THE RADIO AND ELECTRONIC ENGINEER

The Journal of the British Institution of Radio Engineers

FOUNDED 1925 INCORPORATED BY ROYAL CHARTER 1961

"To promote the advancement of radio, electronics and kindred subjects by the exchange of information in these branches of engineering."

VOLUME 25 MARCH 1963 NUMBER 3

ELECTRONICS AND PRODUCTIVITY

THIS is National Productivity Year, and in response to the appeal by His Royal Highness The Duke of Edinburgh, every professional engineering institution in Great Britain is arranging meetings which have the aim of increasing productivity. The Institution's contribution to National Productivity Year is to devote the 1963 Convention to the role that Electronics can play, and is playing, in increasing Productivity. The Convention will be held at the University of Southampton from 16th–20th April.

The theme was suggested by the President of the Institution when he attended a meeting of the Council on 8th February 1962. Subsequently the Convention Committee has arranged a programme to include papers and discussions of interest to engineers in many different fields of industry.

Firstly, the Convention will provide a forum for electronic engineers who are engaged in what may broadly be called "industrial electronics" to hear and discuss papers on new ideas and developments; techniques of considerable interest will be described and, as always at Conventions of this kind, engineers in other branches of electronics will undoubtedly find many ideas presented which are relevant to their own specializations. Such engineers form therefore the second group which the Committee has in mind and, as the real value of a Convention lies in the exchange of differing points of view, it is hoped that many engineers from outside the immediate field of industrial electronics will attend.

The third group of engineer for whom the Convention will prove a valuable source of information will be the electronic engineer working in another industry or the engineer of another discipline who is adopting electronic techniques to his industry's problems. This group will certainly reap considerable benefit from the Convention as indeed was the case at Brit.I.R.E. Conventions held in 1954 and 1957 on "Industrial Electronics" and "Electronics in Automation"; it is with this group in mind that the Committee has invited several papers describing specific applications of electronics to a variety of industries.

From their titles the sessions on two of the days, namely "Sensing Devices, Measurement and Telemetry" and "Control and Information Processing", might appear to be of interest primarily to the first and second groups, and the third full day on "Industrial Applications of Electronic Systems" to be intended mainly for the "user" engineer. However, the lines of division have not been rigidly drawn since in some of the papers the techniques of the system described are closely joined to a particular industrial application—but are not necessarily limited only to that application. Instances of this include the "control and information processing" paper describing the computer control on a steel mill and in other papers describing measuring techniques.

The full list of papers now published on page 194 forms an impressive overall picture of present and future applications of electronics to the solution of production problems. The adoption of some of these techniques to wider fields of use could well be immediate—in others the electronic engineer is describing new aids to production, the engineering and development of which will require the closest collaboration of the user to suit best his particular need.

The Committee believes that the Convention represents one of the most important contributions which the Institution has made to the advancement of British industry. J.L.T.

The 1963 Convention

"ELECTRONICS AND PRODUCTIVITY"—SOUTHAMPTON, 16th-20th APRIL

The Institution's Contribution to National Productivity Year

TIME TABLE AND PROGRAMME

Tuesday, 16th April. INTRODUCTORY SESSION

AFTERNOON	Opening Address by J. L. THOMPSON, *Chairman of the 1963 Convention Committee.*
3.00–3.45	Survey papers:
	"Techniques for the Optimization of Controlled Processes." Professor A. D. Booth (*University of Saskatchewan*).
3.45–4.15	TEA
4.15–5.00	"Productivity and Technical Progress." J. Broderick (*Mullard*).
EVENING	
5.30–7.00	Institution Reception for Delegates.

Wednesday, 17th April. "SENSING DEVICES, MEASUREMENT AND TELEMETRY"

MORNING	*Chairman:* J. R. HALSALL
9.45–10.45	"Mass Flow Measurement with Turbine-type Flowmeters." I. C. Hutcheon and L. S. Duffy (*George Kent*).
	"An Electrodeless Conductivity Meter for Process Control." E. Harrison and P. F. Roach (*U.K.A.E.A.*).
10.45–11.15	COFFEE
11.15–12.45	"Process Analysis and its Application to the Oil and Chemical Industry." R. T. N. Bowen (*Elliotts*).
	"Activation Analysis in Steel Production." A. L. Gray, G. I. Crawford and G. D. Smith (*Plessey Nucleonics*).
	"Determination of Sulphur Content of Hydrocarbons by Bremsstrahlung Absorption Measurement." T. B. Rowley (*Isotope Developments*).
12.45–2.15	LUNCH
AFTERNOON	*Chairman:* P. HUGGINS
2.15–3.45	"Leak Detection and the Use of the Halogen Detector." T. S. Worthington.
	"Some Applications of Inexpensive Photoelectronics." R. D. Carter-Pedler (*Photoelectronics (MOM)*).
	"The Development of Eddy Current Testing Techniques for Tube Inspection." D. Terry (*Accles & Pollock*).
3.45–4.15	TEA
4.15–5.15	"A Frequency Meter with Continuous Digital Presentation." P. Wood (*Plessey*).
	"Voice Frequency Telemetry." K. A. Newman (*E.M.I. Electronics*).

Thursday, 18th April. Sessions on "CONTROL AND INFORMATION PROCESSING"

MORNING	*Chairman:* W. RENWICK
9.45–10.45	"The Application of Numerical Control to Machine Tools." H. Ogden (*Ferranti*).
	"A Tape Controlled Machine Tool System." P. H. G. Burgess (*Wickman*) and R. L. Duthie (*Parmeko*).
10.45–11.15	COFFEE
11.15–12.45	"Hybrid Digital/Analogue Servomechanisms." G. B. Kent (*Newman Electronics*).
	"Curve Generation by Interpolator Attachment to a General-Purpose Computer." J. S. Sibbald (*Ferranti*).
	"Some Principles and Circuit Techniques for Controlling Machine Tools from a Central Digital Computer." D. A. Bell and P. M. Will (*AMF British Research Laboratory*).
12.45–2.15	LUNCH
AFTERNOON	*Chairman:* Professor A. D. BOOTH
2.15–3.45	"The Economic Justification of On-Line Computers for Process Control." W. T. Lee (*International Systems Control*).
	"A Multi-Function Static Switching System for Industrial Applications." C. G. Cargill (*A.E.I.*).
	"The Development of 'ARCH'—A Modular Computer Control System." G. B. Cole and S. L. H. Clarke (*E-A Data Processing*).

3.45–4.15	TEA
4.15–5.15	"Design Features of a Digital Computer for Industrial Process Control." J. A. Freer (*International Systems Control*).
	"Low-cost Electronic Digital Arithmetic." D. W. Thomasson (*Ultra Electronics*).

Friday, 19th April. Sessions on "INDUSTRIAL APPLICATIONS OF ELECTRONIC SYSTEMS"

MORNING	*Chairman:* Professor E. E. ZEPLER
9.45–10.45	"The Analogue Computer as an Aid to Industry." A. W. O. Firth (*Redifon Flight Simulator Division*).
	"A High-efficiency Low-frequency Power Source for Vibration Excitation." B. H. Venning (*University of Southampton*).
10.45–11.15	COFFEE
11.15–12.45	"Data Logging in Power Generating Stations." W. E. Willison (*Elliotts*).
	"Computer Control Hierarchy in a Steel Plant." J. F. Roth (*Elliotts*).
	"Automatic High-speed Measuring System for Complex Products and Shapes." J. A. Sargrove (*Automation Consultants and Associates*).
12.45–2.15	LUNCH
AFTERNOON	*Chairman:* M. JAMES
2.15–3.45	"The Footprint and Wheel Measuring System for Steel Tubes and Bar." P. Huggins and R. Ashford (*T.I. Steel Tube Division*).
	"Automatic Control of Billet Saw Stops." G. Cooper (*Lancashire Dynamo Electronic Products*).
	"Electronics and Economics in the Handling of 35 mm Colour Negative Film." D. M. Neale, J. H. Coote and A. A. Large (*Ilford*).
3.45–4.15	TEA
4.15–5.15	"Limiting Waiting Time by Electronic Control of Materials Handling." Gordon Clark (*Mullard Equipment*).
	"Electronic Instrumentation for Increased Productivity in Petroleum Refining." W. H. Topham (*B.P. Refinery (Kent)*).
EVENING	
7.00	**Convention Banquet at the Guildhall, Southampton.**

Saturday, 20th April. Visit to Esso Oil Refinery, Fawley.

Synopses of Papers to be presented at the Convention

This is a further selection from the papers which have been accepted for presentation during the Convention. Other synopses were published in the February issue of *The Radio and Electronic Engineer*.

Tuesday afternoon 16th April. INTRODUCTORY SESSION

Productivity and Technical Progress

J. BRODERICK. (*Mullard.*)

Productivity is a word with many meanings, some quite specific and others general. In its most general form, it is used as one of the best indexes of the economic stature of a country and of its rate of economic growth. Although we tend to think of higher productivity as the result of more efficient work in factories, it is in fact influenced by an enormous variety of other things. Investment and expenditure on scientific research are commonly associated with the effort to raise productivity, but seen from an historical point of view, the key influences may well be modes of organization or general educational levels which enable the community to make full use of its technical knowledge.

One of the prime determinants of productivity growth is the rate at which technology penetrates everyday life. Although we use terms like "The Industrial Revolution" to describe major technological breakthroughs of past times, the penetration of a new technology is a much slower process than is generally realized. Speeding up this process is one of the main ways to raise national productivity and some of the issues can be exemplified in relation to the application of electronics to industrial and commercial processes.

Wednesday 17th April. SESSION ON "SENSING DEVICES, MEASUREMENT AND TELEMETRY"

An Electrodeless Conductivity Meter for Process Control

E. HARRISON AND P. F. ROACH. (*U.K. Atomic Energy Authority.*)

An electrodeless form of conductivity meter giving continuous indication and suitable for the range 0–1 mho/cm is described. The instrument is capable of working in highly radioactive and corrosive liquors with the minimum of attention. Tests in solution gave an error of ± 0·3% full scale. Fixed resistors, used to simulate a liquid loop, show an absolute accuracy of ± 0·1%.

Some Applications of Inexpensive Photoelectronics

R. D. CARTER-PEDLER, B.A. (*Photoelectronics (MOM) Holdings.*)

The paper describes three examples:

(i) An arrangement of photoelectric counters to discriminate between various sized articles passing through a laundry flat ironing machine and count the total of each. The combination of simple transistor photoelectric units and cold cathode tube timers provide accurate data from which incentive bonus is paid and work studies are made.

(ii) Textile cloth laying for multiple cutting (mass-production tailoring). A mobile edge alignment system fitted to a motorized reel transporter lays a length of cloth exactly over the previous length laid out on the cutting table to avoid any dragging or stretching of the fabrics. Thus the same patterns can be cut from a wide variety of materials, and the shapes, when cut, fall naturally into position when the garments are made up.

(iii) A method for continuous alignment for paper sheet in the manufacture of corrugated paper makes corrections more rapidly as the linear speed increases. Simple photoelectric sensing combined with discharge tube timers regulates the extent and frequency of correction, but the delay timer is designed to receive a signal from a tacho-generator on the machine which automatically adjusts the pre-set time in accordance with speed changes, so that each correction is realized at the sensing station before another correction can be made, despite speed changes.

Process Analysis and its Application to the Oil and Chemical Industries

R. T. N. BOWEN. (*Elliott Brothers (London).*)

Quality Control in the chemical and oil industries is usually understood to mean the control of a final or intermediate product within the limits of a laid-down specification. The specification states the tolerance allowed for certain physical properties or chemical purity of the product.

Quality control instruments fall roughly into two groups:

Instruments which measure chemical composition.

Instruments which measure a physical property.

The first group may be further divided into two:

(*a*) General purpose analysers which can measure several components in a plant stream or which can be used to detect one of a variety of compounds. For instance, gas chromatographs, infra-red analysers, X-ray fluorescent analysers, automatic titrometers etc.

(*b*) Specific analysers which measure one particular element or compound only. For instance, oxygen analysers, sulphur analysers, moisture monitors.

Examples of the second group are: viscometers, initial and final boiling point analysers.

The word *control* is a little loosely applied to these instruments as they are really sensing or measuring devices which feed conventional control loops and in many cases are used to monitor the plant streams rather than control them. However, most of the instruments mentioned have been used for control purposes, usually in a cascade system. As computer control technique develops, the form of plant control will assume immense importance.

Voice Frequency Telemetry

K. A. NEWMAN, LL.D. (*E.M.I. Electronics.*)

This is a frequency analogue telemetry system working over telephone lines and using audible tones in time division multiplex to convey information. The paper describes the various units and application of the system to such schemes as water undertakings, gas distribution, oil pipe lines and an automatic weather recording station are discussed.

Activation Analysis in Steel Production

A. L. GRAY, B.SC., G. I. CRAWFORD, PH.D., AND G. D. SMITH, C.G.I.A. (*Member*). (*Plessey Nucleonics.*)

The general principles upon which activation analysis is based are discussed briefly and the advantages of neutrons as the activating radiation are indicated. The essential constituents of an activation analysis system are then considered in more detail, and it is concluded that for "process control" analysis the use of a neutron generator based on the D-T reaction is most appropriate. The design and operation of such a generator are then described. The properties of the radiations emitted by the active nuclei and of the detectors which can be used to identify and measure these radiations are reviewed and consideration is then given to the requirements of the electronic circuitry required to analyse the output from these detectors.

The principles outlined in the first section are then illustrated by reference to equipment designed for the rapid measurement of the oxygen content of metals, which is described in some detail. The advantages of this type of measurement over conventional techniques are indicated.

The extension of the method to other problems in steel production is then discussed. Possible methods of analysing many of the elements of interest to steel makers are presented. The type of analysis and the problems encountered in the processing of the data obtained in such analyses are illustrated by a particular problem in the analysis of the materials on a sinter strand. The measures required to provide a solution to these problems are then discussed.

Thursday 18th April. SESSION ON "CONTROL AND INFORMATION PROCESSING"

The Application of Numerical Control to Machine Tools

H. OGDEN, B.SC.(ENG.). (*Ferranti.*)

The application of numerical control to machine tools has moved at a great pace over the last decade, and the results of vigorous technical endeavour by the electronics industry has set the scene on a revolution in methods of production which will ultimately affect every manufacturing industry.

The stage has been reached where the main efforts in the direction of continuous three-axis machining is concerned not with the machine tool control itself, but with the methods of input data preparation. Programming techniques using general-purpose digital computers have been developed to achieve area generation by sub-routine programming methods.

Techniques of measurement and servomechanisms have been applied to point-to-point numerical co-ordinate positioning systems for boring, drilling and punching applications. Advantage has been taken of high performance servomechanisms and high accuracy measuring systems to improve machine productivity and/or accuracy with a reduction of dependence on skilled operator attendance.

Basic linear electronic measuring systems derived from optical or induction techniques has enabled a numerical display of machine movement to be presented to the operator with versatility in respect to reset and reversal facilities, such that no arithmetic is necessary and any method of drawing dimensioning can be accommodated by such a system. The range of application of such measuring systems will be discussed.

A range of numerical inspection methods has been developed from the simple linear numerical display of 2D position to 3D point-to-point and continuous control. A simple application of electronics for the drilling of printed circuit boards will be illustrated.

Hybrid Digital/Analogue Servomechanisms

G. B. KENT, B.SC.(ENG.) (*Associate Member*). (*Newman Electronics.*)

The paper is an introduction to hybrid digital/analogue systems utilizing the application of digital techniques to conventional closed loop servo systems. The term hybrid is used to describe the systems where the error signal is derived digitally and then converted into an analogue signal for the operation of the power source. This results in an improved system accuracy in comparison to that obtainable using conventional analogue methods, without attracting many of the difficulties associated with the pure digital approach.

An introduction consisting of a comparison of the salient features of analogue and hybrid servo systems provides a background against which the underlying principles and design features of both velocity and positional servo systems are discussed. The constituent elements of each system are derived in block diagram form with parallel references to their analogue counterparts. The concept of basic logical elements is introduced and used to illustrate the *modus operandi* in preference to the pure circuit diagram approach.

The paper concludes with an assessment of the advantages accruing to hybrid servomechanisms and brief reviews of the various fields of application of such systems.

A Tape Controlled Machine Tool System

P. H. G. BURGESS (*Associate Member*) (*Wickman*) and R. L. DUTHIE (*Associate Member*). (*Parmeko.*)

Magnetic tape is employed to carry instructions for table movement which have been originated by the programming craftsman moving a stylus over a model. The application of the system to die sinking is described; the modest accuracy of $\pm$ 5 thousandths of an inch required in this work is considered to be typical of many applications for which machine tool control systems of much higher accuracy would be uneconomic.

Curve Generation by Interpolator Attachment to a General-Purpose Computer

J. S. SIBBALD (*Associate Member*). (*Ferranti.*)

The paper deals with continuous path control of machine tools by means of a general purpose computer with a simple off-line linear attachment whose design is described.

This arrangement replaces curve generation by means of a G.P.C./differential analyser technique, and is made possible by the increasing availability of high speed computers enabling curve generation to be carried out economically on the customer's own computer. In this way he is able to supervise the progress of workpieces from the drawing board to the shop floor.

The use of a modern large capacity computer also enables a simplification of the planning information required at the programming level so that simple geometric language may be employed to define the profile and appropriate sub-routines can be called up from the computer stores. From these data cutter off-set values are computed and combined with cutting sequence information to give the tool centre path.

Finally, the paper shows how the required variations in the cutting speed (due to acceleration and deceleration) are dealt with in the general purpose computer programme.

Some Principles and Circuit Techniques for Controlling Machine Tools from a Central Digital Computer

D. A. BELL, M.A., PH.D. and P. M. WILL, PH.D. (*AMF British Research Laboratory.*)

The advantages and economics of time-sharing a computer between a number of machine tool control systems are considered. Some of the circuit techniques required are described, e.g. encoders, quantizers, function generators etc.

