect approved weapon system operational and maintenance concepts and plans and are consistent with weapon system and subsystem criteria and requirements which establish the need for any particular function or activity and with the engineering data which define equipment functions. This family of diagrams establishes a hierarchy of system functions and activities which describe those events which are performed by the equipment, by people, and by people and equipment working together. The Type I diagrams use a conventional flow symbology of blocks and circles connected by flow lines which show the relationship both within and between groups of functions and activities. A sample of Type I data is illustrated in Figure 1. It should be noted that a numbering system has been assigned to the units of performance contained on the diagrams which reflects such things as geographical area, groups of functions, subsystem involved, etc.

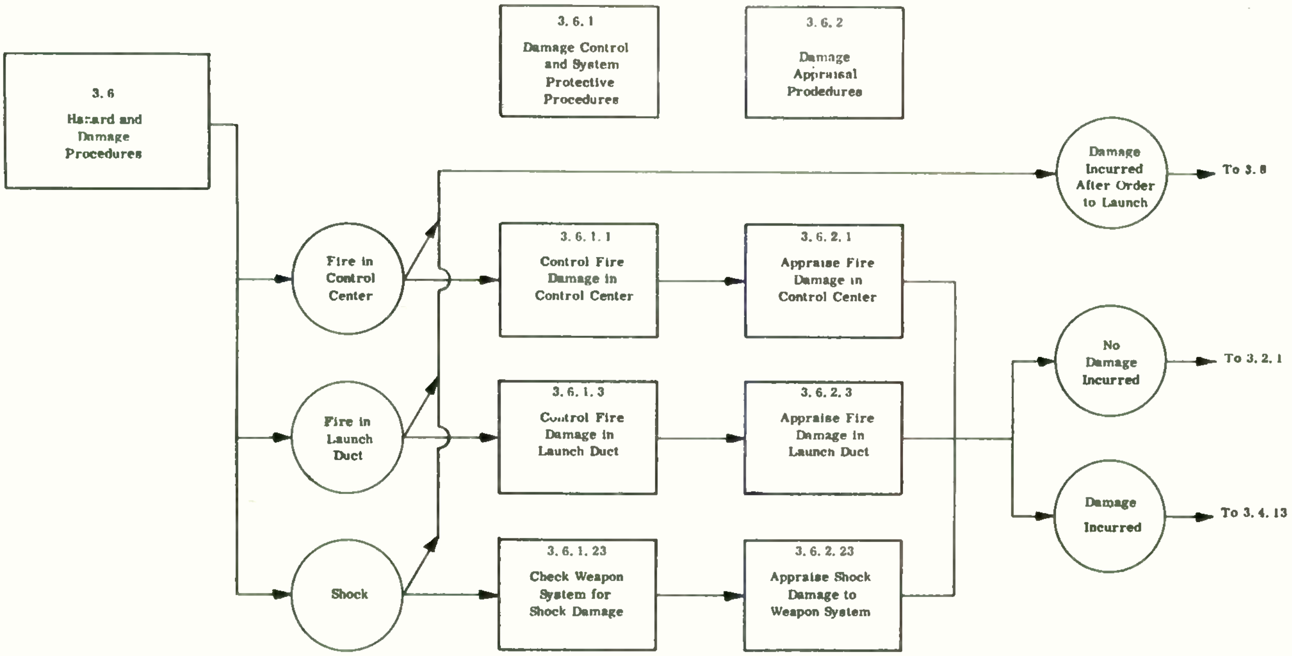

3. 6
Hazard and Damage Procedures
3. 6. 1
Damage Control and System Protective Procedures
3. 6. 2
Damage Appraisal Prodedures
Fire in Control Center
Fire in Launch Duct
Shock
3. 6. 1. 1
Control Fire Damage in Control Center
3. 6. 1. 3
Control Fire Damage in Launch Duct
3. 6. 1. 23
Check Weapon System for Shock Damage
3. 6. 2. 1
Appraise Fire Damage in Control Center
3. 6. 2. 3
Appraise Fire Damage in Launch Duct
3. 6. 2. 23
Appraise Shock Damage to Weapon System
Damage Incurred After Order to Launch
To 3. 8
No Damage Incurred
To 3. 2. 1
Damage Incurred
To 3. 4. 13

Figure 1. Type I Functional Flow Diagram

Type II data provides the first breakdown of system operational and maintenance activities in terms of the equipment involved and the participation of personnel in the performance of these activities. Personnel performance associated with each activity identified on Type I diagrams are portioned into tasks i. e., the next lowest unit of performance, and the initial specification of human performance requirements is accomplished. A tabular format is utilized listing derived equipment and personnel performance characteristics against system function. Typical of data elements so tabularized is the activity description, which lists the input and output states of the activity, the equipment required and used during the performance of the activity, the location where the activity is performed, and the frequency with which the activity is performed; and general personnel characteristics, which describe the personnel tasks to be performed, the time required to perform each task, the perceptual, judgemental, and motor skills demanded during performance of the tasks, the difficulty, criticality, and hazards of the task, crew and communications requirements, and the training requirements for the task. Criteria for most of these data elements is either provided by the customer or developed in an integrated fashion by all contractors so that the personnel subsystem for the total weapon system uses a standard base.

The Performance Standards Analysis Type III data are a further reduction of the personnel performance data generated on Type II data sheets. In general, Type III data are developed through an analysis of each task which is directed toward a detailed specification of the personnel actions associated with task performance, and performance standards. Figuratively speaking, the Type III data describes "what" has to be done to accomplish a given activity, and Type III data defines "how" to do it. A tabular format is also utilized for Type III data. For each task listed in the Type II data, the task elements, i. e., the next lowest unit of performance, comprising the task are listed along with their corresponding inputs or sources, e. g., equipment cues, verbal cues, etc. These task elements are then further analyzed for their performance standards which are also listed on the Type III format. Typical of performance standards data elements are minimum performances required, measurement techniques for evaluating performance, likely deviations from performance standards, and consequences of deviations from performance standards.

This basic data "package" is developed progressively in consonance with the development of the weapon system equipment and engineering data and with the need for data elements of the personnel subsystem. These needs will be described in subsequent sections of this paper. Before leaving this section, however, it should be noted that initial documentation of basic data should be made in the early stages of system development in order to be timely for personnel subsystem products, and this initial documentation must be updated and refined as the various system and equipment specifications are released and system concepts take shape.

Personnel Requirements

Using the Operation/Maintenance Activities Analysis data (Type II) described previously, tentative groupings of tasks into positions (all tasks assigned to one man) are made. This grouping is based on consideration of task performance as being selectable, trainable, and/or guidable when the individual tasks are combined with other tasks to form a position. The results are formally presented in a Qualitative Personnel Requirements Information (QPRI) report consistent with format requirements of military specifications. [3, 4]

The purpose of the QPRI (or QQPRI) program is to develop the requirements for personnel to support system operation and maintenance. Specifically, such bits of information as position definition, personnel qualifications, and manning estimates are derived as a result of the QPRI effort. The comprehensive QPRI program should assure a personnel package which will complement the hardware components of the system. It specifies the kinds of persons who are needed to provide total system performance, the specific performances each person will contribute, the qualifications required in these individuals, and how many individuals of each type are required by the system.

The Type II data will provide a detailed description of all personnel performance required for system operation and maintenance without regard to who must perform them. Performance of a human in a man-machine system is a function of his performance capability with respect to the skill, knowledge, and procedural requirements of the task at hand. The tasks are "clustered" into position on the basis of equipment and facility considerations, functions, sequence of operations, homogeneity of qualifications, and similarity of circuitry.

Each personnel position derived is examined to determine its relationship to the military speciality codes currently in existence. When an existing Air Force speciality matches the qualifications for a given position, it can be utilized simply by adding specific system training.

Upon receipt of personnel requirements information, the military agency responsible for providing the personnel to operate and maintain the weapon system will usually analyze the information to determine whether the specialities and the manning estimates comprising the QPRI are within the current inventory of available military personnel. If such is not the case, the agency will match the specialities requested as closely as possible with available specialities and call for the training program to make up the deficit between the two.