A Multi-function Static Switching System for Industrial Applications

C. G. CARGILL, B.SC. (*Associated Electrical Industries (Manchester).*)

The paper will examine the design considerations involved in developing a static switching system for use in industrial applications. Some examples of the use of the system in industry are given (mining and steel industries). The design features which must be considered to produce a unit and ancillary equipment to permit customers to produce schemes of their own devising are mentioned. Finally, the economics of static switching systems and systems employing relays are contrasted.

The Development of ARCH—A Modular Computer Control System

G. B. COLE, B.SC.(ENG.), AND S. L. H. CLARKE, B.A. (*Associate Member*). (*E-A Data Processing.*)

ARCH is a hybrid analogue-digital system of computers for control applications. The need for both a design philosophy, and a range of modules with which a computer control system can be built without specialized circuit, or logical know-how is discussed. This is particularly important because the process knowledge which is required in the specification of systems is unlikely to be possessed by the computer specialist. The development of this philosophy is described with the economic and engineering implications which affect it.

The basic structure of the resulting system is described in some detail, together with some novel features of individual modules. In particular the "link" units which connect together the various controllers and regulators in a system to form a hierarchy of control are discussed.

Low-cost Electronic Digital Arithmetic

D. W. THOMASSON (*Associate Member*). (*Ultra Electronics.*)

The desirable performance characteristics of low-cost electronic computing devices are examined, and some design problems are considered, with special reference to the dangers of over-simplification.

From this is developed the concept of a range of computing devices derived from standard sub-units, covering both industrial and business applications.

Some practical cases are finally considered, to illustrate the difficulties likely to arise, and the need to defer to machine limitations in order to maintain economic manufacturing prices.

Friday 19th April. SESSION ON "INDUSTRIAL APPLICATIONS OF ELECTRONIC SYSTEMS"

Computer Control Hierarchy in a Steel Plant

J. F. ROTH, B.SC. (*Elliott Brothers (London)*.)

In a paper presented at the Brit.I.R.E. Symposium on "Recent Developments in Industrial Electronics", in 1962, a method of using a number of interlinked computer systems for the control of real time processes was described. The principles explained in this paper have since been realized in a hierarchy computer information and control system recently installed at the Spencer Works of Richard Thomas and Baldwins.

The present paper shows how the philosophy of the hierarchy system operates in practice, and what modifications have been found necessary. This has highlighted the importance of accurately specifying the requirements of the various sections of the hierarchy and of defining what facilities are really necessary for operating the plant with the minimum of dislocation in the event of a failure in one of the sections. It is shown that for this purpose a considerable amount of redundancy is essential. Lessons learnt concerning the physical form of the hardware and its maintenance will also be discussed.

The way in which the complete system is employed to collect plant data, issue shop floor instructions and prepare management reports, will be described. An assessment will be made of the advantages the system has provided in the manufacture of steel strip and indicate what further developments are envisaged.

Automatic Control of Billet Saw Stops

G. COOPER, B.SC. (*Lancashire Dynamo Electronic Products.*)

In hot sawing and shearing operations in a steel works, the length to be cut off is decided by the position of a stop which arrests the moving piece. It is frequently necessary to alter the measured length quickly and accurately and the alteration may be one of twenty or more feet.

The stop is designed to measure lengths from ten feet to seventy feet. It has six stopper heads mounted on a carriage which travels on wheels parallel to the roller table. Each head is raised and lowered by an individual air cylinder. The travel of the carriage need only be the distance between a pair of heads, that is, ten feet. It is obtained by a motor driven screw working in a nut mounted on the carriage. By selecting the appropriate head and setting the carriage in the ten feet range, any length from ten feet to seventy feet can be obtained.

The control equipment works on a digital system, the digital value of the required position being set up by the selector already mentioned. This is compared with the actual position of the carriage which is obtained from the digitizer operated directly from the moving carriage through a steel wire. The digitizer is of an electromagnetic type, which eliminates the possibility of trouble with contacts or filament bulbs. The logic units are of the standard transistorized plug-in pattern, which have the advantage of ease of servicing coupled with reliability in bad operating conditions.

Electronics and Economics in the Handling of 35 mm Colour Negative Film

D. M. NEALE, B.SC., J. H. COOTE AND A. A. LARGE. (*Ilford.*)

Large-scale printing of colour negative photographs has hitherto been relatively cumbersome and expensive. The Ilfocolor 35 mm system uses extensive mechanization to provide the customer with a better and cheaper service.

Continuous rolls of developed negatives pass through an automatic proof-printer which positions each picture, exposes a 1 : 1 "contact" print under photo-electric control, numbers and edge-signs the print and then advances to the next negative. A further machine automatically cuts, card-mounts and numbers the individual negatives and these are then returned to the customer with the processed strip of colour "contact" prints. Negatives chosen and returned by the customer for enlarged printing are loaded into a magazine on an automatic enlarger. The printing operation is again controlled photo-electrically and the number of prints determined by binary-coded perforations in the card-mount of the negative.

The design of the requisite electronic controls calls for photo-electric exposure control systems which are completely stabilized against photo-cell sensitivity variations. This greatly reduces wastage of time and materials.

Also, servo-controlled printing lamp intensities are needed so that dense and thin negatives can be printed in substantially equal times. Machines therefore operate continuously at optimum speed.

The Analogue Computer as an Aid to Industry

A. W. O. FIRTH. (*Redifon Flight Simulator Division.*)

The principles of analogue computation will be discussed briefly, and the application to more straightforward industrial engineering problems illustrated. The concept of simulation, so much an integral part of analogue computation, will be explained.

One use of the analogue computer, of increasing importance, is in the education of engineers for industry; it is of considerable importance to industry that personnel are able to make use of the latest computer techniques.

More advanced applications of the analogue computer, particularly those related to productivity, will be discussed—namely, the use of the analogue computer in solving optimization problems, and the use of the analogue computer in the study of production organizations. The latter topic is one of the most recently proposed applications for analogue computers, and can be extended to a whole field of economic simulation studies.

Automatic High-Speed Measuring System for Complex Products and Shapes

JOHN A. SARGROVE (*Member*). (*Automation Consultants and Associates.*)

A high-speed self-adaptive system of inspection is described and the reasons for evolving it are explained. Sequential measurement on a transfer-line is used to measure reliably to effective accuracies between $\pm$ 1 to 2 microns on complex shaped parts. Each measured dimension is stored up in analogue memory devices and later read-out as the object reaches the sorting or pass-reject gates. The decision to pass-or-reject the object is carried out in a self-adaptive logic system in which the subsidiary dimensions are compared with the dominant dimensions. The tolerant limit relating to the subsidiary dimensions are automatically shifted instantaneously for each object as the dominant dimension varies within tolerance, thus achieving critical inspection for true shape.

Machines are described for measuring conical and spirally grooved complex shaped objects such as twist-drills in more than one place. Also measurement of more abstract factors are referred to such as inspection at speeds of 200 measurements per minute involving computing for "density" and pass-reject action on this factor. Practical results have been obtained in machines which have inspected millions of objects. The cost of such apparatus is justified by reduction of scrap and manual inspection labour in production precision parts.

Limiting Waiting Time by Electronic Control of Materials Handling

GORDON CLARK, B.A. (*Mullard Equipment.*)

Situations often occur in industry where several machines or operatives have a common source of material. If demands for more material are dealt with in random sequence, or in accordance with a fixed routine, there will be cases where a machine or operator runs right out of material while others get their demands met more quickly than necessary.

A solution to this problem is to have an electronic system which acts like a queue and controls the routing of material so that demands are dealt with in the order in which they arise. Several versions of electronic queue are described together with their application in industry.

Problems which bear resemblance to queues also arise in connection with conveyors; it is sometimes necessary to identify packages as they are fed to a common conveyor, and then divert them at the appropriate collecting points along the route; sometimes inspection points cannot be located at the same place as the reject diversion mechanism, and a delay varying with conveyor speed must be introduced.

It is shown that electronic queuing systems can be used with advantage in these situations also.

Electronic Instrumentation for Increased Productivity in Petroleum Refining

W. H. TOPHAM, B.SC., A.R.I.C. (*BP Refinery (Kent)*.)

Electronic systems are replacing pneumatic equipment for automation process control in refineries. Conventional methods for measuring quality are being replaced by semi-automatic instruments, and the installation of continuous quality analysers on plants has led to significant improvements in efficiency. The design requirements for this type of equipment are considered and some typical instruments are described. Continuous analysers in combination with data logging and computing equipment can provide process data for control engineering and design studies. Quantity measurements and blending have been improved by the use of automatic level and temperature gauges, and flow meters with electrical pulse outputs. New instruments for plant inspection and maintenance may increase the on-stream efficiency of plants, and the use of electronic computers for operational planning makes the most effective use of available resources.

U.D.C. 534.88: 639.2.087.7

The Fisheries Application of Sonar

By

R. E. CRAIG†

Presented at the Symposium on "Sonar Systems" in Birmingham on 9th–11th July 1962.

Summary: Sonar has proved invaluable in pelagic fisheries as a direct detector of fish. In demersal fisheries the value of sonar except in Arctic cod fisheries is still mainly indirect—its chief value is for navigation and determination of the type of sea bed. Real advance seems to depend on methods which will improve definition and interpretability really substantially—to the stage of giving some sort of recognizable picture of the undersea scene.

1. Introduction

This paper is directed to experts on the engineering aspects of sonar. Much has been written about the application of sonar to fisheries but most of this has been directed by engineers and fishery scientists to fishermen, to make them aware of the value of acoustic methods and to help them in the use of their machines and in the interpretation of their indications. Here the problem will be examined in another way and some of the problems that arise in fishing mainly in Scottish waters—a world survey would be too ambitious—will be described for the information of engineers.

The Scottish fleet consists of some fifteen distant and middle-water trawlers, ranging from 130 to 180 feet in length and costing around a quarter of a million pounds apiece. There are 150 near-water trawlers, vessels ranging from 70 to 130 feet and costing perhaps £80,000 each. Two groups of smaller vessels are important. There are about 180 vessels in the drifter-seiner class, wooden vessels of 65 to 85 feet valued at about £25,000 and 170 ring-netters and small seiners costing around £10,000 and lying in the range 40–70 feet. In addition there are about 2000 smaller vessels. Thus a large number of vessels, some of which will be shown to have the greatest need for acoustic aids, are small vessels in which space and electrical power tend to be limited and capital for instruments is naturally related to the small earning power per boat.

By far the greatest value at the present day, *in this country*, is tied up in *demersal fish*. These are fish like cod, haddock, whiting and flat fish of various sorts. They are captured by trawl or Danish seine net, and these are nets dragged over the seabed. Thus the most valuable fisheries are for fish lying within 20 ft of the sea bed, and probably most lying on or within 3 ft of it. Certain of these fish are caught by baited lines, particularly halibut in water too deep for trawling. Some of the demersal fish are known to leave the bottom on occasion. Cod and whiting certainly do. This however does not alter the practical fact that their commercial capture by British vessels is dependent entirely on bottom fishing nets.

The other important type of fishing is for *pelagic fish*—herring, pilchard, sprat and mackerel. Though these fish do lie on the bottom for considerable periods, it is when in the upper water that they provide commercial fisheries for British operators. Herring are indeed caught by bottom trawls in large numbers by continental countries, but the market for them caught in this way is not yet available in this country, for various reasons, some practical and some merely traditional.

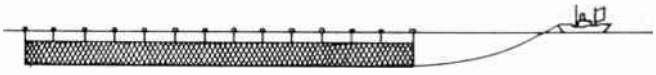


Fig. 1. Drift net—an entangling net 1 mile long by 45 ft deep.

2. Pelagic Fishing

The principal pelagic fishery for British operators is herring—elsewhere sprat, mackerel, menhaden, sardine, and tuna are of great commercial importance. Roughly a million tons of herring are landed annually from the North Sea alone. The herring has a complex pattern of life. Its formations include (a) a sea-bed phase, whose details are not yet understood, (b) a mid-water phase, generally in the form of small tightly-packed shoals, (c) an upper water phase, generally in the form of loose diffuse shoals. The upper water phase occurs generally at night and herring fisheries other than by bottom trawl are pursued therefore at night. The method of capture is by drift net or ring net, while in other countries the purse seine is used, actually a very big version of the ring net.

Figure 1 shows the drift net method. It is required to shoot the nets in a position where herring will become entangled with them. The nets take perhaps three-quarters of an hour to shoot and about three to ten hours to haul. They extend over about a mile. Thus the problem facing the skipper is to select out of a

† Department of Agriculture and Fisheries for Scotland, Marine Laboratory, Aberdeen.

30 or 40 mile square a strip of a mile in length where herring will be swimming in the upper water after his nets are shot. Most Scottish drifters try to do this by echo sounder, seeing the fish in the mid-water phase, and assuming they will rise during the dark hours.

Any good echo sounder will show the mid-water shoals. To recognize them as herring and predict their behaviour is the skipper's vital contribution to the problem, as is the choice of an area to search. Since it takes him about five hours to make one traverse of a 40 mile square his searching power is limited and he compensates for this by knowledge of the last night's catches and by radio contact with other boats, or research vessels.

Drifter operations depend therefore on the detection of mid-water shoals. These shoals are not either evenly or randomly distributed, but are aggregated or concentrated into groups. If the distribution were indeed random, we could expect the paths between them to have a frequency related exponentially to the path length. In practice there is found to be an excess of long paths and an excess of short paths. In the Scottish fisheries it seems to be typical that a group of shoals has a diameter of the order of five miles, and the expected number of shoals detected by echo sounder is about three to five per group.

If the vessels were equipped with horizontal echo rangers capable of detecting shoals at a range of one mile, the total number of shoals observed would be increased. However the chance of detecting a typical group of shoals would only be increased by about 20% (for uni-directional search) as instead of having to intersect a 5 mile circle, the ship could pass as much as one mile to a flank and still make sonar contact. This gain is however very problematical since the echo ranger indications are on the whole less reliable and certainly less interpretable than those of a vertical sounder.

Even if the ranger allowed the detection of the odd lone shoal between groups it is doubtful if this would have any effect other than to divert the drifter from her search for a general area of high fish population.

2.1. *Ring Net Fishing and Purse Seining*

As a complete contrast to drift net fishing, which is an open sea method, and still the largest source of herring, is the inshore ring net fishery conducted by pairs of boats. Here the object is to stalk an individual shoal, sometimes in quite shallow water. Once found, the shoal is encircled by the ring net which is then closed and hauled to one of the boats (Fig. 2). The searching is done quite effectively by echo sounder, but during the encircling movement the vessel is not able to keep the shoal position clearly identified. Here there is great scope for a new device of modest range, and great precision, that will allow the ring net skipper

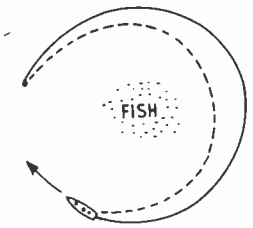


Fig. 2. Encircling nets—ring net 600 ft × 100 ft deep; purse seine 1500 ft × 120 ft.

to have the best possible information about individual shoals. These vessels are small and limited in the size and complexity of the instruments they can carry, and the problem is not simplified by the fact that they are often working among shallow water and rocks which themselves may give rise to complex echoes.

In other countries (for instance Norway) there are large purse seine fisheries. These can be described accurately enough as ring net fisheries on the grand scale, using a very large deep net in deeper water. In such fisheries sonar comes most completely into its own. Attention is being paid to individual large shoals, and the water is deep enough to avoid interfering sea-bed echoes in most cases. In these fisheries both echo sounding and echo ranging are used on a big scale. The impression from reports is that currently available equipment is felt to be satisfactory at present.

2.2. *Pelagic Trawling*

The superficially obvious way to capture fish in mid-water and near the surface is to tow a trawl net at the appropriate depth. This method has found no application in Scottish fisheries, but is gradually developing elsewhere using two boats. The method is not always efficient, perhaps because mid-water fish are able to avoid the net.

In pelagic trawling, net-mounted transducers are finding application for setting the depth of the net to correspond with the depth of the fish.

3. Demersal Fishing

In this commercially vital group of fisheries echo sounding has found great application as a means of navigation, and for the approximate but useful information it can give about the nature of the sea bed. As a useful means of finding fish, however, existing techniques have not proved themselves more than marginally useful, in spite of considerable efforts to devise helpful methods of display.

Since there seems to be no doubt that existing echo sounders can detect fish provided they are about 2 ft or more off the sea bed this conclusion may seem unduly pessimistic, and the situation deserves further examination. Figure 3 shows the otter trawl, the most important demersal fishing gear. This is towed behind the vessel at speeds normally about 2 to 4 knots depending on the power of the towing vessel. The net

is kept open by the otter boards in a horizontal direction. The headline is raised by floats, and sometimes by other means using the forward speed to provide lift. The footrope is weighted to keep it on the bottom, and may be protected by heavy rollers known as "bobbins" to make it roll over the sea floor and so be less likely to catch on obstructions. The weighting and nature of the footrope depends on the type of fish sought, a heavy footrope scraping the bottom is effective for flat fish while a lighter touch is sufficient for haddock, whiting or cod, and is easier to tow.

The trawl is usually towed for two to three hours and so the tactics involve the choice of a strip of bottom about 10 miles in length over which the tow will be made.

Fig. 3. Otter trawl. Towed 5–10 miles. Horizontal gape about 40 ft; vertical gape 5–15 ft.

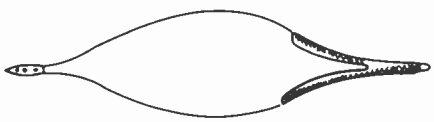


Fig. 4. Danish seine net. Hauled to vessel by winch through 1 mile. Length round net 150 ft, height 10 ft.