Personnel Training

Having described how a portion of the desired performance potential is obtained through selection (via the QPRI), we will now consider another means to this end — training. Training is the primary means through which the performance potential is provided in an individual's repertoire when selection cannot provide it.[5] Training is necessary and most useful when the individual must learn such things as complicated motor and perceptual skills, and principles for decision making and system functioning. Those elements of performance to be provided by training will be subject to a training analysis which will result in a development of training standards and curriculum development. Curriculum, as used here, is the structure based on the training program which specifies subject matter to be learned, time estimates for each subject, training aids and materials necessary, and techniques for presenting information. A necessary adjunct to the curriculum is a system for curriculum and trainee evaluation.

Using data elements contained in the basic data package as a base, a training analysis is conducted to determine the requirements for training, i.e., all the tasks for each position which must be obtained through training. A major step in the training analysis is the review and subdivision (if necessary) of each required performance capability into elements that can be taught and evaluated by some known and feasible means. From this point the analysis will determine the teaching methods and training materials required to instill such capabilities. Recommended course curricula will then be developed on the basis of the teaching method and training materials specifications.

The development of training devices is in itself such a complex and special problem that it is usually treated apart from the curriculum development effort. The concept of the training device must be developed by considering a number of factors, some of the most salient of which are time requirements, cost, state-of-the-art, training principles, and treatment and order of similar materials. Training devices and equipment in support of individual integrated weapon system, and unit training programs on the Titan II system were developed consistent with ground rules laid-down on all Titan II contractors by the Air Force via specification.[6] In summary, it can be stated that training equipment must be designed to provide or invoke experiences within the individual which enable him to vary his responses in the direction of an indicated standard.

Personnel Job Aids (Manuals)

Job aids, or manuals, have an important role in providing performance development in personnel. They are employed as stimulus materials to foster reliable personnel performance. Job aids are used for fostering personnel performance for several reasons. They may be used to reduce stringent selection criteria, or to reduce requirements for training by substituting the job manuals method for the training method. Another reason for employing job manuals is to foster the reliability of personnel performance on the job. In addition, they often act as a "back-up means" for personnel performance.

Job manuals typically contain two different types of information. The first of these, performance guidance information, is that information designed to foster the desired performance without changing the personnel requirements. Thus, a trouble-shooting chart is a performance guidance means because the performance of the technician using the chart is essentially the same as the performance of the trouble-shooter who makes his own analyses. Performance guidance materials in job manuals may also be used as a means for providing on-the-job training for installed systems where modification or retrofits require the acquisition of new skills, knowledge, and procedures.

The second type of information characteristically in job manuals is reference materials, used to provide personnel with an "extension of their memories" where the information series are too long for memorization or too complex for recall. This type of information is typically a fixed procedure and knowledge elements which are primarily factual in nature, and usually presented in a form such as schematics, calibration charts, and parts lists.

For the most part, the information contained in Type III basic data fulfills the needs of job manuals. All that is usually necessary is to transcribe the Type III procedural data to the format(s) prescribed by the military in various specifications [7, 8, 9] and add such other job aids as schematics, charts, equipment descriptions, and so forth.

Human Engineering

Human engineering, as it is known in the profession today, is that aspect of personnel subsystem development concerned with fitting the machine to the man. The successful design of equipment for human use requires consideration of the following basic human characteristics: sensory capacities, mobility and muscle strength, intellectual abilities, common skills and capacity for learning new skills, capacity for team or group effort, and body dimensions — and, in addition, of the effects of working environments upon human performance. By considering these characteristics, human engineering ultimately reduces the problem of selection and training requirements, discussed earlier, and thus provides for more effective utilization of personnel in the system.

It is not the purpose of this section to delineate the many principles involved in the human engineering of inertial guidance systems. The reader is referred to a paper on this topic delivered by the writer during the 1962 IRE Winter Convention on Military Electronics to be published in the IRE Transactions of this convention.[10] The topic is introduced here so as not to destroy the continuity established earlier of describing personnel subsystem "products."

Basic data, developed during the early stages of system development, is useful to the human engineer in assigning equipment means to man-machine functions, within constraints established by the military in human engineering criteria specifications.[11] Human engineering can be started immediately after the system functions have been identified and allocated to man and/or machine. The requirement for human engineering should be evident for most performances designated to be accomplished by man-machine combinations. For example, if the function of aligning an inertial guidance system to azimuth is shared by man and device, the knobs and dials on the device must be designed for fine adjustment to facilitate the accuracy with which such azimuth alignment must be made. Many such requirements can be identified after the functions have been allocated to man-machine combinations. More detailed requirements will come after functions have been analyzed for tasks and task elements. For example, such analyses may reveal that a meter on the equipment will not provide information sufficiently accurate for the man's task, e.g., the units on the meter may be larger than the adjustment required of the man.

Maintenance Analysis

Using the system functional flow diagrams (Type I data) developed as part of the Basic Data program described earlier, a maintenance analysis program is inaugurated in order to derive the functional requirements for ground support equipment. From this point on, the personnel subsystem Basic Data program and the Maintenance Analysis program complement and support each other.

A typical maintenance analysis program, in addition to presenting the requirements for ground support equipment, delineates inspection requirements, maintenance allocation requirements, spare parts requirements, replacement intervals, and time significant item requirements, to mention a few. This process begins with an analysis of each operational ground equipment end item to functionally define the maintenance requirements for these equipment units. Further analysis will result in recommendations for maintenance ground equipment required to inspect, test, service, adjust, calibrate, appraise, gauge, measure, repair, overhaul, assemble, disassemble, handle, transport, safeguard, record, store, actuate, or otherwise maintain the intended functional operating status of the operational equipment.

Personnel Subsystem Test and Evaluation

In order to determine whether or not the criteria used for the development of personnel subsystem products have been met, a Personnel Subsystem Test and Evaluation program must be designed and implemented. Under the guidance of the military,[12] such a program is currently in effect at most of the facilities of Titan II contractors and at Vandenberg Air Force Base in California. It is comprised of a human engineering evaluation of each end item of non-standard equipment used in the weapon system, and a personnel activities evaluation of each applicable personnel activity listed in the Performance Standards Analysis Data (Type III data) as defined by the Basic Data program.

Typical of items to be evaluated are the following:

a. Adequacy of design and procedures in relation to operability and maintainability.

b. Adequacy of trade-off points for automated versus manual operations.

c. Adequacy of the number of personnel specified as necessary to perform any given activity.

d. Effect of equipment arrangement on crew efficiency.

e. Adequacy of test equipment and special tools authorized for use in the activity.

f. Adequacy of transporting and handling equipment.

g. Communications efficiency.

h. Efficiency of the sequence of procedural steps prescribed for each activity.

Personnel subsystem evaluations are also designed to identify deficiencies in the equipment or procedures which would tend to degrade efficiency and safety during operation and maintenance of the weapon system. Typical of deficiencies to be identified are the following:

a. Design features which lead to inefficient operation and maintenance.

b. Error inducing equipment design features and procedures.

c. Design features or procedures which constitute a hazard to the safety of personnel or equipment.

The evaluation itself usually takes place in an environment that is the same or close to the environment which the ultimate user will experience. Even such factors as temperature, humidity, light, and sound are controlled to approximate those to be experienced by the user if the evaluation is to be representative. It is also extremely important to simulate the "types" of people who will ultimately use the system during evaluations insofar as experience, training, education, anthropometry, motivation, and other characteristics are concerned.

The uncertainties encountered in weapon system development make this type of test program necessary. As such, it becomes the responsibility of the human factors engineer to determine whether operations performed by humans in the system meet performance predictions, and, where they do not, to determine the nature and extent of the deviations in order that design and procedural improvements can be made.

Conclusion

The foregoing sections of this paper have described a development system for "producing" the personnel products and processes for a complex weapon system, and have identified the major units within such a system. It has been shown that personnel subsystem development closely parallels the development of its system counterpart, the hardware subsystem. Both subsystems have specified performances required of them and both contribute to total system reliability and effectiveness. In addition, both have measurable inputs and outputs and both require maintenance.

The products and processes described herein are all necessary to provide the human performance required by the weapon system. Personnel subsystem development uses as a point of entry the engineering data generated in design trade-off studies and system functional analyses. These design studies are normally accomplished by an integrated group of engineering/human factors personnel whose main objective is to develop a design which effectively utilizes human capabilities in combination with available or projected equipment techniques, functions, and capabilities so that optimal system performance is achieved.

These design efforts also result in a delineation of the functions which can be most reliably and economically performed by humans and those which are more appropriately mechanized or automated. The human functions are then analyzed to derive the specific duties and tasks which must be performed in order to operate, maintain, and control the weapon system. The duties and tasks are further studied to determine job structures, performance requirements, and the selection, training, manning, and support or job-aids implications.

The machine functions are also analyzed to determine the scope of the human engineering effort required to assure optimum man-machine interface. In the field of missile systems, human engineering generally concerns itself with the design for operability of ground operating and ground support equipment, and the design for maintainability of the airborne equipment together with its associated ground equipment. The total personnel subsystem is then tested and evaluated to assure that man-machine relationships developed are optimum; to assure that the reliability designed and built into the hardware subsystem is not degraded to the extent that the weapon system loses its usefulness as an instrument of combat.

Bibliography

1. AFBM Exhibit 60-17, "Personnel Subsystem General Requirements Specification for WS-107A-2 Titan II Weapon System," Revision A, 10 October 1960.

2. AFBM Exhibit 60-26, "Basic Data for Titan II Personnel Subsystem Development," Revision A, 10 October 1960.

3. AFBM Exhibit 58-18, "Qualitative Personnel Requirements Information," 12 October 1959.

4. MIL-D-26239A, "Data, Qualitative, and Quantitative Personnel Requirements Information (QQPRI)," 14 April 1961.

5. Miller, R. B., "Human Engineering Design Schedule for Training Equipment," WADC Technical Report 53-138, June 1953.

6. AFBM Exhibit 59-17, "Training Equipment Provisioning for Air Force Ballistic Missiles and Military Space Systems," 1 November 1959.

7. MIL-M-5474C (USAF), "Technical Manuals: General Requirements for Preparation of," 30 April 1960.

8. MIL-M-9864 (USAF), "Technical Manuals: Operation and Organizational Maintenance (Missile Weapon Systems)," 30 June 1959.

9. MIL-C-9883A (USAF), "Check Lists for Missile and Space Systems Operation and Organizational Maintenance," 26 October 1960.

10. Ferwerda, Gerald J., "Human Engineering and Inertial Guidance Systems," presented during 1962 IRE Winter Convention on Military Electronics, 7-9 February 1962, Los Angeles, California.

11. MIL-STD-803 (USAF), "Human Engineering Criteria for Aircraft, Missile, and Space Systems, Ground Support Equipment," 5 November 1959.

12. AFBM Exhibit 60-20, "Personnel Subsystem Test and Evaluation Program," Revision A, 10 October 1960.

AN EXPERIMENTAL STUDY OF MISSILE CONTROL SYSTEM PARAMETERS

J. J. Seider and H. L. Williams
Orlando Aerospace Division
Martin Marietta Corporation
Orlando, Florida

Introduction

Experimental studies reported in this paper were performed as part of a program to establish the design parameters in the control loop of an air to ground missile system. The control loop consisted of a pilot, joystick, command link from the aircraft and the control surface actuator in the missile. One objective of the experimental studies was to determine how long the joystick should be. A second was to determine the feasibility of including a lead or type of quickening in the command link to compensate for "time lag" due to the pilot and the equipment. Still a third was concerned with establishing the transfer function governing degree of sensitivity between given angular movements of the joystick and responses of the missile control surfaces.

Length of the joystick affects control sensitivity in that the knob of a longer joystick must be displaced further to obtain a given response of the missile control surfaces. In studying the problem of how long to make the joystick and thereby obtain the proper control sensitivity, accuracy of control was investigated with stick lengths of 2.18, 3.18 and 4.18 inches.

The use of lead to compensate for the "time lag" in the system may be accomplished by momentarily increasing the amount of "command" given to the control surfaces of the missile. In arriving at the "amount of lead" to use, a time lag in the system of approximately 0.2 seconds was assumed. Accuracy of control was then compared with that obtained without lead in the command link.

The particular lead circuit studied consisted of an RC network with a time constant between 0 and 0.20 seconds depending upon the position and rate of movement of the joystick. In an experiment subsequent to those reported in this paper, different time constants were investigated. The results of the later experiment were consistent with those reported below.

The transfer functions studied are described by the curves of figure 1 in which stick deflection in degrees is plotted against normalized missile response. It can be seen that the sensitivity of transfer function 2 at the point of greatest difference is approximately 25 percent less than that of transfer function 1. The primary objective behind the comparison of the two functions was to determine if the reduced sensitivity of transfer function 2 would make pilot performance less erratic and thus improve accuracy of control.

Apparatus

A block diagram of the apparatus employed in the study is presented in figure 2. The apparatus simulated a four degree of freedom air to ground missile system. Missile dynamic equations were programmed into a Page Analog Computer. The missile dynamic equations took the form:

$$\ddot{Y} = A(x)\delta_y + B\dot{Y} + CY - D\dot{X} \qquad (1)$$

$$\ddot{X} = A(y)\delta_x + D\dot{X} + CX - D\dot{Y} \qquad (2)$$

which define angular motion in a vertical and horizontal plane respectively.

The variable function A and the constants B, C and D are determined by a specific missile configuration. The functions described by δy and δx are the pilot's control inputs and take the form of a pulse train with a constant period. The average values of the pulse trains are represented by the transfer functions of figure 1.

The missile kinematic equations employed in the study took the form:

$$\ddot{H} = EY \qquad (3)$$

$$\ddot{V} = FX \qquad (4)$$

where E and F are constants determined by a specific missile configuration. The symbols H and V represent the horizontal and vertical acceleration components respectively.

These accelerations were combined in the computer with programmed acceleration, errors ε_H and ε_V as follows:

$$\ddot{H} + \varepsilon_H$$

$$\ddot{V} + \varepsilon_V$$

The resulting functions were used to obtain H and V as shown later in equations 5 and 6.

The missile was represented by a point of light on the cathode ray tube which was located in the rectangular enclosure shown in figure 3. The joystick was placed in front of and to the left of a conventional pilot's seat as shown in the figure. Voltages generated by movements of the control stick were transmitted through the control mechanism to the computer. Inputs to the cathode ray tube were then generated by the computer.

The joystick as indicated by figure 3 was operated by the subject's left hand. The left hand operation was necessary inasmuch as the pilot's right hand is used to control the aircraft. Firing of the simulated missile was accomplished by the push button shown in the subject's right hand.

A sketch of the simulated missile, target presentation and the cathode ray tube is presented in figure 4. Diameter of the point of light was 2.4 millimeters. The target was represented by a red, transparent circle having a line width of 3.2 millimeters. The inner diameter of the target circle was also 3.2 millimeters. The cathode ray tube was a 27-inch commercial type.

Prior to the start of a simulated flight problem, the missile was centered in the target. At the instant of firing, constant programmed errors (E_h and E_r) caused the missile to diverge from the target at one of several predetermined accelerations. These errors were combined with the corresponding missile accelerations. Then the resulting functions were solved by the computer in the following manner:

$$H = \iint (H + \epsilon_h)\, dt\, dt \qquad (5)$$

$$V = \iint (V + \epsilon_v)\, dt\, dt \qquad (6)$$

The horizontal (H) and Vertical (V) missile displacements were then presented on the cathode ray tube. Subjects using the joystick were required to correct observed errors by returning the missile to the center of the target and maintaining it in this position.