The Danish seine net is used only by small vessels, up to about 80 ft in length. The net is not dissimilar in shape to the trawl (Fig. 4). It is however shot and then hauled to the boat using the winch. The distance covered per haul is about 1 mile though the effective distance swept is probably much less since the net is not held open by boards, and closes during the later part of the haul. The time to shoot and haul is about three-quarters of an hour.

3.1. *Distribution of Catches*

Research vessels often take part in comparative fishing experiments in which a series of trawl hauls are made on the same ground. From a large number of such experiments the variance of the $\log_{10}$ of numbers caught per haul is found to be of the following order:

Haddock	0·05	Cod	0·06
Flatfish	0·05	Whiting	0·20

This gives 95% of the catches of, for example, haddock lying within a factor of 3 of the geometric mean. (Because of the mobility of most fish there is not any "fishing out" of the experimental strip during a typical experiment.) If we consider a very large area such as the North Sea, or the Faeroe Plateau, we find naturally a much larger variance in catch of a species. Thus roughly 95% of the catches lie within a factor of 12 of the geometric mean catch. This demonstrates the obvious fact that there are good and bad fishing areas, and that by trawling in various places certain regions can be specified as being more productive of particular species of fish.

The vital question in demersal fish detection is thus how far there are differences in population to be discovered on a practical scale.

Pursuing this problem, the writer tried to discover, by the time-honoured technique of asking the experts, how the variance of catches increased as the size of the area increased. It seems that this could only be answered by unreal simplification. The model presented, at least for the North Sea, was rather of a sea area broken up by variations of the depth and sea-bed structure into discrete "fishing grounds". A "fishing ground" is to be visualized as any region of uniform sea bed and conditions suitable for fishing. It might be merely a strip perhaps five miles by half a mile or a region as much as thirty or forty miles square. Such "fishing grounds" are usually well known and are given distinctive names by the fishermen.

At the present state of knowledge it appears that the variance of catch at any time within a "fishing ground" is about the same size as the variance observed in catches on a selected strip within a ground. It is not therefore, in general, practical to indicate (even after the event) that one selected beat within a fishing ground is the best one to concentrate on. Thus even if an echo sounder gave a reliable indication of demersal fish (which it certainly cannot at present do) there would apparently be nothing useful in this information so far as selecting where to fish within a recognized ground.

This leaves the acoustic techniques to be justified by allowing a choice to be made between two or more adjacent fishing grounds. If, for example, the two grounds differed by a factor of two in their productivity, a matter of perhaps two hauls in each ground would give a good chance of deciding which would be the most profitable one to fish. For an acoustic method to be useful it would have to indicate the catchable population about as accurately as a pair of trial hauls, and it would have to give this indication much more quickly than these hauls, which after all are themselves productive even if not profitable.

So far as the North Sea and the fisheries with which the writer is familiar are concerned, present acoustic techniques are not only inadequate for this critical requirement, but except for whiting and occasionally cod, they give no practical information whatever about fish. This therefore is the design problem. It might

even be realistic to say that there is not yet a problem at all and that North Sea demersal fisheries are not at present suffering from the lack of means of fish detection so far as present fishing methods are concerned. However, it may well be agreed that there is in this situation at least a very real challenge to engineers and scientists, a challenge to find ways of assessing fish populations to the very high degree of accuracy that is needed to provide significant benefits.

After these somewhat pessimistic remarks it must be emphasized that a good echo sounder is an excellent thing to have aboard any fishing boat. It gives an awareness of the character of the sea and the sea bed. On occasions also, for instance, the writer has fished in regions characterized by heavy echo traces which appear to be caused by small fish several fathoms off the sea bed, and the fishing has been good. It seemed certain that the fish caught were not those giving the echo trace, but nevertheless a whole characteristic pattern has appeared—perhaps only of local and seasonal interest—which can be recognized again, or perhaps seen to change from year to year. Thus at least some knowledge of the sea is gathered in and over the years facts about echo traces are accumulating. Some of these facts we can already explain in terms of fish. The rest we might hope in time to explain in terms of fish.

4. Future Developments

There is a natural desire on the part of all associated with fisheries to increase the efficiency of fish catching, to reduce the heavy work and long hours still associated with the profitable pursuit of fish. To this end continuous thought is being given to the development of better gear and better methods. Our knowledge of fish behaviour and shoaling habits, and of the way fish react to the approach of fishing gear is not adequate as a background for the design of new gear, and most strenuous efforts are being made to fill these fundamental gaps in human knowledge. The difficulty is obvious—all the events under examination go on beneath the sea and out of reach of ordinary observations. In shallow water skin divers have contributed much knowledge about the behaviour of fish in daylight. In deeper water and at night, flash cameras attached to the net have made some contribution. In addition to this, ordinary echo sounders can be applied to give information, though they are severely limited by the difficulty of interpreting their indications with certainty.

Hence the field is wide open for more and more precise acoustic methods—which can have much greater than optical range, and can work in darkness. One development of this kind has been described at this symposium‡ and represents the kind of progress that may prove invaluable. At a much simpler level of engineering there seems to be application for echo sounders of much higher resolving power. This can most easily be achieved by using frequencies in the region of ½ to 1 Mc/s and very narrow beam widths. Equipment of normal power for use from vessels would find applications, and miniaturized equipment suitable for attachment to the nets could also give valuable information.

These net-mounted equipments can be either self-contained recorders, or transmit their information to the ship by cable or a secondary acoustic link. It must be remembered that electric cable from ship to net is always inconvenient and should be avoided where possible. The vital need now is for better definition, however achieved, so that we can distinguish between, say, a fish 2 ft long and a fish 1 ft long, and between either of these and a small shoal of sprats or shrimps.

The principal need for miniaturized equipment is for research, and development of fishing gear. When better gear exists it will become apparent whether such specialized acoustic aids are necessary and useful in their commercial application.

So far as ordinary ship-mounted equipments are concerned, even the pelagic fisheries would benefit from improved definition, which would make the display easier to interpret, and allow tactics to be based on more significant indications.

Manuscript first received by the Institution on 28th June 1962 (Paper No. 794 SS 11).

POINTS FROM THE DISCUSSION

Dr. H. Maass†: When the echo sounder installed on board a fishing craft shows the presence of shoals of fish in a certain depth it has been for long a problem to adjust the net being towed several 100 metres behind the ship to the actual depth in which the fish are shown in the echo sounder recorder. This is done usually by altering the ship's speed.

In order to render this method effective the Atlas Echo Sounding Equipment "Netzsonde" was developed (Fig. A). In this system a transducer is fastened on top of the opening of the net sounding vertically and controlling hereby the mouth of the net and its position with regard to the bottom. The transducer is connected by a cable to a second

† Atlas Werke A.G., Bremen.

‡ V. G. Welsby and J. R. Dunn, "A high-resolution electronic sector-scanning sonar" (Sonar Systems Symposium paper).

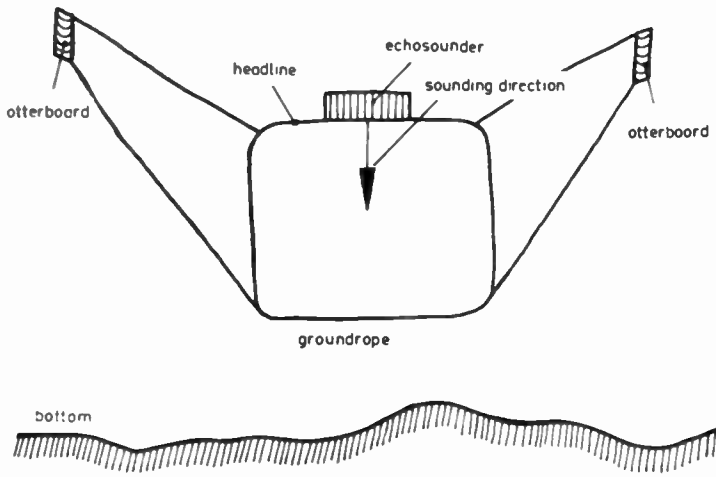


Fig. A.

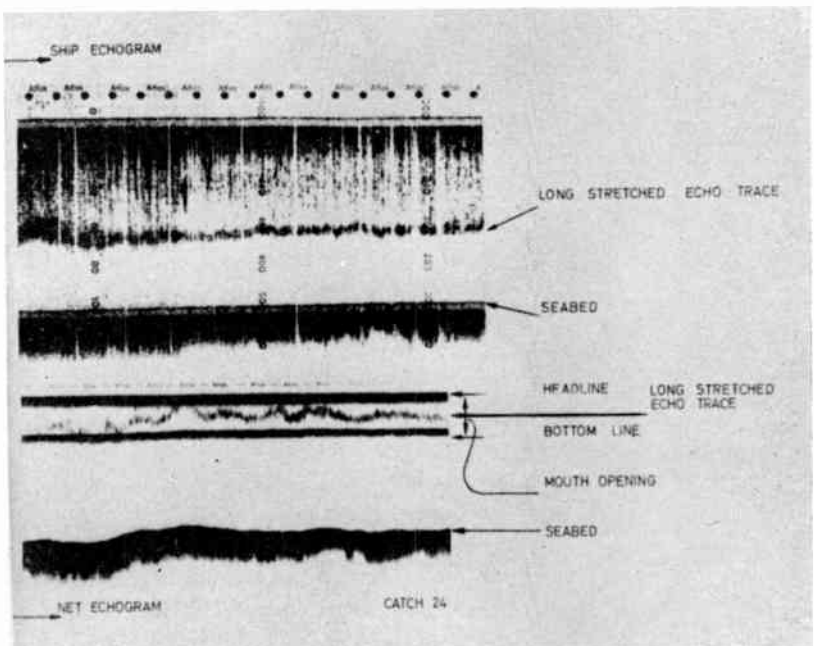


Fig. B.

echo sounding equipment installed on board the fishing craft (Fig. B).

In using this method it has become possible to correct the ship's speed according to the "Netzsonde" tracings in such a way that in the end the opening of the net is brought to a height on top of the bottom which is identical with the height of the shoal of fish previously indicated on the recorder of the ship's depth sounder (Fig. C). One is hereby in the position to control the fish going into the net. This method will considerably improve the efficiency of pelagic fishing.

Besides this the "Netzsonde" equipment offers in trawl fishery the advantage that in areas with rough bottom conditions the exact position and height of the net with regard to the bottom can be continuously observed and the net can thereby be protected from damage.

The author (*in reply*): I agree that the headline oscillator or "Netzsonde" shows promise as an aid to pelagic trawling, and also, as I mention in the paper, as an aid to studies of fish behaviour in relation to any type of trawl.

Mr. M. Schulkin†: Have you considered the possibility of using passive or listening sonar for commercial fisheries?

Advantages over active sonar are:

(1) Active sonar pings might be scaring the fish.

(2) It would eliminate the problem of bottom return and interference.

(3) Specific fish might be recognizable by their individual sounds. If the fish which are sought do not make sounds, then perhaps the fish which are their feed-fish do.

† Avco Corporation, Washington D.C.

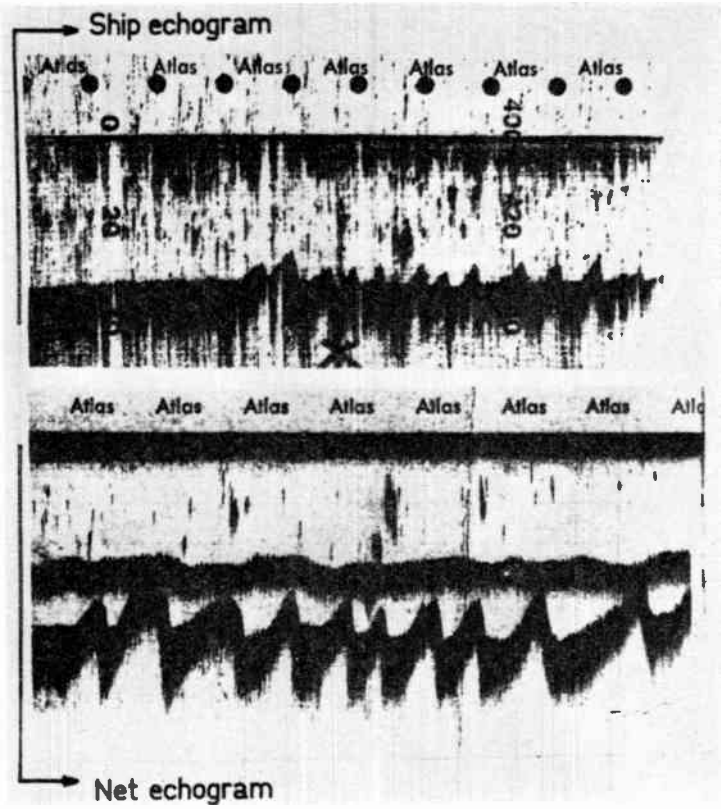


Fig. C.

The author (*in reply*): Detecting fish by the sounds they make is a different technique altogether from active sonar. What limited experiments we have carried out have been discouraging and offer no prospects of early success. This is not however to deny that better techniques might show promise in the future for some fisheries.

Dr. D. H. Cushing†: Echo sounders are used intensively by trawlermen working in the Barents Sea, Iceland and on the Grand Banks. Purse seine fishermen use them for finding herring in the Norwegian Sea and off Iceland; indeed both these fisheries are asdic fisheries. In the North Sea all herring fishermen use echo sounders to find fish. It is true though that trawlermen in the North Sea do not use echo sounders so directly. They use them to describe the different characters of the bottom.

I disagree with Mr. Craig profoundly on the real use of echo sounders to fishermen, apart from finding the depth. There is, of course, a variance in catches, but an echo sounder in the Barents Sea allows a trawlerman to find fish more quickly and to find when to haul his trawl. In other words an echo sounder allows a good fisherman to differentiate himself from a poor one. I am not surprised that his correlation between trawl catches and the paper record failed; this is because the echo sounder records the range of fish just above the bottom and not their depth. The correct correlation is between trawl catch and those traces which are known to be below the headline height. Many traces which appear to be below the headline height are in fact above it.

The author (*in reply*): Dr. Cushing emphasizes the value of echo sounders in the, principally cod, fisheries of the Arctic. These fisheries I excluded from my discussion because I had no personal experience of them.

In my paper I tried to explain briefly that fish aggregations had to be on a suitable scale to make echo detection tactically useful, and it is clear that the Arctic cod population satisfies the required conditions better than the mixed population of the nearer water fisheries.

We seem to agree completely about the applications to herring fisheries.

Regarding the attempt I mentioned to correlate trawl catches with paper records, I am no more surprised than Dr. Cushing that this failed. Long experience has taught both of us that the paper record is completely useless as an indication of trawlable fish in the near waters. *I felt it necessary however to emphasize this point to the present audience.* We have also attempted to correlate the indications of commercial 'scope display with trawl catches, in North Sea and Faroe waters. These attempts have also failed.

My general conclusion remains that we could, in general, benefit from more refined equipment, but that even then there are many fisheries which could not obtain tactical advantage from it.

† Ministry of Agriculture, Fisheries and Food, Fisheries Laboratory, Lowestoft.

U.D.C. 621.9

A Survey of the Techniques Evolved for the Measurement of Position in Numerically Controlled Machine Tools

By

R. BELL, M.Sc.(Tech.)†

Presented at the Symposium on "Recent Developments in Industrial Electronics" in London on 2nd–4th April 1962.

Summary: The techniques described are divided into analogue and digital methods and the two categories are further subdivided where it is considered logical. The inherent sources of error in machine tools are briefly discussed and the basic types of control system are explained before entering into the detailed discussion of the transducers.

1. Introduction

A numerically-controlled machine tool is an automatically-controlled machine, where the information which determines operating conditions is derived from magnetic tape, paper tape or punched cards (manual controls are often also available). In the extreme case the functions of the operator are restricted to initial setting-up of the work-piece and the removal of the end product.

The principal elements of an axis of controlled movement are shown in schematic form in Fig. 1. From this diagram, it will be readily appreciated that the resultant workpiece accuracy can never be better than the accuracy of the displacement transducer employed.

The first numerically-controlled milling machine was developed at the Massachusetts Institute of Technology in 1951.[1] This machine employed digital displacement transducers and a measuring accuracy of $\pm 5 \times 10^{-4}$ inches was attained.

The acceptance of numerical control had brought definite benefits but has also posed a new set of design problems for the machine tool designer. One of these problems is that of automatic displacement measurement. To meet the combined requirements of accuracy, resolution and compatibility with the closed loop operation of machine tool drives, it has been found necessary to evolve a range of techniques which form a marked departure from previously accepted practice. This paper will survey the many fascinating and extremely ingenious techniques evolved to overcome the problem of displacement measurement.

2. The Range of Application of Numerical Control

The range of application of numerical control to machine tools is very wide and of course the accuracy of the displacement transducer employed is a function of the particular process which is being controlled. Perhaps the most stringent example of accurate positioning is the jig borer; in this case a positional accuracy of $\pm 10^{-4}$ in is often required. The accuracy of positional control may have to extend over a traverse of up to 30 in. in any axis. A typical example of this class of machine tool is shown in Fig. 2; in this system the input command is derived from punched cards or manually set dials. The quoted time taken for the system to align itself to a given set of co-ordinates is not greater than 30 seconds. An interesting contrast to the jig borer is the horizontal boring machine shown in Fig. 3, where the required accuracy need not be better than 1×10^{-3} in, although the same transducer is employed.

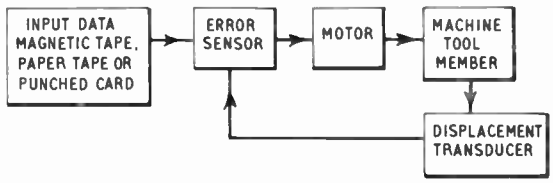


Fig. 1. The general form of the basic closed loop for one axis of a numerically-controlled machine tool.

The application of numerical control to contour milling calls for both accuracy and speed of response, the most striking example of this class of machine being the "Green Linnet" vertical miller,[2] shown in Fig. 4. This machine is still one of the best examples of a machine tool design specifically for numerical control. The cutter is capable of 25 ft traverse in the longitudinal axis and 7 ft traverse in the vertical axis. the positional accuracy is stated to be $\pm 2 \times 10^{-3}$ in over 18 in and $\pm 5 \times 10^{-3}$ in over the full range of the machine.

The precision control of angular rotation is illustrated by reference to the rotary table shown in Fig. 5. The table diameter is 30 in and accuracy of setting is ± 3 seconds of arc. Here manual and automatic input commands are available.

The measurement of the work-piece calls for the employment of accurate, high resolution position transducers. An example of a machine developed solely for inspection measurements is shown in Fig. 6. This machine measures in increments of 5×10^{-4} in and the specified accuracy is $\pm 10^{-3}$ in over 24 in.

† Faculty of Technology, University of Manchester.

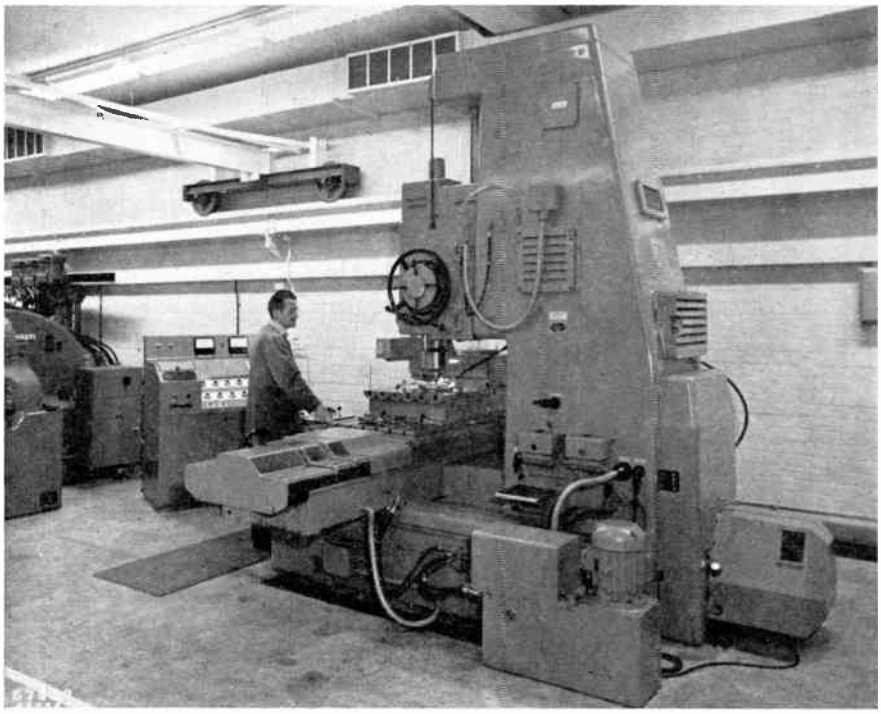


Fig. 2. Jig borer.

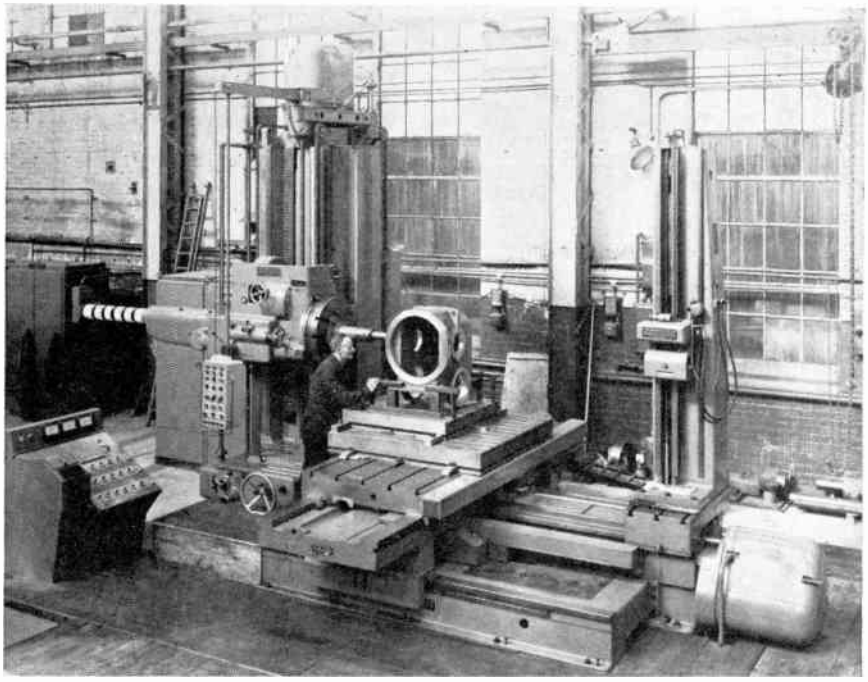


Fig. 3. Horizontal boring machine.

3. Factors Influencing the Accuracy of the Finished Work-piece

It is necessary to examine briefly the inherent limitations of the machine tool and the cutting process which cause the finished work-piece to be less accurate than the displacement measuring systems used to determine the movements of the machine tool elements. A major limitation is the static deflection of the machine tool members which will change as the distribution of the masses of the elements is altered during the machining operation. The magnitude of these errors and the cost of the machine are closely aligned. An example of static errors for a knee-type milling machine is quoted in Fig. 7. It

should be noted that this class of machine does not give the maximum accuracy possible with machine tools.

The fluctuating cutting forces[3] will add to the static errors mentioned above and this will be particularly evident if a harmonic of the cutting force can excite a mechanical resonance of the machine structure. The magnitude and spectrum of the cutting forces will vary with the material being machined. In the case of light alloys, the cutter fluctuations are characterized by a

Fig. 4. "Green Linnett" vertical miller.

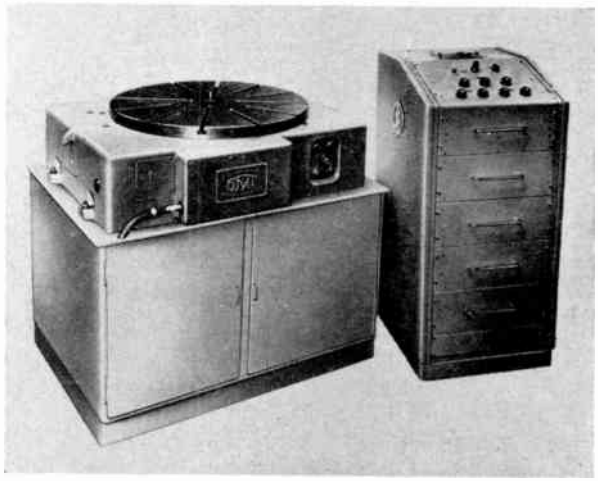


Fig. 5. Precision controlled rotary table.

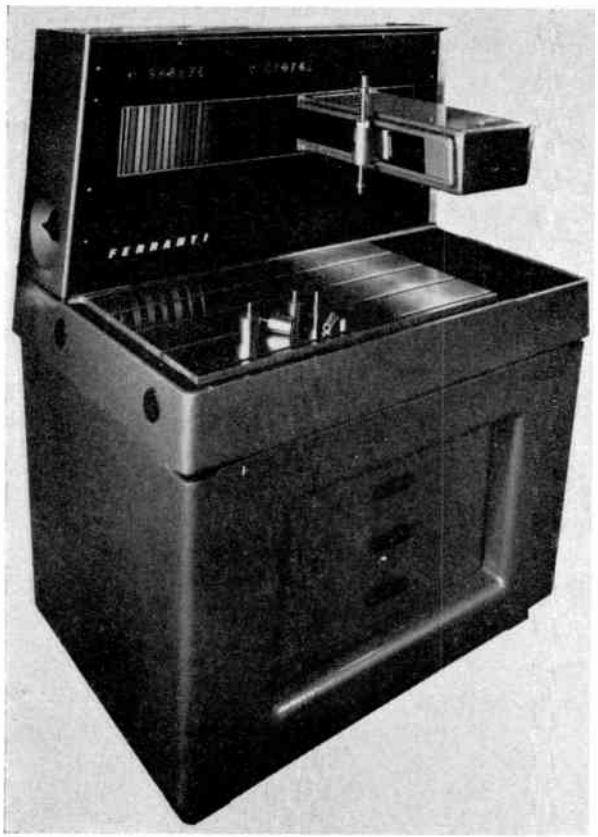


Fig. 6. Inspection machine.

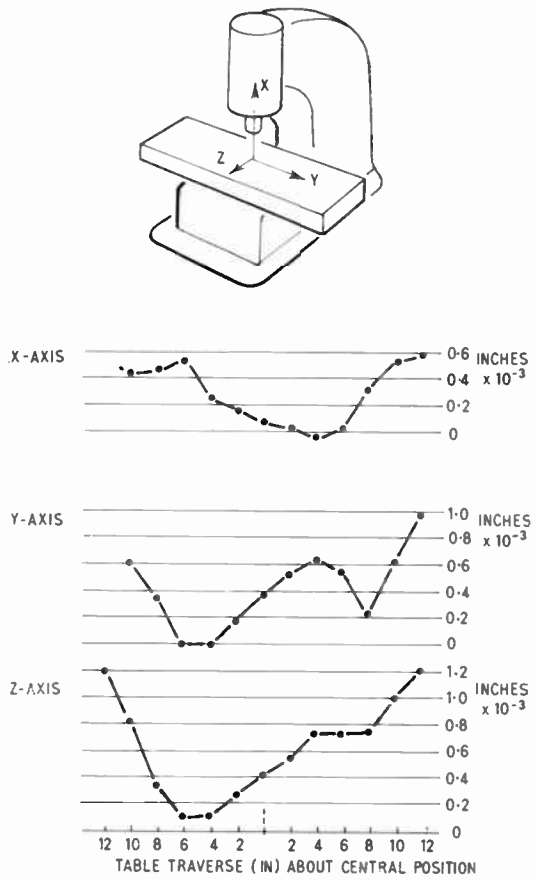


Fig. 7. Unloaded static errors for a knee-type milling machine

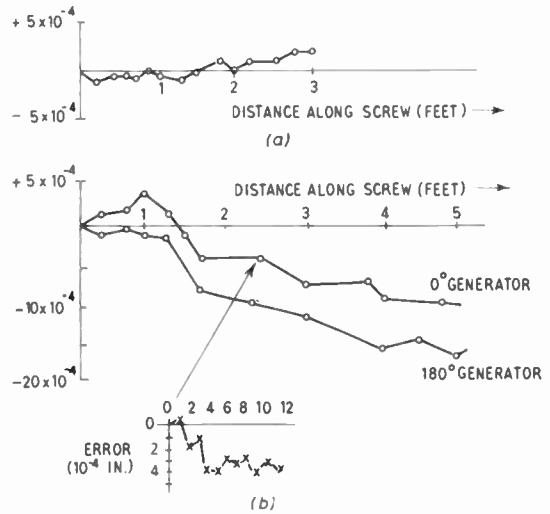


Fig. 8. Lead-screw errors.

(*a*) Cumulative error in pitch of a master lead-screw.

(*b*) Cumulative error on a quality screw

Standard accuracy of a lead-screw $\begin{matrix}+0{\cdot}002\\-0{\cdot}003\end{matrix}$ in any 8 ft.

Special accuracy $\begin{matrix}+0{\cdot}006\\-0{\cdot}001\end{matrix}$ in. in any foot.
$\begin{matrix}+0{\cdot}001\\-0{\cdot}002\end{matrix}$ in. in any length.

high fundamental frequency (of the order of 100–150 c/s) and low peak amplitude (say 500 lb). High tensile steels present greater forces (5000 lb peak force), with fundamental frequencies of the order of 10 c/s.

Under severe cutting conditions, the wear of the cutters can contribute errors in contour machining, but it is not easy to correlate work-piece errors and cutter wear.

Non-linearities in the machine tool drives, i.e. backlash and stiction, will further aggravate the realization of maximum accuracy. However, the improved mechanical design of drive members can minimize backlash and the development of hydrostatic and roller bearings has provided a means of eliminating drive stiction.

One source of error sometimes overlooked is the misalignment present due to departure of the axes of movement from the desired orthogonal alignment.

The machine tool drive element which merits individual mention is the lead-screw. The lead-screw and nut is the most commonly used method to translate motor shaft rotation into linear table movement and is also often used in conjunction with a rotary transducer to provide the output displacement information. The lead-screw is subject to two forms of error, i.e. the progressive error over its complete length and cyclic errors due to imperfections in the thread form (see Fig. 8). It is possible to compensate for the progressive errors and reference will be made to these techniques later in the paper. Cyclic errors, however, are rarely compensated as this error term is more susceptible to wear effects than is the progressive error. The introduction of "recirculating ball" nuts will minimize wear and also eliminate backlash. Any end-play in the lead-screw bearings will provide a source of error outside the scope of the measuring system. The error shown in Fig. 8 can be reduced by the use of ground lead-screws and "recirculating ball" nuts. The accuracy quoted for one machine employing a ball lead-screw, with 0·333 lead, hardened and ground form, is within 0·0005 in. in any twelve inches and not greater than 0·003 in of the whole traverse.

Finally, the effect of temperature must be considered and an approach to this problem is to employ displacement transducers made of materials possessing similar thermal coefficients of expansion to that of cast iron or steel, but no simple solution is readily available.

4. The Role of Position Transducers in the Numerically-Controlled Machine Tool

A very large number of numerical control systems for machine tools have been evolved. In the wide range of system design, it is possible to bring out certain salient points which are of interest. The first point is that most machine tool applications call for displacement transducers capable of accuracies of the order of 10^{-3}–10^{-4} in and resolution of the order of 1 part in 10^4 to 1 part in 10^6. This specification is completely outside the scope of any one analogue scale, an analogue scale being defined as scale giving a continuous measure over its working range. The linearity required to give the performance stated above is virtually impossible to obtain, particularly when one considers the performance of the circuitry coupled with the basic transducer. It is possible to devise transducers with accurate resolution to 1 part in 10^3 or better. Thus if the basic scale length is, say 10^{-1} in and this scale can be repeated with the required precision, then obviously one can devise round this class of element a control system capable of the necessary accuracy. This class of transducer can be produced in two forms, the repetition of scale can be done by repeated manufacture of the basic scale or by using an element in a rotary form coupled to a lead-screw of the required accuracy and pitch.

The alternative to the analogue transducer is the digital transducer which gives a quantized measure of position. Here the necessary requirement is a pattern produced with sufficient accuracy to give the necessary basis for accurate signal production. A digital transducer can be made in two forms, the incremental element or the coded element. In the former the output of the transducer consists of a series of pulses,

each representing one unit of displacement. The coded transducer gives a pattern of pulses through a multi-channel output system. The output of this class of transducer is absolute within the length of the pattern. The incremental transducers require some means of identifying the sense of the movement measured; the coded transducer inherently monitors the sense of the movement. Both these types of transducers can be employed in either linear or rotary form.

In a continuous system, the machine tool element is required to keep in step with the input command throughout the whole machining operation. In the "point-to-point" or "co-ordinate setting" class of system, the machine tool is required to take up a series of commanded positions; the elements need not follow any specific locus in between points, the ultimate co-ordinate accuracy being the only significant parameter. This essential difference explains why some transducers employed on "point-to-point" systems cannot be used in continuous path systems. The common characteristics of these transducers is that they are unable to give a fast measure of displacement.† This range of transducers, however, will allow the machine to arrive at the required co-ordinates more quickly than the alternative manual manipulation of gauges or scales. On one system,[5] it is possible to set the co-ordinate *automatically* to several thousandths of an inch and the operator uses a microscope to arrive at the accurate setting. This setting is then recorded magnetically and this dimension can be repeated from the recording. This system is not a conventional co-ordinate setting system.

One feature of the "point-to-point" system worth mentioning is the fact that the machine can traverse at maximum speed between required settings and so the fine measure of position need only be employed when the element is very close to its ultimate position.

In a controlled machine tool system, the choice of the input data form and the nature of the position transducer are closely related. The input data will consist of either a sequence of incremental position instructions or a series of statements of absolute position with respect to a given datum point. A combination of these two concepts is possible, for example, a system where the input data consist of a digital measure of the basic analogue scale is commercially available.‡

The resultant motion of the machine tool element is continuous and therefore there must be somewhere in the control system either a digital-to-analogue conversion or an analogue-to-digital conversion. If an analogue transducer is employed, then the input data must be converted into a compatible analogue form. The use of a digital transducer of the incremental type will require a digital-to-analogue converter to give an analogue error signal. If a coded digital transducer is employed, then it may be necessary to have an additional system element which makes the input and feedback information compatible.

5. Analogue Position Transducers

The use of a repeated analogue scale in a position control system carries with it the potential danger of signal ambiguity if the linear range of the transducer is exceeded. The system designer has three possible methods to use to avoid this fault. Firstly, the system can be designed to operate continuously within the linear range of the transducers with a fault sensing system to cover emergencies. Secondly, several analogue scales can be employed of different sensitivities. The final choice is the use of a secondary digital scale to count the integer number of analogue scales to be traversed. If the transducer is to be used in an absolute system, then only the second or third alternatives can be employed. The third method is a hybrid analogue-digital system, but this will be classified here as an analogue system. It should be noted that the first of these three techniques would satisfy the requirements of an incremental data input system. In this type of control system the maximum instantaneous and short-term displacement errors are controlled and therefore only one scale is significant.