The response capability of the missile at maximum deflection of the joystick was always greater than the error acceleration. Values of the error accelerations used and the response capabilities of the missile are included in the description of the experimental program.

The Experimental Program

Four experiments were conducted in the course of the experimental program. The factors investigated in experiment 1 are listed in table 1. Although of secondary interest, error accelerations and direction of the error acceleration were taken as treatments and their effects evaluated in the analysis of results. The missile response capability in this experiment was 158.2 mm/sec^2 on the face of the cathode ray tube.

A treatments X subjects experimental design was used. In this design each subject was tested at every combination of treatments. Treatments were presented in a balanced order to compensate for any learning on the part of the subjects. Each subject was given 32 problems at each combination of lead, transfer function and stick length. These were presented in groups of four corresponding to the four error acceleration-direction combinations. Within groups of four the order of presentation was random. The first 24 of the 32 problems were used to train the subjects under the new parameter combination. The last 8 at each combination were taken for analysis.

As Table 1 indicates four subjects were used. All were Martin Marietta employees with recent jet pilot experience. The subjects had been used in earlier experiments of a similar nature and thus were already familiar with the basic control task.

Flight time of the missile was set at 20 seconds for all problems. Subjects were scored on the basis of the total time the missile was maintained within an imaginary circle taken about the center of the target. The diameter of this imaginary or criterion circle in experiment 1 was 40.6 mm. Subjects were at no time aware of the limits of the criterion circle. Their instructions were to keep the missile centered within the 3.2 mm diameter of the inner target circle.

Time for scoring purposes was measured by the computer and displayed on a digital voltmeter at the end of each problem.

With the exception of the missile response capability, error accelerations and the diameter of the criterion circle, all parameters and conditions in experiments 2 and 3 were identical to those in experiment 1. Values of the new parameters are listed in table 2. The diameter of the criterion circle was changed in experiments 2 and 3 to compensate for the easier control tasks provided by the new error acceleration and missile response capabilities.

Experiment 4 was conducted to compare the effects of the 2.18 and 3.18 inch stick lengths. The error accelerations, missile response capability, diameter of the criterion circle and direction of error accelerations, were the same as in experiment 3. A balanced experimental design was used in which each subject was tested at each stick length. The same four subjects employed in earlier experiments were used in experiment 4. Lead and transfer function 1 were used.

Results of Experiments 1, 2, and 3

The mean values for the parameters of major interest in experiments 1, 2, and 3 are presented in table 3. The difference in time within the criterion circle for lead versus no lead varied from a maximum of 2.0 seconds in experiment 1 to 1.1 second in experiment 3. Differences for transfer functions and stick lengths were in general smaller.

In the analysis of the data from experiments 1, 2, and 3, tests for homogeneity of variances showed that the variances for lead in all three experiments were heterogeneous. An attempt was made to find a suitable transformation of the data. This failed so the Wilcoxon Matched Pairs Signed Ranks Test was used which did not require homogeneity of variances. The results in all three experiments indicated that the use of lead

Table 1

Factors Investigated in Experiment 1

Factor	Level
Lead	Lead and no lead
Transfer Function	Transfer Functions 1 and 2
Stick Length	2.18 and 4.18 inches
Error Acceleration	98.2 and 32.7 mm/sec^2
Direction of Error Acceleration	180° and 300°
Subjects	4

Table 2

New Parameter Values in Experiments 2 and 3

	Error Acceleration (mm/sec^2)	Missile Response Capability (mm/sec^2)	Diameter of Criterion Circle (mm)
Experiment 2	6.7 and 2.2	10.7	10.8
Experiment 3	27.9 and 9.3	45.0	20.3

Table 3

Mean Values for Experiments 1, 2 and 3

Source	Mean (Secs)		
	Exp. 1	Exp. 2	Exp. 3
Lead	11.5	17.8	15.4
No Lead	9.5	16.3	14.3
Transfer Function 1	11.1	17.4	15.1
Transfer Function 2	10.1	16.8	14.5
2.18 inch Stick Length	10.6	17.5	15.1
4.18 inch Stick Length	10.6	16.7	14.5

in the command link to compensate for system "lag" significantly improved accuracy of control over that obtained without lead. All tests of significance in the experimental program were made at the 0.05 level.

The variances for other parameters were homogeneous, so the data were evaluated by analysis of variance. The results are presented in table 4. The difference between transfer functions, between stick lengths, and between direction of error acceleration were not significant. In every case differences between error accelerations and among pilots were significant.

F-ratios for the various interactions are not presented because of the large number involved.

Results of Experiment 4

The mean values for the 2.18 and 3.18 inch stick lengths obtained in experiment 4 are presented in table 5. As the table indicates, the difference is very small.

All variances were homogeneous, so the analysis of variance was employed in the evaluation of the data. The results in table 6 show that the difference between stick lengths was not significant.

Conclusions

It was concluded on the basis of the experimental results that significant improvement in

Table 4

Summary Analysis of Variance, Experiments 1, 2 and 3

Source of Variance	Degree of Freedom	F Ratio		
		Exp. 1	Exp. 2	Exp. 3
Lead	1	--*	--	--
Transfer Function	1	1.4	9.5	1.43
Stick Length	1	Less than 1	4.9	1.8
Error Acceleration	1	135.8**	76.3**	34.6**
Direction of Error Acceleration	1	7.9	4.8	7.2
Pilots	3	7.2**	11.6**	17.8**

*Heterogenous Variances

**Significant at 0.05 level

Table 5

Mean Values for Experiment 4

Source	Mean (secs)
2.18 inch Stick Length	16.3
3.18 inch Stick Length	16.6

Table 6

Summary Analysis of Variance, Experiment 4

Source of Variance	Degree of Freedom	F-Ratio
Stick Length	1	0.1
Error Acceleration	1	614.1*
Direction of Error Acceleration	1	1.6
Subjects	3	1.3

*Significant at the 0.05 level

accuracy of control can be obtained by incorporating a lead circuit in the command link. The increase in accuracy to be obtained is expected to range from 10 to 20 percent. This finding verifies the results of experimental studies which indicate that compensating for lag improves accuracy of control.

The subjects performed equally well with the two transfer functions, indicating a high degree of adaptability to changes in sensitivity between movements of the joystick and responses of the missile. This adaptability was also apparent in the investigation of stick lengths, in that no significant difference in accuracy of control was found with the 2.18, 3.18, and 4.18 inch sticks. It was concluded that relatively large changes in control sensitivity can be compensated for by training of operating personnel. In contrast the effects of lag could not be overcome by training.

It was demonstrated by the experimental program, of which the present series of investigations was a part, that excellent results in establishing design parameters can be obtained by organized experimentation.