The second and third possibilities would be applied in control systems, where the input data is expressed in absolute form, i.e. as a measure of the displacement of the machine member from an arbitrary datum point.

A very wide range of analogue techniques have been developed and will be classified as follows:

(i) Electromagnetic analogue transducers
(ii) Resistive analogue transducers
(iii) Capacitive analogue transducers
(iv) Optical analogue transducers.

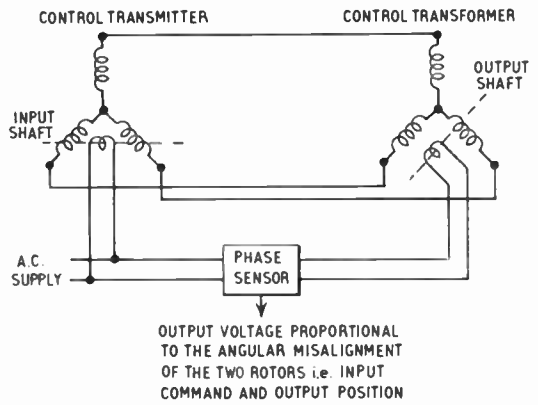
It is interesting to note that before the advent of electronic techniques mechanical gauges were employed to give an accurate measure of displacement. Indeed, more than one system has been developed for point-to-point control using mechanical gauges or indexing.[6, 7, 31]

5.1. *Electromagnetic Analogue Position Transducers*

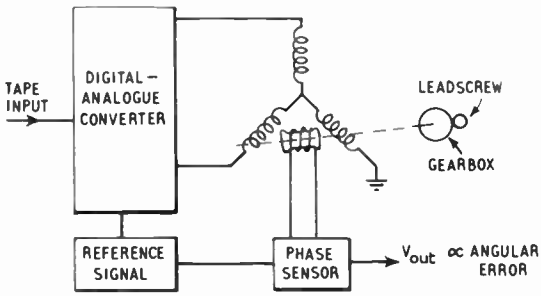
The most commonly used transducers for position measurement in machine tool control are those belonging to the synchro range. The absence of wear effects and continuous resolution of the rotary electromagnetic element make them superior to the resistive

† An example of this type of control is given in Fig. 17 (*a*), where the Pratt & Whitney system is shown in schematic outline.

‡ In the A.E.I. Numeritol control system, the analogue scale is 0·1024 in pitch to allow a simple digital-to-analogue conversion.



(*a*) A synchro position transducer system with mechanical input.



(*b*) Synchro position transducer system with a digital input.

Fig. 9. Synchro transducers.

potentiometer. Synchros display very good linearity; for example, an error of ± 7 min of arc is comparable to a very high quality potentiometer of 0·03% accuracy. Comprehensive discussions of the characteristics of synchros are available in the literature[8, 9, 10]; only the adaptation of these techniques to machine tool control will be discussed in this paper.

The synchro transmitter and the synchro transformer have three equi-phased stator windings and one rotor winding. Normally, a position sensing system consists of a synchro transmitter and a synchro transformer as shown in Fig. 9(*a*). The rotor of the transmitter is normally energized with a 400 or 1100 c/s supply, and a misalignment between the two rotors can be measured by sampling the output of a phase-sensitive rectifier employing the transmitter rotor signal as reference and the rotor signal of the synchro transformer as input.†

Input information is normally fed in by rotating the rotor of the synchro transmitter. In a numerical control system, this mode of input insertion would create certain problems. Systems have been evolved therefore, where the synchro transmitter is replaced by a special-purpose digital-to-analogue converter, designed to deliver the appropriate signals to the stator windings of the synchro transformer. This approach is illustrated in Fig. 9(*b*) and has been described by Cooney and Ledgerwood. It is possible, however, to specify either transmitter shaft rotation or voltage selection for the stator windings of the synchro transformer. A further variant on the basic method is employed in the numerical control system discussed in reference 11.

In the G.E.C. system, the digital-to-analogue converter and synchro transformer combination is employed.[7] Three channels of synchro feedback are employed with measuring ranges of 10, 1 and 0·1 in. Interpolation to one part in a hundred on the least significant digit gives a measuring system whose range is from 0 to 99·999 in. To achieve this measuring system, the synchros are connected to the table leadscrew via a gearbox. The position servo is normally driven by the error signal from the least significant digit synchro, unless as in point-to-point applications, top traverse speed is required. In this case, the most sensitive synchro signal is ignored until the commanded position is approached.

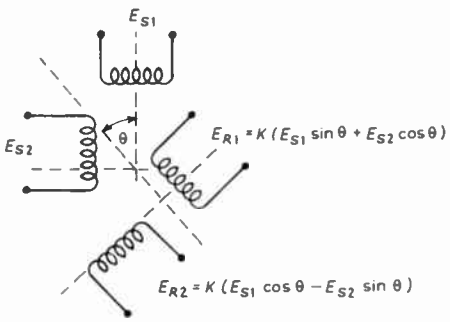
It is possible to employ synchros capable of delivering a restricted amount of torque but the accuracy of this element falls below that of the system described above. It has been stated[8] that the error of a typical multi-speed synchro system could be as good as (+1·5′, −2′), the error of a torque synchro can be of the order of ± 40′ of arc.[12]

An interesting development of the synchro was employed in the Sperry-Orcutt dividing engine, which consisted of a 36-pole synchro, giving positioning to ± 2 seconds of arc.

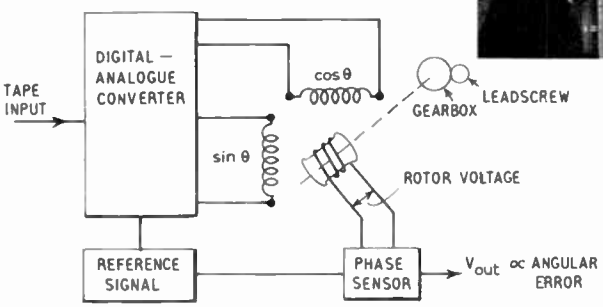
The resolver synchro, as its name suggests, is an electromagnetic device for resolving a voltage into two components which are 90 deg out of phase. The basic element is shown in schematic form in Fig. 10(*a*). When used as an error-sensing element in a position control system, the function of the resolver is simpler than its use as a computing element. The commanded position has to be expressed as two voltages which are the reference voltage multiplied by the sine and cosine of the appropriate shaft angle. When the resolver rotor takes up the commanded angle the output from the phase sensitive rectifier is zero. The accuracy of the resolver can be improved by employing additional windings which are provided by the feedback resolver. For a description of these techniques and their significance, see references 8 and 13.

One of the more significant developments in machine tool measuring systems is the extension of the resolver concept into a linearized form called the "Inductosyn".[6, 8, 14, 15] The Inductosyn elements (see Fig. 10(*c*) and (*d*)) consist of sets of thin, flat conductor windings plated on glass. The moving member

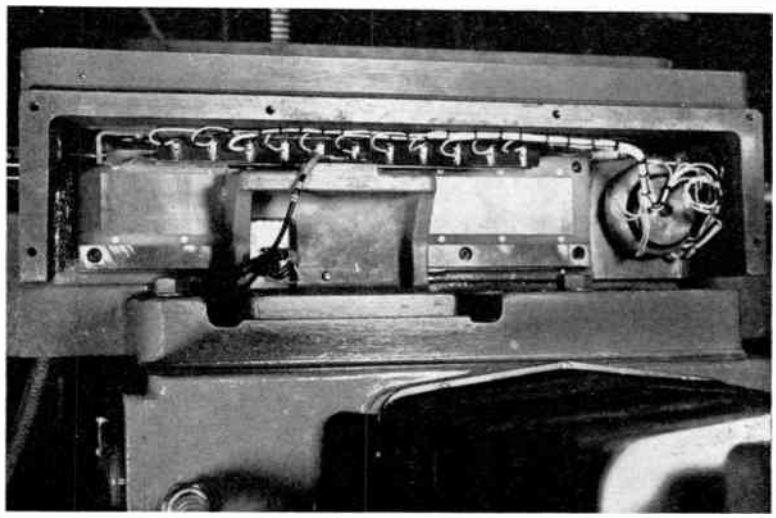
† Control transformer rotor signal $\propto \sin(\theta_t - \theta_o)$. $\propto \theta_t - \theta_o$ for smaller angles.



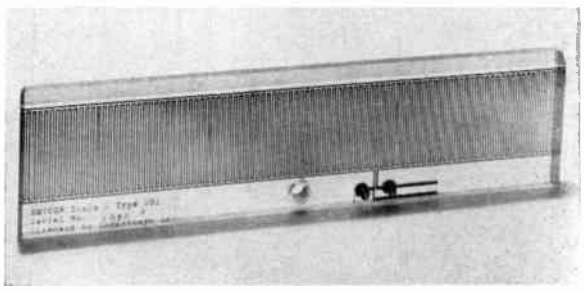
(*a*) The winding configuration on a synchro resolver.



(*b*) The synchro resolver as a position transducer with digital input.

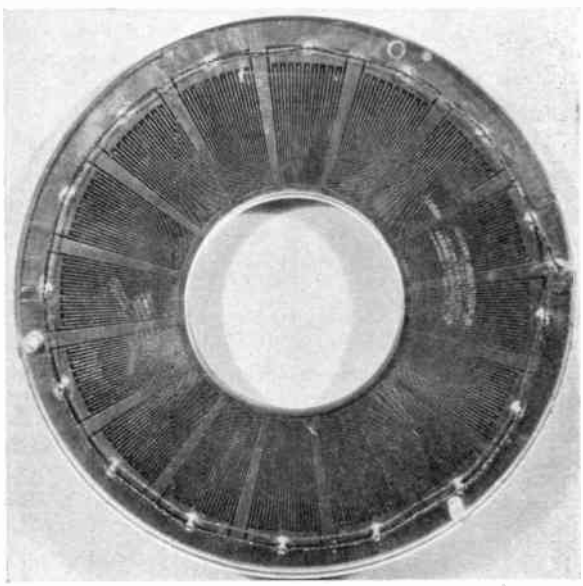


(*c*) Linear Inductosyn *in situ*.



(*d*) Inductosyn scale.

(*e*) Rotary Inductosyn *in situ* in machine shown in Fig. 5.



(*f*) Rotary Inductosyn scale.

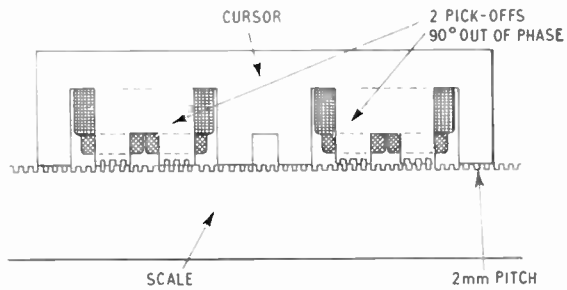
Fig. 10. Resolver transducers.

attached to the machine element has one rectangular winding plated on it. The pitch of the winding is 0·1 in and the total length of this member is 10 in. This section is directly analogous to the resolver rotor winding as shown in Fig. 10(*b*). The equivalent of the two stator windings shown in Fig. 10(*b*) are plated on to the shorter member of the Inductosyn pair. These two windings are 90 deg out of phase. If these two windings are energized with signals whose amplitudes are proportional to the sine and cosine of a given electrical angle, the servo will move until the moving member of the Inductosyn pair takes up the position within the one winding pitch commanded by the voltages supplied to the "stator pair". The working gap between the two glass plates can vary between 5 and 15×10^{-3} in and the range of the scale can be lengthened by joining up a series of 10 in members (with 4×10^{-3} in separation). The linear Inductosyn can measure to an accuracy of 10^{-4} in with a sensitivity of 10^{-5} in and a repeatability of $2{\cdot}5 \times 10^{-5}$ in. In its rotary form the inductors are made with 3, 7 or 12 in diameters. The 3-in discs give angular accuracy of ± 15 seconds of arc (for a 144 pole pattern), the 7-in disc (360 pole pattern) will measure ± 2 seconds and the 12-in disc will give ± 1 second. When comparing this performance to a resolver, it must be borne in mind that a resolver is a 2-pole device and 5 deg angular rotation of a 3-in disc will be equivalent to one revolution of a normal resolver. A view of a rotary Inductosyn, *in situ*, is shown in Fig. 10(*d*).

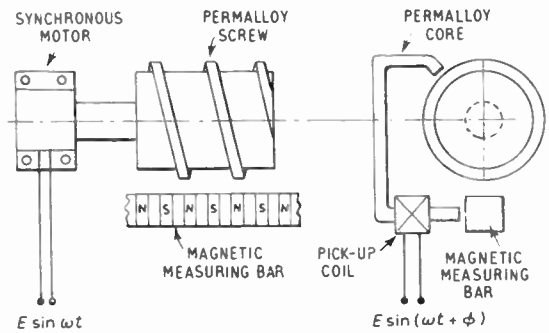
A recent development in the design of the Inductosyn elements is the production of a compound pattern which allows the use of coarse/fine systems without additional resolvers geared to the lead-screw. In order to avoid the problem of differential expansion of the glass and work-piece material, Inductosyn elements are made in a form where a conductor of copper is plated on to a base plate of any of the common metals via a $2{\cdot}5 \times 10^{-3}$ in insulating layer.[16]

An interesting example of a multi-speed control has been described using resolvers for co-ordinate setting equipment.[17, 18] Five resolvers are used, the most sensitive being a linear Inductosyn. The fine resolver (Inductosyn) measures in cycles of 0·1 in, the resolvers then span distances of 1, 10, 100 and 1000 in per revolution. This makes an interesting comparison with the E.M.I. multi-speed system, where the coarsest measuring device is so constructed to make error compensation possible.

The E.M.I. system uses an autotransformer as its coarse measuring member.[19, 20] The taps of the autotransformer are operated on a make-before-break principle and so there is no loss of signal and as the taps can be varied along the winding, it will be seen that it is possible to use the movement of tap



(*a*) The S.E.A. magnetic transducer.



(*b*) Gisholt "Factrol" magnetic transducer.

Fig. 11. Analogue magnetic transducers.
(Courtesy *Engineers' Digest*.)

position as a technique to correct for cumulative errors in the lead-screw.

Several other electromagnetic transducers have been developed to give analogue signals. One common characteristic of two of these transducers is to produce a castellated metal surface and to count these castellations to give a coarse measure of position and then to employ a technique to generate alternating voltages proportional to the sine and cosine of the electrical angle within the magnetic pattern. The first method, as developed by the Société d'Electronique & d'Automatisme,[21] employs surface castellations of 2 mm pitch. A count of this pattern gives a coarse scale of 2 mm. To obtain the analogue interpolation between integer pitches, two U-shaped "reading heads" are effectively employed (see Fig. 11(*a*)). The head is energized by an a.c. signal and two signals are obtained from the limbs of the head. The castellations of the two limbs are out of phase by a quarter of a pitch with respect to the main scale. It is claimed that the combination of these two signals allows analogue interpolation of the 2 mm pitch to 0·01 mm.

The same basic technique is employed in Gisholt "Factrol" transducer but a different physical form is employed.[22] A stationary analogue signal is picked-off from a magnetic circuit consisting of a magnetic

measuring bar and rotating permalloy lead-screw (see Fig. 11(*b*)). With no relative displacement between bar and screw, the signal from the pick-off will bear a fixed phase relationship to the a.c. drive for the screw. Any displacement between screw and bar will disturb this relationship and allow a displacement signal to be obtained. This technique is said to give accuracy of the order of $0{\cdot}5 \times 10^{-3}$ in.

The last electromagnetic analogue device to be described is the "Nultrax" helical differential transformer. This transformer consists of a long steel rod which has a bifilar coil embedded into its surface, a similar coil is set into a sleeve.[24] Typical dimensions would be $\frac{11}{16}$ in rod diameter, coil lead $\frac{1}{10}$ in, rod length 16 in. Movement between the sleeve and the rod gives two nulls every cycle if the rod is fed by an a.c. signal and pick-off is taken from the sleeve. The interpolation achieved by relying on the form of the voltage pattern obtained at this sleeve will not yield high accuracy and so a phase sensitive system is employed and the rod is rotated. The system is almost analogous to a resolver with, however, one substantial difference, that the rod has to be moved by its own servo. In this system, the coarse scales are provided by three potentiometers. The Nultrax transducer will gives accuracies of the order of 10^{-4} in and has been made in lengths varying from 10 to 145 in. The sleeve contains 50 to 100 turns and thus a considerable integration of error in coil form is attained.

5.2. *Resistive Analogue Transducers*

Potentiometers do not lend themselves readily to high resolution applications. Where they have been employed, their function has been as a coarse measuring element. One exception to this statement is found in the Ekco system[25, 26]. Here the system relies on three coarse potentiometers to give 0–9 for the digits 99–9. The resolutions from 0·000–0·100 is achieved on the fourth potentiometer. Resolution of 2×10^{-4} in is achieved using this technique.

In the E.M.I. 100B system, a ten turn helical potentiometer is employed to give the coarse measure of 100 in. As described earlier, the coarse scales of Nultrax transducer systems utilize potentiometers, but it can be generally concluded that potentiometers are not ideally suited to this field of application.

In one commercially-available positioning system, the limitations of the potentiometers are quoted as a limiting factor to accuracy.[27] There is no reference in the literature to the use of linear resistive transducers.

5.3. *Capacitive Analogue Transducers*

A recent development in this country[28] employs a capacitive pick-off from a cascaded voltage transformer. The schematic form of the transducer is shown in Fig. 12(*a*). The long member of the transducer is a segmented rod. The segments of the rod are connected to tappings on a voltage dividing transformer but are electrically insulated. The use of a tapped transformer enables equal voltage drops to be obtained across any two adjacent stators. If a small pick-off electrode were employed, a discontinuous measure of position would result. The use of a pick-off electrode of equal length to the stator segments allows a continuous measure of position to be achieved. An accuracy of $\pm 10^{-4}$ in over 10 in has been attained. A typical segment length is of 0·5 in, but this is not a fixed parameter.