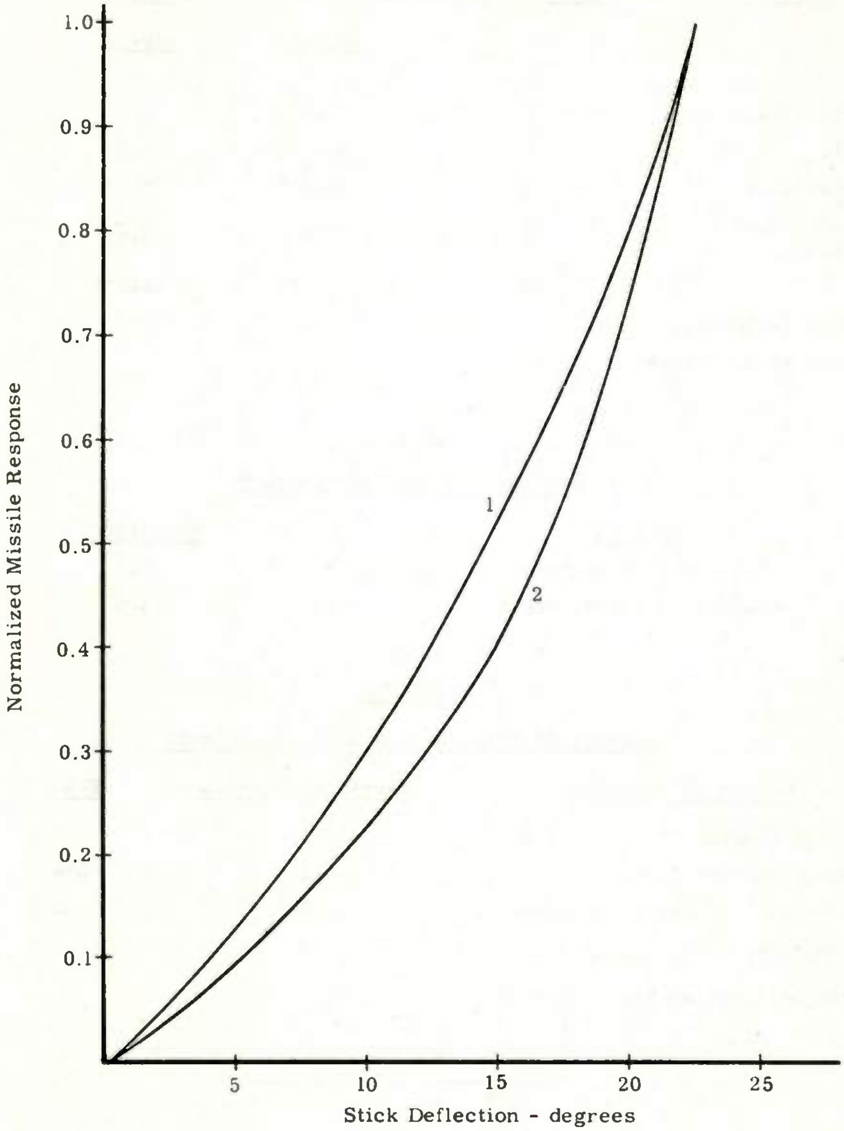


Fig. 1. Transfer functions.

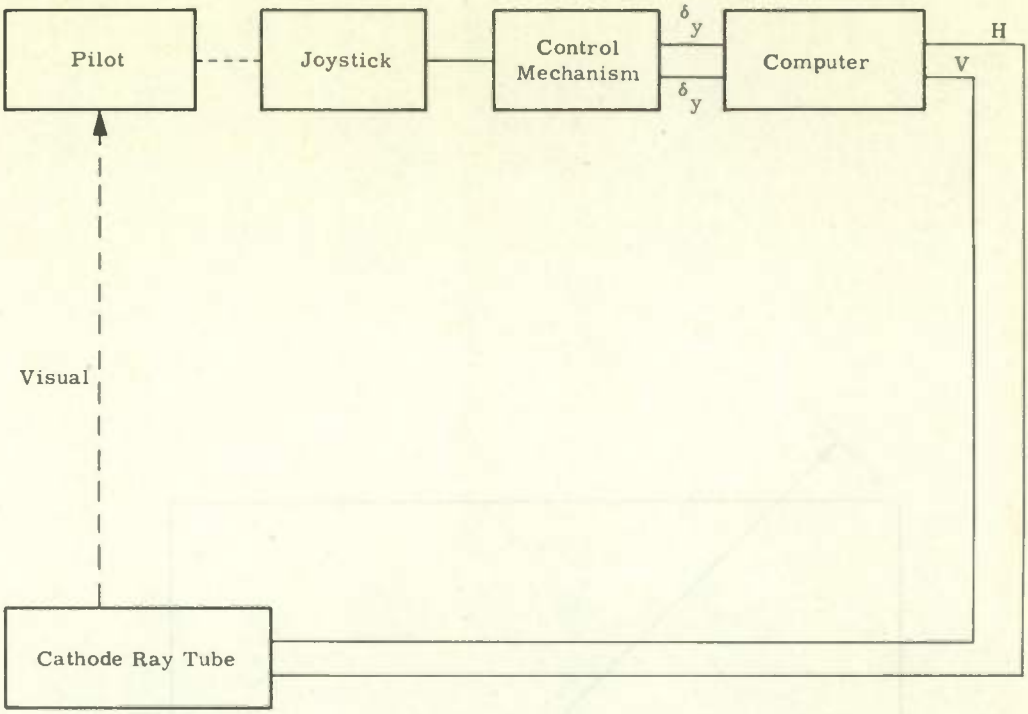


Fig. 2. Block diagram of apparatus.

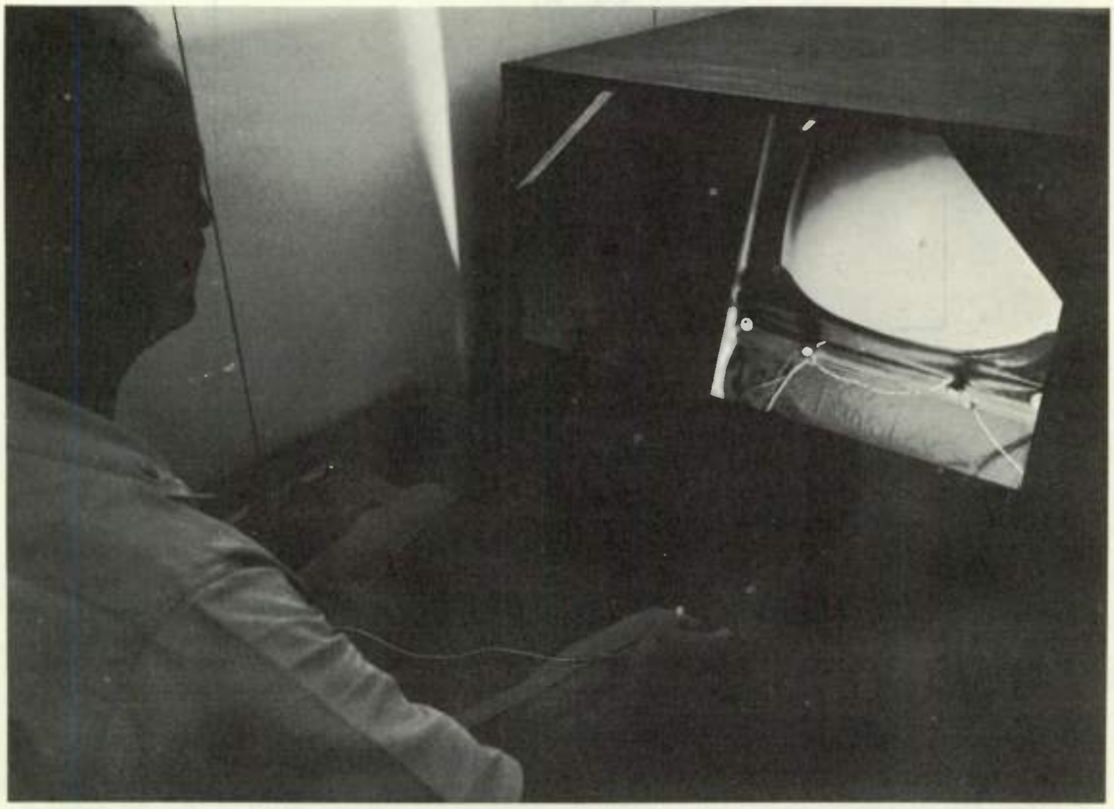


Fig. 3. The simulator.