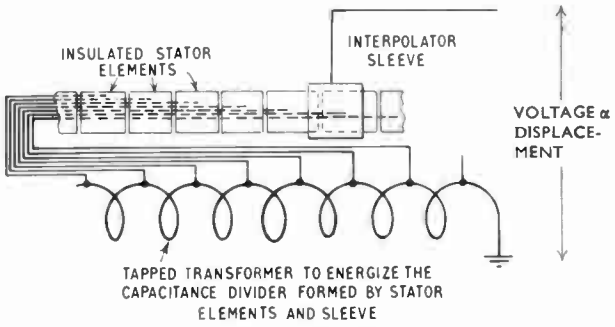


Fig. 12. Analogue capacitive transducers.

(*a*) The Reilly absolute analogue capacitive transducer.

The first capacitive analogue displacement transducer in this country was the A.E.I. Helixyn.[29, 30] This transducer shows a certain similarity to the Nultrax described earlier. The physical differences are first, the pick-off is obtained as a capacitively-coupled signal from the rod and the coils that are wound are three-start, not bifilar. The difference in the operating principle is more marked; here the Helixyn is analogous to the Inductosyn; if one of the three windings is earthed, the sleeve can be energized by sine and cosine signals as in the case of the general resolver system and a null signal can be obtained from the bar (see Fig. 12(*b*)). The air-gap is 0·25 in and the pitch of the analogue scale is 0·1024 (this choice of scale simplifies the digital-to-analogue conversion at the input when binary information is read from the tape). The Helixyn has the merit that it readily lends itself to the insertion of error correction. The rotation of the sleeve by an arm riding on a sine bar adds a predetermined correction.

The third transducer in this class is the Telecomputing Corporation's electrostatic disc transducer.[31] This device (see Fig. 12(*c*)) is basically two concentric glass discs with metallic patterns deposited on them. The stationary disc has three concentric circles etched on it and in the two spaces between these circles are etched two sinusoidal patterns. The moving members form a series of sectors of circles concentric with the axis of rotation. The width of these sectors, where they overlap the sinusoidal patterns on the other disc, is half a wavelength, with half a wavelength spacing.

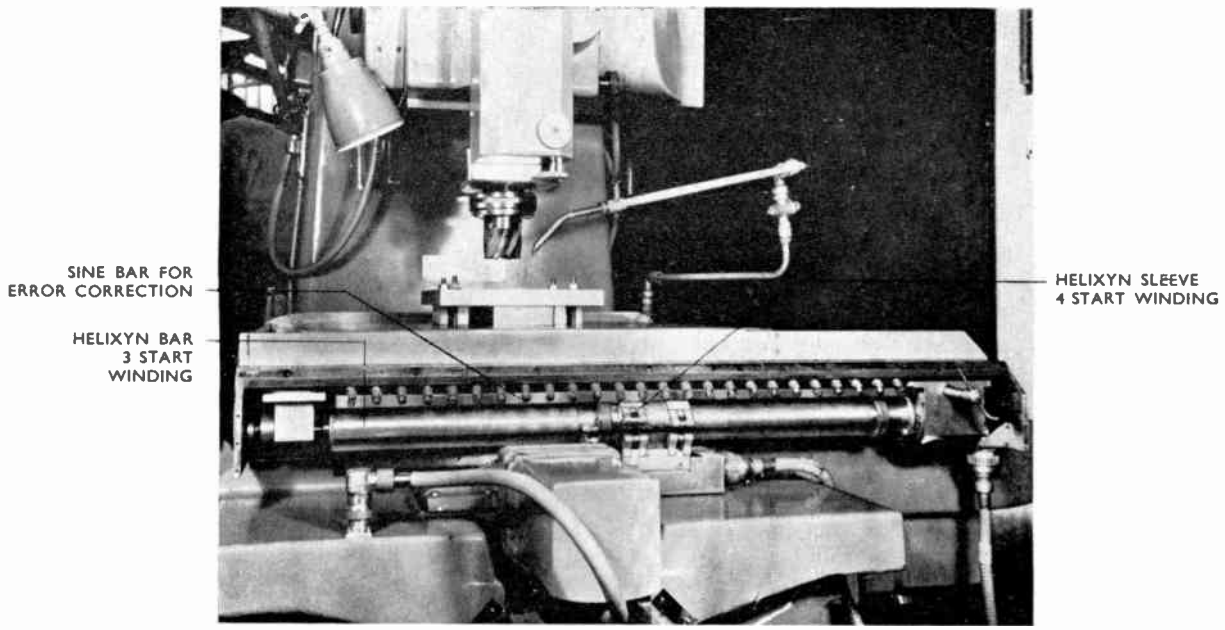


Fig. 12. (*b*) A.E.I. Helixyn capacitance transducer.

The input signal consists of amplitude modulated sine waves generating the sine and cosine of the electrical angle required. These two signals are applied to the glass disc between the plated circles and the output signal is taken off the periphery. The function of the moving plate is to provide variable capacitive coupling between output and input. The output signal is a constant amplitude signal of varying phase angle with respect to the reference supply.

This "resolver" with 200 repetitions of the sinusoidal pattern is able to measure to 6 seconds of arc, i.e. 1·8 deg of mechanical rotation yield 360 deg of electrical phase change. The basic scale is then interpolated to 1 part in 1000. The supply voltage has to be in the range 0·5–30 kc/s and when a coarse-fine pattern is plated on the same discs different supply frequencies are employed to minimize cross talk.

5.4. *Optical Analogue Transducers*

Optical analogue transducers have been applied in this country, the earliest being the Barber transducer.[32] The transducer consisted of three gratings and two photocell heads. A drum grating is rotated (see Fig. 13(*a*)) and used to create moiré fringe patterns in conjunction with two linear gratings. One grating is fixed into position relative to the drum grating, the other is free to move. If the free grating is stationary the two sets of photocells will give out the same frequency signal with a constant phase relationship. Any movement creates a change in instantaneous frequency generated between the moving grating and the drum grating. The phase difference between the photocell outputs can be utilized to produce a misalignment signal. This technique had never been directly applied to a machine but it may be regarded as the first step in the use of diffraction gratings in analogue measurement.

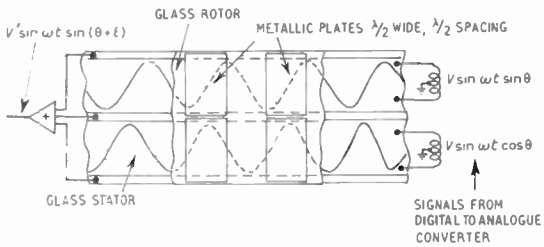


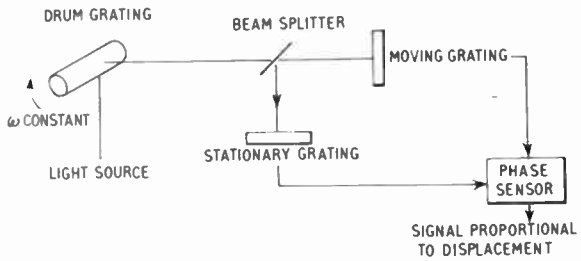
Fig. 12.

(*c*) A linearized sketch of the Telecomputing capacitive transducer.

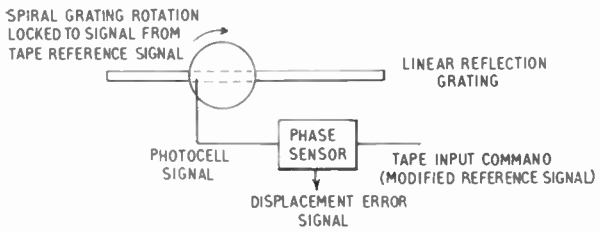
The Ferranti transducer[33] differs in detail from the above. Here, a disc with spiral rulings on it is rotated at constant speed (locked to the tape input information). The frequency signal generated from the moiré fringe patterns is a constant frequency sine wave when the linear grating is stationary (see Fig. 13(*b*)). In this condition, the reference frequency and grating signal frequency are the same, any table movement upsetting this condition and allowing a misalignment signal to be established. Grating pitches used are 100, 40 or 10 lines/in and interpolation gives the smallest digit of input to be equal to 1/50 of a line pitch. The reference frequency is 140 c/s.

The Staveley-N.P.L. transducer[34] is very similar in principle to the two already mentioned. However,

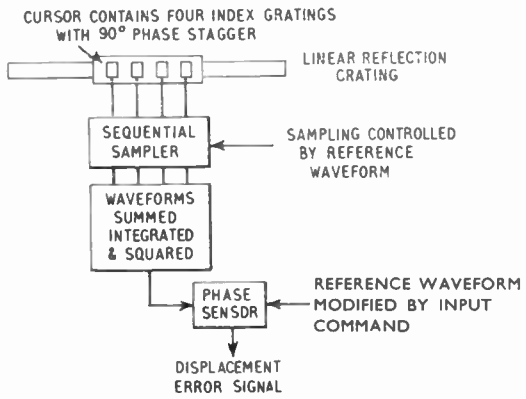
there is one very marked difference: the alternating grating waveform is generated by switched sampling of four photocells, each measuring the intensity patterns from four index gratings. The four index gratings are staggered in phase by 90 deg steps. This obviates the need for any mechanical movement. The sampling will result in a periodic triangular waveform and with zero relative movement the phase relationship between this waveform and the sampling waveform is constant. This phase relationship is altered by either movement at the table or input commands fed in as phase changes to appropriate waveforms. When pitches 100 lines/in are employed, this transducer system will measure to $\pm 2 \times 10^{-4}$ in. This transducer is not a moiré fringe element as the pair of gratings are in line and the light system produces a "shutter" effect rather than a fringe pattern.



(*a*) The basic transducer described by Barber and Atkinson.



(*b*) The Ferranti analogue diffraction grating transducer.



(*c*) The N.P.L.-Staveley analogue diffraction grating transducer.

Fig. 13. Analogue optical transducers.

It is essential to appreciate the reasons for the evolution of analogue grating systems. Later in the paper digital grating systems will be mentioned. Here the effect of error is minimized by sampling over many line pitches and hence integrating errors to a minimum. However, the accuracy of individual line ruling did not merit the use of interpolation techniques. The development of improved techniques for the production of diffraction gratings[35] has made available a supply of relatively coarse gratings (rulings of up to 100 lines per inch); when the accuracy of line rulings merits the introduction of analogue interpolation techniques, a more economic circuit design can be achieved.

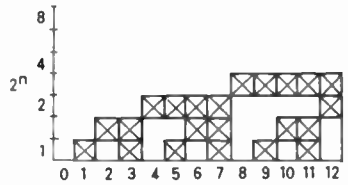
6. Digital Position Transducers

A digital or quantized measure of position can be obtained by one of two techniques. Either an incremental system can be employed where one pulse is obtained per unit of displacement or a coded pattern can be employed to generate sets of pulses which give an absolute measure of position within the length of the pattern.

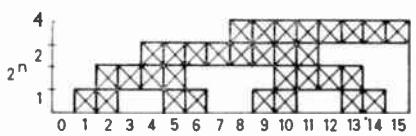
6.1. *Coded Plate Transducers*

A very wide range of digital codes can be employed,[36] examples of two of the most common being shown in Fig. 14. The binary code is the most familiar pattern, where the resultant read-out expresses the displacement in powers of two. The second code, the Gray code, has the advantage over the simple binary code of minimum ambiguity of read-out. This will be readily appreciated when it is realized that the difference between any two successive counts on the Gray code requires a change of only one channel of the read-out.

The coded pattern is an absolute measure within the extent of one length of the scale. Obviously the repetition of the pattern gives a multi-channel mea-



(*a*) Binary coded pattern—4 digits.



(*b*) Gray (or reflected binary) code—4 digits.

Fig. 14. Typical digital coded scales.

suring system, i.e. a count has to be made of the integer number of pattern repetitions and the desired displacement within one pattern. One factor which limits the realization of high accuracy is that the transducer has to be coupled to rotary elements and thus include the errors contained in lead-screws, pinions and gearboxes. This has to be accepted because the cascading of linear coded plates would involve errors greater than the resultant rotary form because there is no integration of pattern error as is common in most analogue or incremental digital transducers.

The design and use of the digitizer and in particular the optical digitizer, has been covered in great detail in the literature.[37] An application of this technique has been described,[38] where the accuracy of measurement is 10^{-3} in/ft. The dominant limitation to the accuracy of measurement is the error present in the pinion driving the digitizer.

An application of Gray code discs has been described,[6] where an overall accuracy of $2{\cdot}5 \times 10^{-3}$ in was attained. In this case, the input data are expressed in binary code and a code conversion is required to enable the error to be sensed.

Systems employing coded discs include the Norden-Ketay,[6] Bendix[22] and the Warner and Swasey.[40] An interesting departure from the more usual transducer employing optical or resistive read-out is the G.E.C. inductive device, which would appear to be well suited to industrial applications.[41]

One class of coded transducer which is really a multi-channel coded system is the decimal scaled transducer. Here an absolute measure of position is obtained from a series of scales in decimal form.

Examples of this technique include the transducers employed in systems developed by Arter,[6] Airmec[42] and Electrosystems.[39] The Arter transducer has an interesting construction—the segments of the decimal scales are etched in the form of a printed circuit on bakelite board.

One feature of all coded digital transducers is that inherently they possess a measure of the sense of the direction of motion. This is not true of a simple incremental scale.

6.2. *Incremental Digital Scales*

The basic incremental scale will provide a pulse output for a pre-determined displacement. It will not be possible to obtain the sense of the displacement. If, however, more than one sensing system is employed and a certain phase displacement is maintained between the sensors, then as is shown in Fig. 15 it is possible not only to sense the direction of movement, but also to improve the resolution of the transducer.

In the example quoted, the two outputs are 90 deg out of phase and comparison of the two waveforms

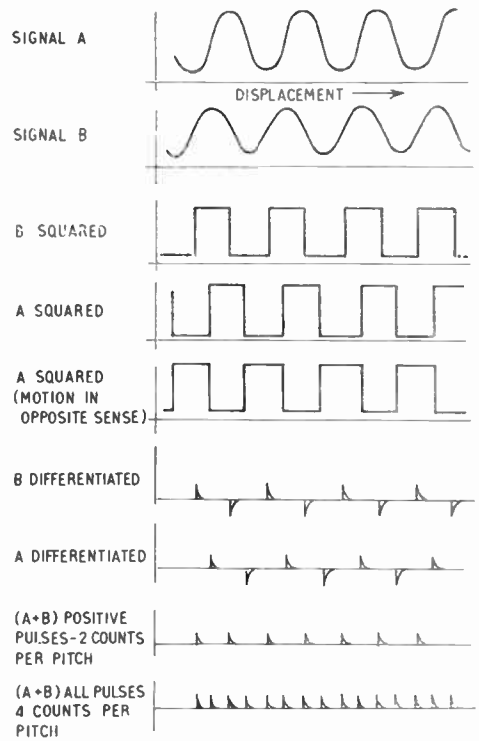


Fig. 15. Direction sensing and scale amplification with incremental scales.

provides a measure of the sense of the displacement and a divide-by-four on the pulse count can be achieved.

A wide range of techniques has been employed to produce incremental position transducers. For clarity, they have been classified as before into the following categories:

(1) Electromagnetic transducers,

(2) Optical transducers,

(3) Resistive transducers,

(4) Capacitive transducers.

6.2.1. Electromagnetic incremental transducers

In the discussion of analogue transducers, it was stated that the synchro type of transducer was the most widely used position sensing element. It is, therefore, interesting to consider two methods which have been adopted to provide digital information from a synchro.

The A.C.E.C. synchro resolver[43, 44] is a specially-designed element which has five stator windings instead of the usual two and one rotor winding. Each stator winding can be excited via one of two switches, the additional provision being that only two of the resultant switches can be made at one time allows one revolution of the resolver to be divided into ten distinct positions where nulls can be established. In

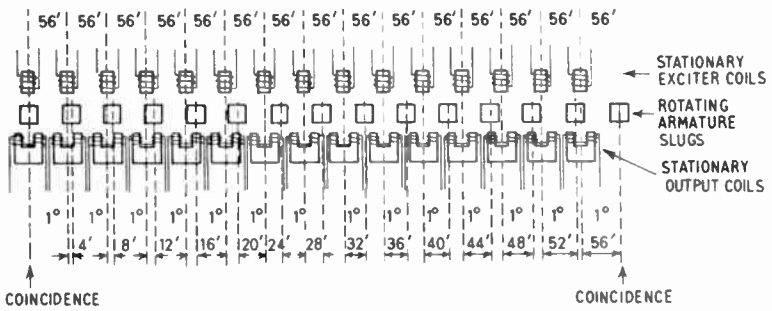


Fig. 16. Schematic of the Cleveland vernier transformer.
(Courtesy *Control Engineering*.)

this system, one revolution of the fine resolver is equivalent to 2 mm. To allow subdivision to 1/200 of revolution, the digital-analogue converter feeding the stator windings can supply twenty voltages. The same principle is applied to the coarse and medium resolvers. The medium resolvers give one revolution per 50 mm and has to be subdivided into one part in 25. The coarse resolver measures 2000 mm per revolution and has to measure up to 1/40 of a revolution.

The transducer developed by Sperry Gyroscope (Canada)[45] could almost be regarded as an Inductosyn with 25 distinct stator windings. The stator consists of metallic "hairpin" patterns plated on glass, rather like an Inductosyn scale. The moving member moves parallel to the stator but here the pattern is subdivided into 25 separate windings of staggered phase. The selection of the appropriate coil in the transducer allows positioning to 10^{-3} in.

The Cleveland vernier-differential transformer[39] may be regarded as a digital device, although not quite in either of the two simplified categories. Rough positioning is provided by a potentiometer to 10 min of arc. In its angular form the transformer has 360 armature slugs on 1 deg centres. On the other member of the transducer, there are fifteen exciter coil/output coil differential transformer structures on 56 min centres (see Fig. 16). When one armature slug is aligned with a transformer structure, the output is zero. The effective vernier system allows a 1 deg interval to be broken down into 4 min intervals. Interpolation over the 4-min interval is achieved by electrical shift of the transformer null output, yielding resolution of 1 second of arc.

Two accurate "point-to-point" positioning systems have been built around the use of accurate magnetic coarse scales. In each case, when the slide has been positioned to the nearest inch on the coarse scale, the fine position is determined by a micrometer screw. The Electrolimit transducer[46] consists of a castellated bar (see Fig. 17(*a*)), and an E-shaped inductive pick-off containing two coils which are used in bridge circuit for determining balance. The castellated surface gives a symmetrical form with 1 in pitch. The accumulated error over the whole length is said to be of the order of 2×10^{-5} in.

The A.E.I. measuring bar consists of a steel bar with circular brass inserts which allow 1 in intervals to be sensed.[47] The system is shown in Fig. 17(*b*), which gives a clear picture of the complete transducer and indicates the use of the micrometers. A fine adjustment is provided by "D" pins sited in the brass inserts.

A Russian transducer has been described,[48] which is similar in physical form to the Société d'Electronique & d'Automatisme transducer. In this case, however, serrations of 0·1 mm pitch are employed and a 20 kc/s input is provided for the E transformer. The resultant sine and cosine terms from the two outputs of reading head are processed to give a pulse for every quarter of the serration pattern traversed, i.e. for a displacement of 0·025 mm. This is essentially the same process as that shown in Fig. 15. Here, however, a phase-sensing technique is added to derive sine and cosine terms from the carrier frequency used to excite the magnetic circuit.

The Bendix incremental transducer[49] is an angular device consisting of two glass plates. On one plate a metallic hairpin pattern is deposited. On the second plate two similar patterns of similar pitch, with a

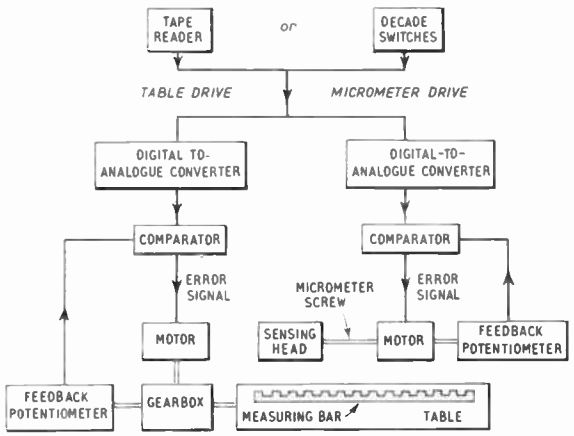


Fig. 17. Magnetic measuring bars.

(*a*) A schematic representation of the Pratt and Whitney use of the "Electrolimit" transducer. Note the use of two control systems—one for the table—one for the transducer.
(Courtesy *Engineers' Digest*.)

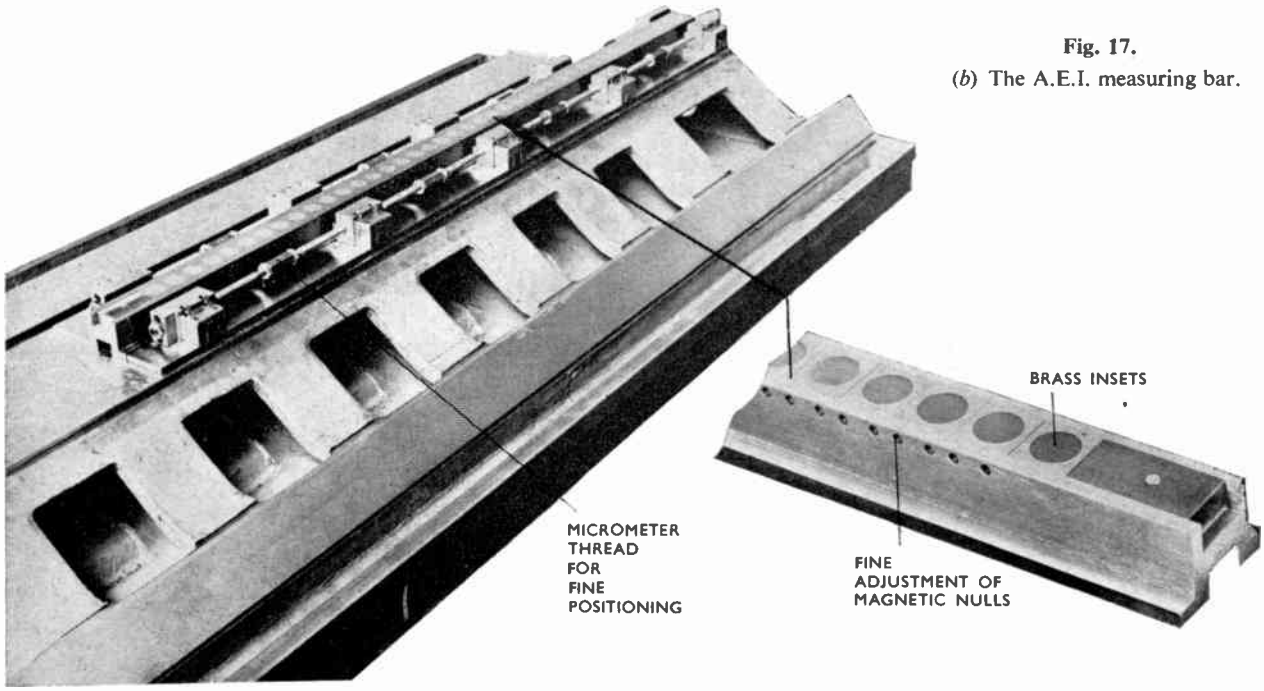


Fig. 17.
(*b*) The A.E.I. measuring bar.

90 deg phase displacement, sense signals which are processed to give both a count for every quarter of a pitch displacement and the sense of the displacement. The transducer gives 1000 pattern repetitions for each cycle and is used to measure via a lead-screw to $\pm 2 \times 10^{-4}$ in.

A similar transducer developed by Hughes Aircraft[50] senses the presence or absence of grooves photo-etched in a steel bar with a spacing of 250 lines per inch. The use of two magnetic sensing heads out of phase by one quarter of a ruling enables this to resolve to 10^{-3} in. The rotary form of this transducer has been developed to measure to 10^{-4} in and has replaced the photo-electric transducer formerly employed.

A transducer based on the use of a square-toothed gear wheel has been described.[51] A gear wheel with 50 teeth rotates 12·5 times for each revolution of the lead-screw. In this system a rotation of the gear of tooth width is equivalent to a table movement of 0·0004 in. Two pick-offs are used to give a direction sensing incremental scale.

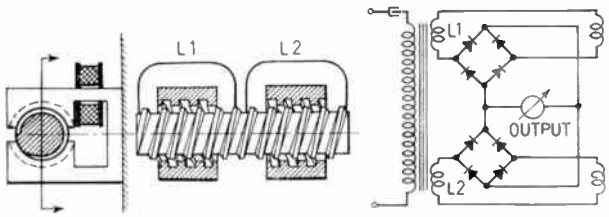
A similar approach has been employed in the development of an electromagnetic system where the repetition of magnetic pattern is sensed by four reading heads, staggered by one quarter of a cycle.[23] The squaring of these four waveforms allows the subdivision of the basic scale by a factor of four. Quasi-linear interpolation† is then achieved by utilizing the appropriate "linear" section of the four sinusoidal patterns generated by the reading heads. The magnetic scale is repeated every 2 in—this of course will give a coarse count of 0·5 in and it is stated that 2×10^{-4} in accuracy can be achieved overall.

A Russian angular transducer[52] works on a similar principle to the method described above, but here two sets of corrugated surfaces (in this case gear profiles) are provided and are staggered by 90 deg in phase. Rotation of the gears creates alternating variations of the impedance of the coils used to sense displacement. A method of compensation is also described for eccentricity of the gear wheels. For an application to gear hobbing machines, the gear wheels were 1000 mm diameter and had 1000 teeth.

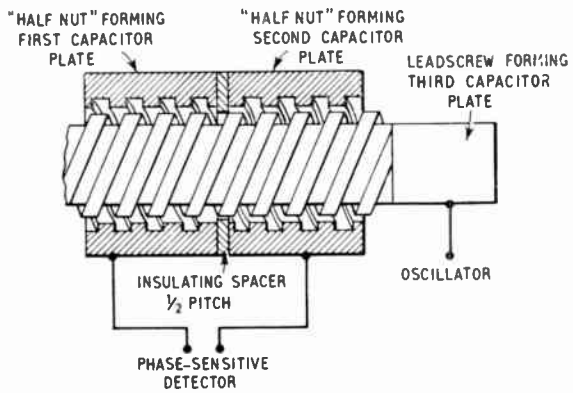
A further Russian transducer (see Fig. 18) has been described[53] which is of a similar nature to that described above. Here a split nut is mounted on the lead-screw. The impedance of the coils in each half-nut varies as the thread of the lead-screw rotates. This transducer is employed in the same mode as the Nultrax transducer described earlier. The thread pitch is 5 mm and the positional accuracy attained is 0·005 mm. The two half-nuts are half a pitch out of phase.

Perhaps the most accurate angular transducer employed in machine tool control is the magnetic scale developed by Stepanek.[54] Basically, the transducer is merely a disc with a ferromagnetic disc coated around its periphery. If a constant frequency is recorded on the periphery as the disc rotates, a scale will be created where resolution will be a func-

† C.f. "Numill" system.[60]



(*a*) Russian variable inductance transducer.



(*b*) Mullard variable capacitance transducer.

Fig. 18. Transducers using a split-nut on lead-screw configuration.

tion of frequency recorded and disc speed when recording. There are obvious limitations to this simple concept but sophisticated compensation techniques have been evolved, so that the cumulative error is less than 1 second of arc.

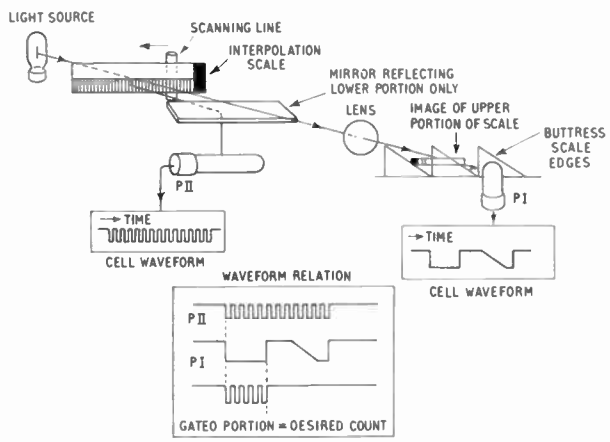
6.2.2. Optical transducers

The simplest incremental optical transducer is the disc with a series of holes drilled on a constant diameter, with the variable light intensity sensed by a photocell. Such transducers have been employed,[55] but the weakness of this technique is that very little light can pass through the small aperture corresponding to a fine subdivision of the disc periphery.

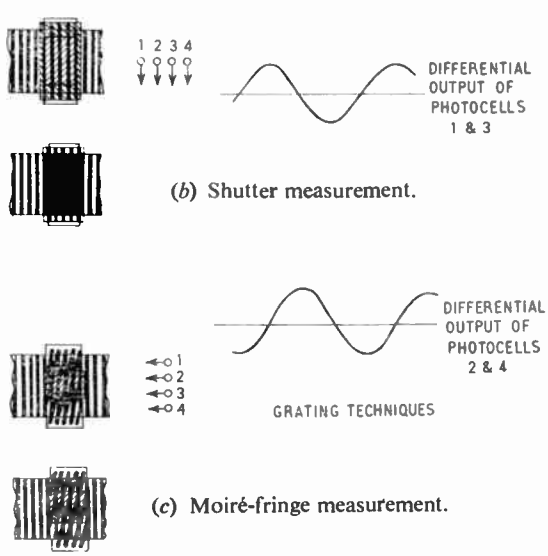
The use of optical grating techniques overcame the limitation of the simple method of counting by sampling the intensity variations produced by large numbers of fine rulings. The basic methods employed to date are shown in Figs. 19(*b*) and 19(*c*).

The highest resolution achieved by the use of the simple technique of counting a sucession of opaque sections on a scale is 10^{-4} in. This transducer[56, 57] employed an accurate buttress scale machined in a rod, with a pitch of 0·1 in and a pitch error of 2×10^{-5} in. A count of this pattern gives a coarse measurement. The fine measurement relies on a grating 2 in long with 1000 rulings. The projected image of this grating pattern can be projected to fit exactly between adjacent vertical edges of the buttress scale. A combination of the coarse count and a count of projected rulings not obstructed by the buttress thread results in a transducer measuring to 10^{-4} in. A sketch of this transducer is given in Fig. 19(*a*).

Diffraction gratings have provided the basis for many measuring systems.[58] Two are shown schematically in Figs. 19(*b*) and 19(*c*). In each case the effect of a half a pitch displacement is shown. The Ferranti system[59] relies on the detection of the moiré fringe patterns created by inclining two gratings at a small angle. Four photocells are used to monitor the fringe pattern. The four cells work in two pairs to minimize the effect of photocell dark current. This technique, coupled with 2500 lines per in gratings, gives a direction-sensitive transducer with one pulse per 10^{-4} in.



(*a*) Mullard optical incremental transducer.
(Courtesy *Machinery Lloyd.*)



(*b*) Shutter measurement.

(*c*) Moiré-fringe measurement.

Fig. 19. Incremental optical techniques.

The Numill transducer[60] is very similar to the Ferranti transducer. In this case, however, the two gratings are parallel and the light intensity changes are obtained by monitoring the "shutter" pattern obtained as relative movement takes place between the gratings.

A Japanese coarse-fine optical system has been described,[61] where the coarse scale is formed by indenting a mechanical scale on the table which is sensed optically. This coarse scale gives 1 pulse/mm. A fine scale comprising a conventional moiré fringe scale gives an output of 1 pulse/micron.

The technique of scanning an engraved scale with a photoelectric microscope is employed in the Société Genevoise positioning system.[5] This method is virtually the same as the coarse measuring channel described above. Here, however, this provides the fine measurement and the coarse and medium transducers are synchros. It is stated that the accuracy attained is $\pm 10^{-4}$ in.

Grating methods are further described in recent papers by Shepherd[33] and Leslie,[62] in which the current work of Ferranti and the National Engineering Laboratory respectively is reported.

6.2.3. Resistive transducers

The only transducer of interest in this category is an incremental disc.[63] The disc has two concentric sets of 240 conducting segments out of phase by a quarter of a pattern. With the appropriate gearing this transducer gives one pulse per 0·01 mm.

6.2.4. Capacitive transducers

The only transducer under this heading is the capacitive equivalent manufactured by Mullard[64] of the Russian transducer mentioned earlier (see Fig. 18(*b*)). The transducer consists of a bar cut to the form of a lead-screw and wound with a bifilar coil, together with a non-contacting sleeve or nut containing a secondary bifilar winding. The sleeve slides over the bar.

As the sleeve is moved along the bar, a series of nulls in the induced electrical currents of the secondary are detected, the spacing of these nulls being equal to the helical pitch. For accurate systems this pitch is generally 0·1 in but leads of 0·05 in for very high resolution and 0·2 for less accurate purposes are possible. Taking the case of a 0·1 in pitch, interpolation between nulls is performed by rotating the bar a fraction of a turn equivalent to the fraction of 0·1 in required, and the picking up the null point by moving the table carrying the sleeve along. Thus a rotation of 3·6 deg corresponds to 1×10^{-3} in, and 0·36 deg to $0{\cdot}1 \times 10^{-3}$ in; this angular movement is not difficult to obtain with a servo-motor driving through a reduction gear. A sketch of this transducer is given in Fig. 19(*b*).

7. Conclusions

A very wide range of techniques is covered in a survey of this nature and therefore it is difficult to derive a set of clear-cut conclusions. The possibility of accuracy classification has been left unexplored. The reason for this is the shortage of detailed assessments of transducer performance. Of all the transducers reviewed in this paper, only three clear statements of periodic and cyclic errors are available in the literature. It can, however, be concluded that in virtually all cases, the control system displacement transducer contributes less error than either the machine tool members or the cutting process.

The improvement of metal removal techniques and machine tool design may however lead to a position where the cyclic errors of displacement transducers are reflected more significantly in the surface finish of the end product.

The achievement of the ultimate accuracy is not in itself the end product desired by industry for a large number of production applications. Therefore the trend of future developments in the majority of cases where economy rather than accuracy is at a premium may be seen by studying the recent Russian and Japanese work on open loop feed drives.

8. Acknowledgments

The author wishes to thank Mr. D. L. Leete for providing the material for Fig. 7 and all the many firms who also provided material.