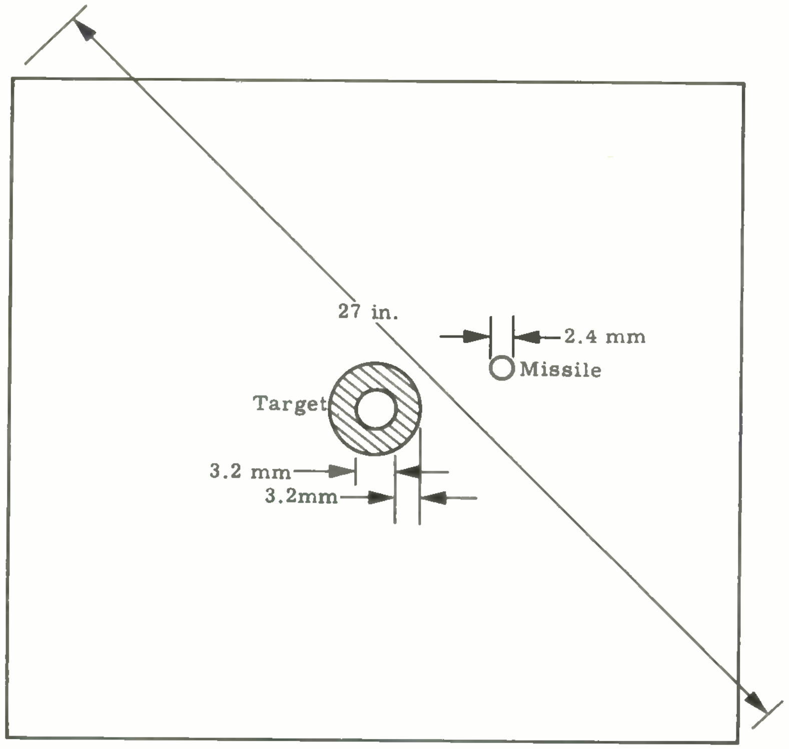


Fig. 4. Size of missile and target on cathode ray tube.

A SIMULATION STUDY OF
OPERATOR CAPABILITY IN ROBOT VEHICLE CONTROL

M. Chomet, N. Freeberg, and A. Swanson*

Airborne Instruments Laboratory
A Division of Cutler-Hammer, Inc.
Deer Park, Long Island, New York

I. Introduction

Requirements of future space systems are almost certain to include remote control of vehicles thousands of miles from the point of control origin.

Earth-based control of a vehicle which would traverse the lunar surface has recently been under consideration. Earth-based control of space vehicles engaged in rendezvous and docking maneuvers for purposes of join-up, inspection, or repair, represent other likely areas for remote control.

Human operators will be part of the control loop in such systems; not merely as backup for system malfunction, but because most of the required control inputs could not be anticipated reasonably--that is, there is little probability of our mapping the lunar surface to a degree which would allow devising a preprogrammed course for a surface roving vehicle. Human decisions and control responses will thus be advantageous and based largely upon some form of pictorial information displayed to an operator (such as, video, radar, or infrared displays). (See Figure 1 for a block diagram of such a control situation.)

Given this dependence upon a pictorial display, there are known to be rather unique operator problems which stem from interpretation of such visual information combined with the presence of a lengthy control-display time lag imposed by signal transmission time. This time lag is about 2.6 seconds for a signal traversing the distance from earth to moon and return. If a remotely controlled vehicle on the lunar surface carries a video visual sensor, it would be about 3 seconds (including scanning time necessary to produce one frame) before an operator could observe the effect of a control input. Few people are accustomed to dealing with such lengthy control-display time lags especially when coupled with primary dependence upon an "artificial" pictorial image.

In earlier work (references 1 and 2) at Airborne Instruments Laboratory (AIL) it was found that, under such conditions, operators show poorer driving accuracy than when there is no delay in the system. When driving between numerous obstacles operators can become geographically disoriented with a simple video display and a 3-second control delay. In addition, anticipatory control inputs are limited (3 being about maximum for experienced operators) so that control impulses generally deteriorate to repeated stop and start driving.

The present study was undertaken to determine, more precisely, operator capability under a 3-second control-display time lag condition as a function of different video display techniques and the complexity of the control task required.

The simulation of the control of a lunar robot vehicle from earth must take into account the time lag inherent in the system. This time lag, defined as the time between the issuing of a control on earth and the receipt of any verification that the command was received by the vehicle on the lunar surface, consists of two basic delays. The first delay is caused by the round-trip transmission of radio signals between the earth and the moon and is about 2.6 seconds. The second delay is the result of information being presented to the operator on a TV display. This delay is caused by the scanning time necessary to produce one frame. Even higher frame rates would still result in some delay. However, there is a practical upper limit to increasing picture frame rate even before consideration of power and bandwidth. Previous studies (reference 2) of this problem have come to the general conclusion that frame rate will be quite low for earth-moon transmission. For this study a frame rate of 4 per second was hypothesized. The resultant system time lag, therefore, is the summation of the individual delays, or about 3 seconds.

II. Apparatus

A. Simulation of System Time Lag

The 3-second system time lag existing in the control-display loop can be lumped at any one point in the loop for purposes of simulation.

For the simulation described in this paper, the entire time delay was lumped into the control portion of the loop. No attempt was made to reduce the TV frame rate, or delay the transmission of the video picture.

* The authors wish to acknowledge M. Pollack, B. Cantone, and T. Paige for their assistance in carrying out this study.

The 3-second delay was produced by reducing the control stick movements to a binary code in which form they entered electronic shift registers (Table I). The output of the shift registers energized relays which routed appropriate power to the vehicle (Figure 2). The shift registers used in the simulation were constructed of standard AIL digital modules. Each shift register was 20 bits long. The shifting rate was such that the total transfer time was 4 seconds, giving a resolution of 0.2 second per bit. The shift registers were tapped at each bit. Therefore, the delay in the simulator could be varied from 0 to 4 seconds in 0.2 second increments.

Table I

Coding of Control Commands

Command	Sw 1	Sw 2	
Stop, no power	0	0	
1/2 speed fwd	1	0	
Full speed fwd	1	1	0 - Switch in normal condition
Reverse	0	1	1 - Switch in momentary condition

B. Simulator Vehicle

The simulator is a dual-tracked, electronically powered chassis on which a TV camera is mounted. It is about 14 inches long, 12 inches wide, and 12 inches high (not including the overhead viewing globe). Use of separate reversible motors for each track permits individual as well as concurrent track motion in forward and reverse direction. This track motion permits the vehicle to make turns and other tractor-like maneuvers.

Power is fed to the vehicle through a slip-ring assembly and a cable connected to an overhead dolly. This combination permits complete freedom of movement in the simulator room. The video information is transmitted back to the TV console using a standard AM technique on one of the TV channels (generally channel 3 was used). A liquid pen attached to the rear of the vehicle was used to trace its course.

The simulator room is about 24 by 15 feet in size. Solar illumination effects (if desired) are simulated by the use of high-wattage lamps, and by adjustment of the video brightness and contrast controls.

It is important to bear in mind that this vehicle does not represent a lunar vehicle nor any attempt at lunar vehicle design, but rather a convenient means of simulating such a vehicle.

C. Views

Three operator views or presentations were chosen for evaluation in this study. The first was a forward, wide-angle (55°) view, the second an overhead (360°) circumferential view, and the third a stero forward view.

The first view was obtained by the normal lens as installed on the closed-circuit TV system. This lens system was used in earlier studies (reference 2) on the effects of control delay on operator performance. Figures 3 and 4 show the operator views for both the forward and the circumferential views.

The second presentation (circumferential) used a gazing globe suspended just above the vehicle, and a flat mirror in front of a 45° viewing lens. Figure 5 is a photograph of the vehicle with the globe installed. The overhead or circumferential view resulted from recommendations of earlier studies. These studies indicated that the limited forward view was a hazard in obstacle avoidance. This was particularly the case for the reverse maneuver, during which the probability of a collision would increase since the operator could not see to the rear of the vehicle.

The circumferential view allows an operator to see 360° about the vehicle. When this presentation is used the vehicle is seen near the center of the screen and remains always facing the same position relative to the console. The view can best be described as giving the sensation of driving on a sphere.

The third presentation used split optics to produce stereo images. As shown in Figure 6 a stereo adapter produced two images on the vidicon face plate. These images were then transmitted to the display unit at the control console. The images, which were separated and displayed side by side, each had a polarizing filter placed before them on the viewing screen. These filters polarized the light emanating from the two pictures at right angles to each other. The operator wore a special set of prisms (shown in Figure 7) which allowed him to adjust the image received at each eye and to overlap the images to produce a stereo effect. The prisms have a set of polarizing filters which allow the proper image to be transmitted to each eye.

Although a distinct stereo image could be obtained by this system a number of difficulties in its use forced its elimination as a display variable.

Primarily, the less than ideal resolution and quality of the video image made it difficult to hold the stereo effect for the length of time required by the experimental situation. Ocular fatigue led to constant splitting of the image, while other problems in operator performance were introduced by the restricted freedom of head movement required to maintain the stereo effect. Subject performance became severely degraded and erratic as a result of difficulties in maintaining the stereo effect so that its use was considered an unfair estimation of stereo per se as a display variable. While stereo is apparently of value for fine manipulation in remote handling operations (such as, industrial application for hot-cell work) its value in guidance maneuvers, under the conditions imposed by this study, could not be properly evaluated.

III. Method

A. Experimental Procedure

The subjects (12 adult males) were required to guide the robot vehicle along each of three video-presented courses (Figure 8). Each course was presented under one of two display conditions (stereo having been eliminated as a technique owing to difficulties previously described).

Each course appeared as a white path on the screen which the subject was required to follow "as quickly and accurately as possible" until he reached a designated end-point.

The first video display techniques was a direct or straight view showing a standard "head-on" presentation of the course as it might be seen by an observer looking out from within a vehicle (Figure 3). The second display technique employed the panoramic view which gave the subject a perspective looking down from above the vehicle (Figure 4).

Three levels of course complexity (low, medium, and high) were used, with "complexity" being defined by the number of turns in a 15-foot course. The simplest course had two turns, the moderate one, three and the most complex course four turns. Angularity of the turns was also made successively sharper for increasing levels of complexity. (See Figure 8 for a sketch of the course layouts).

A simple training course was also used (one turn) in order to acquaint the subject with the vehicle controls and the task, as well as to bring him to a predetermined level of ability before beginning the experimental trials. The training criteria chosen required the subject to run the training course in less than 1 minute with no deviation greater than 5 inches from the path. Subjects averaged about 2.5 training trials before qualifying. No visual cues of any sort were provided in the field of view, other than the white course against a dark grey background. All runs were made with a 3-second delay between control input and vehicle response.

B. Performance Measures

Subject performance was measured under each display and course condition by taking: (1) deviations of the actual path, traced by the vehicle, from the video presented path, (2) time to complete the course and, (3) the frequency of Stops made by the vehicle in traversing the course.

C. Statistical Design

The order of experimental conditions was randomly assigned for each subject, who ran each of three courses under two display conditions--for a total of six trials. This 2 × 3 factorial design was employed for the analysis of variance tests. Correlation coefficients were obtained between Deviation and Time scores and between Deviation and Stop scores.

IV. Results

A summary of the analyses of variance for Deviation Scores, Number of Stops and Time to run the courses is presented in Tables II through VII.

For the two variables of Stops required and Time to run the courses, the overall differences between the display techniques and the three levels of course difficulty are highly significant ($P < .001$ confidence level). The scores for both measures (Stops and Times) favored the panoramic display. On the other hand, the difference between displays, as measured by Deviation scores only (Tables II and V) favored the Straight (Direct) view--although at a lower confidence level ($P < .05$). For all performance measures (Deviation Scores, Stops, and Time), there was poorer performance with increasing course complexity--that is, number of turns required. No significant interactions were found.

The relationship between course Deviations and Time to run the course was a positive coefficient of +0.29 (significant at the .05 confidence level) indicating a slight tendency for subjects who ran the course in slower time to be less accurate. No significant relationship was found between scores for Deviations and Stops.

Table II

Mean Deviation Scores by Experimental Condition

	Display Technique	
Course Complexity	Direct View	Panoramic View
Low	6.83	9.58
Medium	7.92	10.17
High	13.75	17.25

Table III

Mean Stop Scores by Experimental Condition

Course Complexity	Display Technique Direct View	 Panoramic View
Low	18.1	12.8
Medium	19.0	16.8
High	24.4	19.6

Table IV

Mean Time Scores by Experimental Condition

Course Complexity	Display Technique Direct View	 Panoramic View
Low	8.92	7.08
Medium	10.33	8.92
High	13.58	10.83

Table V

Analysis of Variance for Course Deviations

Source	Sum Squares	df	Mean Squares	F
Rows (course complexity)	765	2	382.5	15.9 ($P < 0.001$)***
Columns (displays)	145	1	145.0	6.04 ($P < 0.05$)*
Interaction	4	2	2.0	
Within groups	1598	66	24.0	
Total	2512	71		

Table VI

Analysis of Variance for Stops Taken

Source	Sum Squares	df	Mean Squares	F
Rows (course complexity)	514	2	257.0	13.82 ($P < 0.001$)***
Columns (displays)	304	1	304.0	16.34 ($P < 0.001$)***
Interaction	30	2	15.0	0.81 (N.S.)
Within groups	1230	66	18.6	
Total	2078	71		

Table VII

Analysis of Variance for Time

Source	Sum Squares	df	Mean Squares	F
Rows (course complexity)	216	2	108.0	19.6 ($P < 0.001$)***
Columns (displays)	72	1	72.0	13.1 ($P < 0.001$)***
Interaction	6	2	3.0	0.5 (N.S.)
Within groups	366	66	5.5	
Total	660	71		

* = .05 Confidence Level.

*** = .001 Confidence Level.

V. Discussion and Conclusions

In examining the panoramic display technique, it was believed that this provided considerably more information regarding the vehicle's track along the path. The results for course Deviations which indicate otherwise, and favor the straight view for accuracy are, however, explainable.

With the panoramic viewer small deviations of the vehicle from the path do not appear as great as when viewing the path at "windshield level." Thus, the driver tends to accept being "generally" on course with the panoramic viewer.

By contrast, similar amounts of camera lens displacement for a head-on view result in a larger relative displacement on the screen and what appears to the operator as greater error. In addition, since the forward view has a restricted field of vision, the operator is under greater constraint to stay within this field to keep the path in view.

The greater accuracy is obviously achieved for the Direct View by taking greater care in maneuvering and by sacrificing time as well as number of stops.

Where power requirements dictate covering the greatest distance with the least number of vehicle starts and stops, an overhead view can be more efficient. Where accuracy of vehicle movement is critical, on the other hand, an operator should be provided with a "windshield" view. Obviously, each display technique would have its place under different system conditions.

There is little doubt that course "complexity" was properly defined in this study, since the greater the number of turns and the steeper their angle, the poorer the operator performance. That is, more stops, longer time, and less accuracy coincide with increased course complexity.

This study represents an initial attempt at defining display and control problems and possible solutions involved in robot-vehicle displays. Other display techinques--primarily predictor displays--will require evaluation. More effective control techniques than the one used in this study are feasible, and should be carefully evaluated before optimum control-display systems can be recommended for specific robot-vehicle mission requirements.

References

1. S. H. Gross, "Controlling a Lunar Vehicle from Earth," presented at the ARS Space Flight Report to the Nation, New York Coliseum, N. Y., 9-15 October 1961.

2. "Lunar Robot Vehicles - A System Study," Airborne Instruments Laboratory, J-9305A, August 1961.

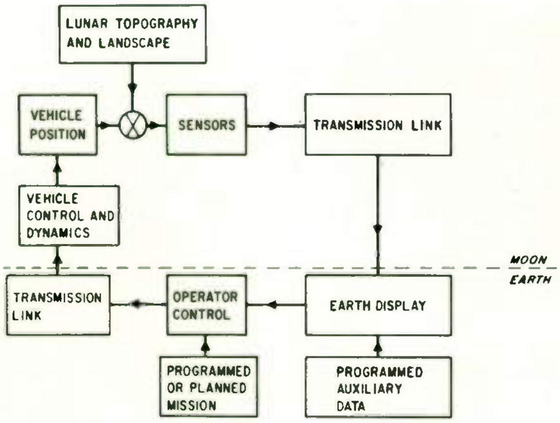

LUNAR TOPOGRAPHY AND LANDSCAPE
VEHICLE POSITION
SENSORS
TRANSMISSION LINK
VEHICLE CONTROL AND DYNAMICS
MOON
EARTH
TRANSMISSION LINK
OPERATOR CONTROL
EARTH DISPLAY
PROGRAMMED OR PLANNED MISSION
PROGRAMMED AUXILIARY DATA

Fig. 1. Remote lunar vehicle control system.

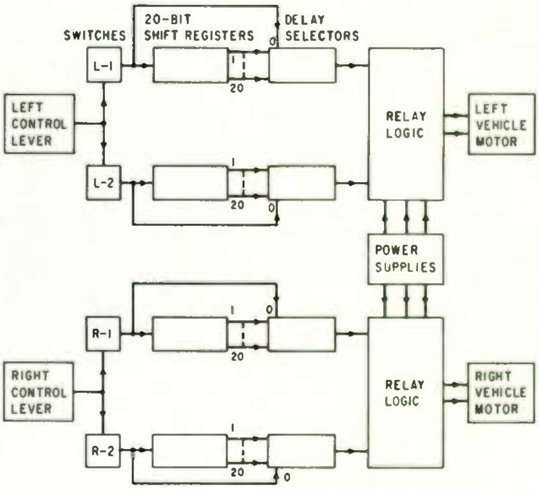


Fig. 2. Block diagram of simulator vehicle control.

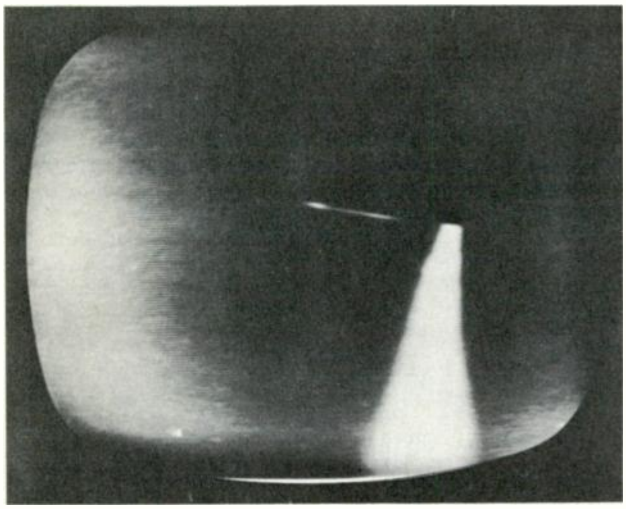


Fig. 3. TV screen, direct view.

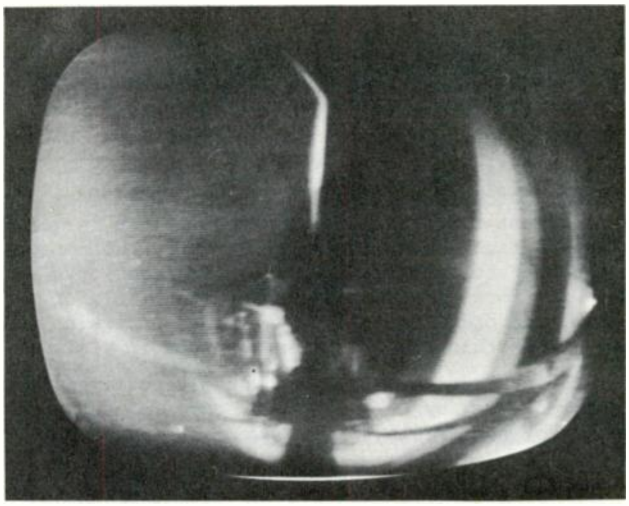


Fig. 4. TV screen, circumferential view.

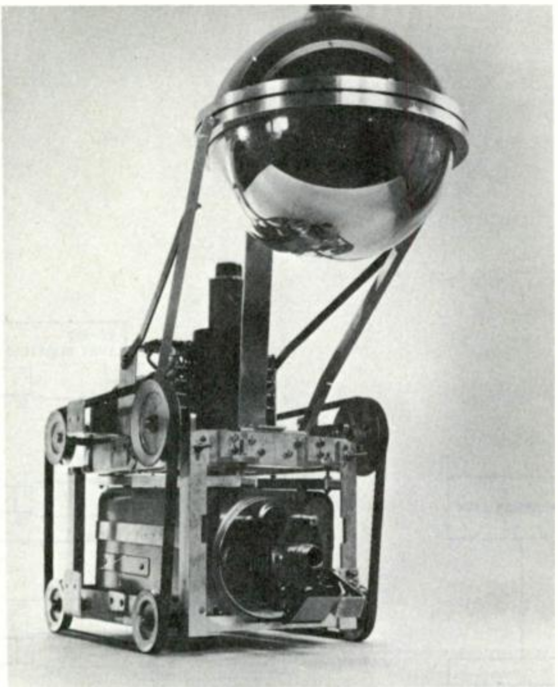


Fig. 5. Model of remote lunar vehicle.

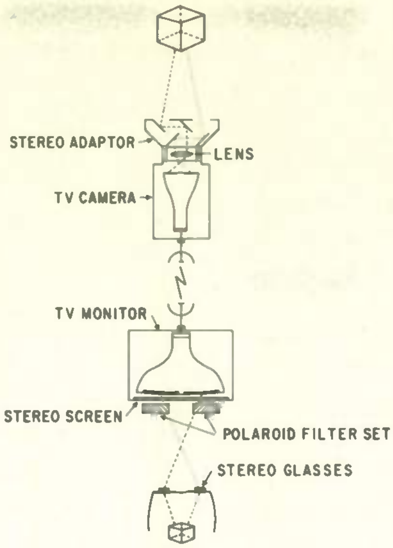


Fig. 6. Stereoscopic TV system.

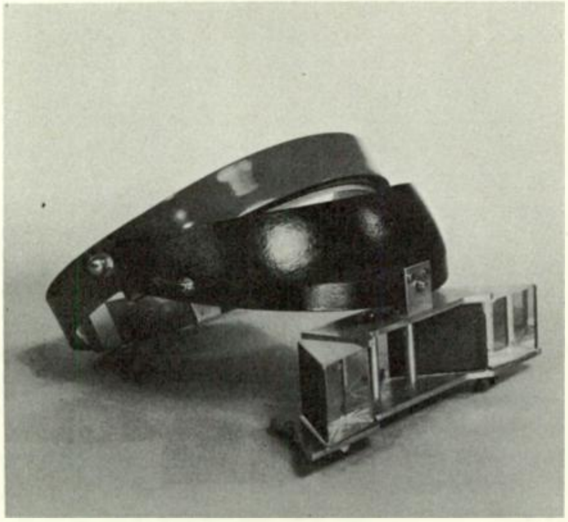


Fig. 7. Polaroid headset.

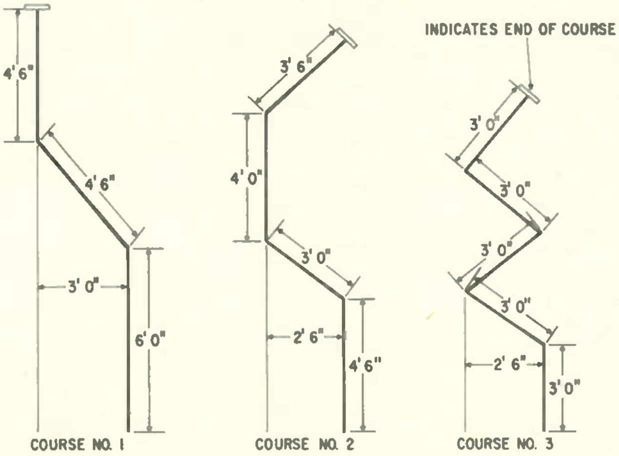


Fig. 8. Course layouts.