9. References

1. M.I.T. Servomechanisms Laboratory Report, "Design, development and evaluation of a controlled milling machine", Ministry of Aviation, Ref. No. P.68610(2).
2. P. J. Farmer, "Fairey-Ferranti three-dimensional contour milling machine operated by a digital control system", *Aircraft Production*, **20**, No. 5, p. 174, 1958.
3. A. J. Sabberwal, "Chip section geometry and cutting forces during the milling process". Ph.D. Thesis, Manchester University, 1961.
4. G. Schlesinger, "Testing machine tools", p. 18. (The Machinery Publishing Co., Brighton, 1961.)
5. "Engine machining at Aston-Martin", *Machine Shop Magazine*, **23**, No. 1, p. 5, 1962.
6. J. D. Cooney and B. K. Ledgerwood, "31 numerically controlled point-to-point positioning systems—Part 1", *Control Engineering*, **5**, No. 1, p. 56, 1958.
7. C. R. Hibbard, "Tape control of a jig borer", *Aircraft Production*, **21**, p. 357, 1959.
8. S. A. Davies and G. K. Ledgerwood, "Electromechanical Components for Servomechanisms", p. 92. (McGraw Hill, New York, 1961.)
9. F. G. Helps, "Data transmission by synchros", *Electronic Engineering*, **28**, pp. 438–45, October 1956.
10. H. A. Dinter, "Importance factors influencing the choice between a synchros and a potentiometer as an angular pick off", *Trans. Amer. Inst. Elect. Engrs II*, **75**, p. 198, 1956; *Applic. and Industr.*, No. 26, September 1956.

11. P. J. Farmer, "Automatic machining, the numericord data control system for machine tools", *Aircraft Production*, **20**, No. 1, p. 28, 1958.
12. J. G. Dixon, "Synchro following accuracy", *Muirhead Technique*, **15**, No. 3, pp. 19–20, July 1961.
13. M. V. Stevens, "A survey of synchro computing resolvers and their testing", *Muirhead Technique*, **15**, No. 1, pp. 6–9, January 1961.
14. P. J. Farmer, "Position identification", *Aircraft Production*, **19**, No. 10, p. 400, 1957.
15. H. J. Finden and B. A. Horlock, "The inductosyn and its application", *J. Brit.I.R.E.*, **17**, p. 369, July 1957.
16. Engineering Report No. 503, Inductosyn Corporation, 729 Carson Street, Carson City, Nevada, U.S.A.
17. G. W. Younkin, "Single time shared drive position two axes", *Control Engineering*, **7**, No. 8, p. 122, 1960.
18. G. W. Younkin, *Amer. Inst. Elect. Engrs*, Paper DP60–747, March 1960.
19. P. J. Farmer, "Data controlled milling", *Aircraft Production*, **20**, p. 102, 1958.
20. F. W. Hartley, "The E.M.I. system of machine tool control". A. D. Booth (editor) "Progress in Automation", Vol. 1. (Butterworth, London, 1960.)
21. "Lecteur lineaire magnetique SEA type 11B 133 pour machines-outils", Société d'Electronique et d'Automatisme Report NC92D, January 1959.
22. R. C. Brewer, "Numerical control of machine tools", *Engineers' Digest*, No. 5, p. 46, September 1959.
23. F. P. Caruthers, "Prepunched key permit rapid set-up of programmes", *Control Engineering*, **7**, No. 10, p. 143, 1960.
24. F. Brouwer, "A critical evaluation of high precision electromechanical linear measuring systems", *Electrical Manufacturing*, p. 57, August 1957.
25. A. D. Booth (Editor), "Progress in Automation", Vol. 1, p. 51. (Butterworth, London, 1960.)
26. K. J. Coppin, "A 'basic' system of position control for the traversing tables of machine tools", *J. Brit.I.R.E.*, **17**, p. 263, October 1957.
27. "Co-ordinate drilling", *Aircraft Production*, **19**, p. 195, 1957.
28. "A new system of linear measurement and positional control", *Machine Shop Magazine*, p. 150, March 1960.
29. D. A. Mynall, "A new idea of precision linear positioning", *Control Engineering*, **6**, No. 6, p. 125, 1959.
30. D. J. Mynall, "Helixyn position control", J. F. Coales (editor) "Automatic and Remote Control", Vol. 4, p. 156. (Butterworth, London, 1961.)
31. J. O. Morin, "Six transducers for precision measurement", *Control Engineering*, **7**, No. 5, p. 107, 1960.
32. D. L. Barber, "Method of displacement measurement using optical gratings", *J. Sci. Instrum.*, **36**, No. 12, p. 501, 1957.
33. A. T. Shepherd, "Moiré fringe measuring techniques". Paper read at Brit.I.R.E. Symposium on "Recent Developments in Industrial Electronics", London, April 1962.
34. B. J. Davies, R. C. Robbins, C. Wallis and R. W. Wilde, "A high-resolution measuring system using coarse optical gratings", *Proc. Inst. Elect. Engrs*, **1073**, p. 624, November 1960. (I.E.E. Paper 3312 M.)
35. J. M. Burch, "Photographic production of gratings for measurement", *Research*, **13**, 217, January 1960.
36. A. K. Susskind, "Notes on Analog-Digital Conversion Techniques". (John Wiley, New York, 1957.)
37. D. S. Evans, "Digital Data". (Hilger & Watts, London, 1961.)
38. "Precision dimensional and position control", *Machinery*, **89**, p. 82, 1951.
39. J. D. Cooney and B. Ledgerwood, "31 numerically-controlled point-to-point positioning systems—Part 3", *Control Engineering*, **5**, No. 3, p. 108, 1958.
40. H. W. Mergler," Numerical building blocks control turret lathe", *Control Engineering*, **7**, No. 7, p. 74, 1960.
41. C. J. Wayman, "Shaft angle digitizer needs no brushes", *Control Engineering*, **7**, No. 8, p. 145, 1960.
42. "Airmec automatic co-ordinate setting equipment", *Machinery*, **93**, p. 616, 19th September 1958.
43. R. Bingen and J. Vroman, "Punched-tape-controlled automatic lathe", J. F. Coales (editor) "Automatic and Remote Control", Vol. 4, p. 5. (Butterworth, London, 1961.)
44. R. Bingen and J. Vroman, "Numerically controlled Belgian lathe", *Control Engineering*, **5**, No. 9, p. 173, 1958.
45. "Drill retrofitted for numerical control actuated by linear hydraulics", *Machinery (New York)*, p. 104, February 1961.
46. P. J. Farmer, "Automatic setting (Pratt & Witney system)", *Aircraft Production*, **19**, No. 12, p. 456, 1957.
47. P. J. Farmer, "Position control (A.E.I. system)", *Aircraft Production*, **18**, No. 5, p. 180, 1956.
48. R. C. Brewer, "Numerical control of machine tools", *Engineers' Digest*, p. 46, September 1959.
49. "Digital servo for numerical control uses unique transducer", *Control Engineering*, **7**, No. 5, p. 155, 1960.
50. R. C. Bell, "Magnetic scale substitutes for light chopper", *Control Engineering*, **7**, No. 10, p. 145, 1960.
51. *Electrical Manufacturing*, No. 8, p. 121, 1955.
52. L. M. Kaufman, "Beskopirniye Sistemy Automatizatsii Stankov", p. 42. (Mashgiz, Moscow, 1959.)
53. E. M. Goloyl'nikov, "Inductive device for angle measurement", *Izmer. Tekhnika*, **4**, p. 9, 1961. (Abstract in *Soviet Technology Digest*, p. 20, July 1961.)
54. J. Stepanek, "Magnetic scales", *Czechoslovak Heavy Industry*, No. 9, p. 3, 1959.
55. J. D. Cooney and B. K. Ledgerwood, "31 numerically controlled point-to-point positioning systems—Part 2", *Control Engineering*, **5**, No. 2, p. 100, 1958.
56. J. D. Cooney and B. K. Ledgerwood, "31 numerically controlled point-to-point positioning systems—Part 1", *Control Engineering*, **5**, No. 1, p. 97, 1958.
57. "Jig-boring machine with electronic positioning", *Machinery Lloyd*, **28**, p. 16, August 1956.
58. J. Guild, "Diffraction Gratings as Measuring Scales". (Oxford University Press, 1960.)
59. Ibid., p. 106.
60. P. F. Fischer, "The numill numerical control system", *Electrical Manufacturing*, September 1958.
61. S. Nishida, Y. Doi and K. Togino, "Numerical control of jig borer 'Jidic' ", J. F. Coales (editor) "Automatic and Remote Control", Vol. 4, p. 42. (Butterworth, London, 1961.)
62. W. H. P. Leslie, "Widening the applications of diffraction gratings by measurement and control", *Inst. J. Mach. Tool Des. Res.*, **2**, p. 393, 1962.
63. T. Kaiwa and S. Inaba, "Japanese tape controlled milling machine", *Control Engineering*, **6**, No. 11, p. 103, 1959.
64. G. Butcher, "Numerical control for marking out and jig-boring", *Automation Progress*, **4**, No. 5, p. 166, May 1959.

POINTS FROM THE DISCUSSION

Mr. J. R. Arrowsmith: It is interesting to note that a Patent Specification was published as far back as 1942 which outlines a number of methods of using the differential capacitor principle to give outputs which change when there is a linear translation of the elements relative to one another. It is also interesting to realize that a semi-automatic inspection device was apparently made even in those days.

Mr. P. A. Jassoy: What advantages are there in using analogue servos as against digital servos?

Do you believe that there are advantages in using coded types of pick-offs, rather than incremental types?—e.g. absolute position measurement, simple digital/analogue converse pulse output, etc.

The author (*in reply*)*:* In current control systems applied to machines the same order of performance is obtainable either by digital or analogue techniques. A significant factor, however, is the economics of circuit design. With computer-prepared tapes there is an obvious simplicity in employing bi-directional counters and digital transducers.

For continuous control there would appear to be no advantage in employing coded transducers. A decimal coded transducer may be useful in co-ordinate setting systems.

Mr. P. Tomkins: Assuming a form of positional control is to be applied to an existing machine tool, does Mr. Bell consider that "dither" techniques to reduce stiction are worth investigation?

The author (*in reply*)*:* Dither can be beneficial under certain conditions. It should not however be included as part of the design of a new controlled machine tool. The dither technique only attempts to overcome bad design.

Mr. P. Huggins: Can Mr. Bell substantiate the very high claims made by the Reilly Engineering Company for accuracy? This seems very high for a device which balances inductance against capacitance, bearing in mind the practical difficulties of obtaining a non-resistive inductor.

The author (*in reply*)*:* The only information available on the Reilly transducer was given in the technical press.

There are certain practical difficulties in the use of this transducer over several feet. However, there would appear to be great flexibility of application for lengths of the order of 10 in or less.

The non-resistive inductor would present no problem. This particular technique has been employed for several years in a high accuracy continuous control system.

U.D.C. 621.396.96.32

Visual Detection in Intensity-Modulated Displays

Presented at the Symposium on "Sonar Systems" in Birmingham on 9th-11th July 1962

By

J. W. R. GRIFFITHS, Ph.D.
(*Associate Member*)†

AND

N. S. NAGARAJA, Ph.D.‡

Summary: In a two-dimensional intensity-modulated display such as the plan position indicator used in radar systems, the observer is looking for and recognizing the signal as an area, or pattern, of brightness differing from that of the surround. By using a closed circuit television system it was possible to reproduce the essential features of this type of display whilst having complete control over the important parameters—signal area, background noise, target presentation time, etc. In particular, the background can be changed from being uniformly illuminated, i.e. the situation studied by many psycho-physiologists, to the more realistic situation appertaining to radar displays, i.e. when the background is completely perturbed by noise.

Threshold signal/noise ratios have been measured for a number of such conditions and compared with those of an equivalent theoretical model. The results suggest that the visual detection system is a sub-optimum one and its efficiency is dependent, among other things, on the area of the signal. This point goes some way to explain some previously observed discrepancies between experimental and theoretical rates of improvement with increase of area.

Table of Symbols

R_B	output voltage signal/noise ratio
ρ	input voltage signal/noise ratio
B	screen luminance
v	law of demodulator
R	amplitude of r.f. signal
C	luminance contrast
r	normalized luminance fluctuation
N	equivalent number of samples
X	time average of $V(t)$ over a period T $= \frac{1}{T}\int_{t}^{t+T} V\,dt$
f_0	r.f. bandwidth
V	demodulator output voltage
$W(f)$	power spectrum of $V(t)$
$\phi(\tau)$	autocorrelation function of $V(t)$
A_1, A_2, A_3	constants relating to power spectrum
Λ	likelihood ratio
s	distance from centre of target
σ^2	variance of luminance of independent element of area
D	number of elements/unit area
ψ_0	mean noise power
P	sinusoidal signal amplitude
$p_s(R)$	probability density of amplitude (signal present)
$p_n(R)$	probability density of amplitude (noise only)
q	$= \frac{1}{N}\sum R_R^2$ mean square of amplitude samples.

1. Introduction

In echo location systems such as radar or sonar the detection of distant or small targets is limited by the random noise inherent in the system. The term detection is used here in the sense of a decision process, i.e. an observer is required to decide whether at any particular range the received waveform comprises noise alone or noise plus a target echo. In practice more than one return is received from every target, thus enabling the decision to be based on a number of observations and resulting in either an increased reliability of the decision or, alternatively, for the same reliability, the detection of smaller echoes. Performance is usually assessed in the latter method, i.e. a standard of reliability is fixed and we define the threshold as the signal/noise ratio required to obtain this standard. A convenient criterion which is often used is the rate of improvement of this threshold for each doubling of

† Electrical Engineering Department, University of Birmingham; at present Guest Professor at the Institute of Radiophysics and Electronics, University of Calcutta.

‡ Formerly at the Electrical Engineering Department, University of Birmingham; now with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India

the number of observations but in practice many experimenters have obtained various values for this improvement varying from 1·2–2·5 dB.[1,2,3,5,8]

To develop a mathematical model of detection it is necessary to know what factors the observer takes into account when making his decision. For instance a simple treatment of this problem assumes that he bases his decision on the output signal/noise ratio, the latter being defined in the following way.

$$R_B = \frac{\text{change in mean output due to the presence of the signal}}{\text{r.m.s. fluctuation in the absence of the signal}}$$

The relation between R_B and the input signal/noise voltage ratio can be simply calculated and hence the rate of improvement when the number of traces is doubled. If N traces are integrated, i.e. stored and then added in the proper time sequence so that returns from each range add linearly, then the value of R_B increases by $\sqrt[4]{N}$. This is of course because, assuming no correlation from trace to trace, both the mean value and the variance will increase by a factor of N and hence the r.m.s. by a factor of $\sqrt{N}$. For small input signal/noise ratios (less than unity) the relation between input and output signal/noise ratios is very simple,[7] namely

$$R_B = \rho^2 \qquad \text{.....(1)}$$

This is exactly true only for the square-law demodulator but the approximation is quite good for almost any non-coherent demodulation provided $\rho \ll 1$.

From the above it may be seen that if the decision criterion is assumed to be a fixed value of R_B then by integrating N successive traces the value of ρ is reduced by a factor $\sqrt[4]{N}$, i.e. for each doubling of the number of traces there is an improvement of 1·5 dB. The fact that a higher rate of improvement of the threshold than this is obtained in certain displays does not necessarily mean that the analysis is incorrect, since the measured threshold of the display may be very much below the optimum absolute threshold when the number of traces is small and approach closer to the optimum as the number of traces is increased. It would appear then, that a more valid basis for comparison would be the absolute threshold itself and that optimum models based on statistical decision theory provide a suitable reference for such a comparison. However even this method requires a statement of the decision procedure used by the postulated observer, and will in general lead to different results for different types of observer. In the experiments which will be described later a forced choice method of selection was used and fortunately in this case there is no ambiguity as to which type of observer should be used to obtain optimum results. Consequently a comparison could be made between the theoretical and practical results.

The objects of the investigation described below were to attempt to find the detection mechanism at work in intensity-modulated displays, to establish the conditions for optimum efficiency of detection and to compare the thresholds obtained in visual detection with those of an optimum equivalent model.

In the case of the chemical recorder[4,5,8] and other two-dimensional intensity-modulated displays the observer is essentially looking for and recognizing the signal as an area, or a pattern, of different brightness (or density) from the surround. When the number of traces is increased the area occupied by the signal increases.

Considerable experimental evidence exists on the relation between stimulus area and visual contrast when the background is uniformly illuminated,[9] and it was suggested by one of the present authors[8] that the rate of improvement in threshold described above could be explained in terms of the depression of the visual contrast threshold with increase in area of the stimulus.

The experiments to be described later make use of a closed-circuit television system in order to simulate an intensity-modulated display such as a p.p.i. or chemical recorder. By this means the size and shape of the target point can be controlled at will, as also the mean intensity and mean-square fluctuation of the background. It was thus possible to determine the threshold for targets in a uniformly illuminated background and then by progressively increasing the fluctuation of the background by adding known amounts of noise, observe the effect of this on the threshold.

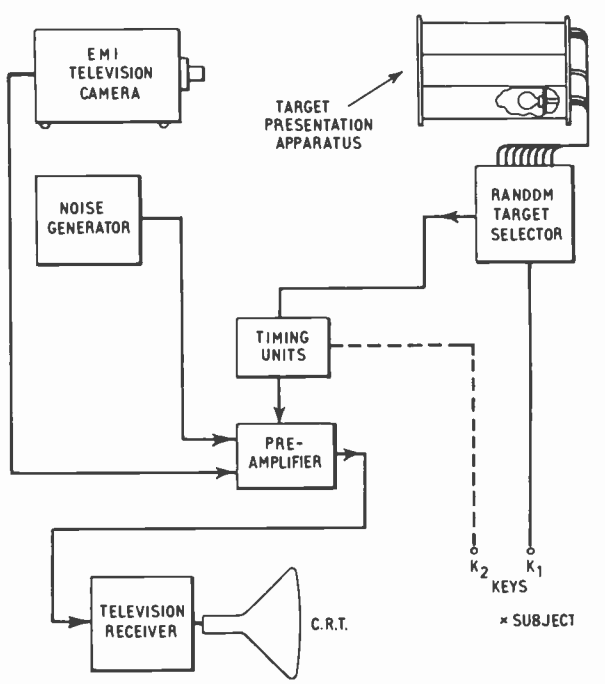


Fig. 1. Block diagram.

2. Experimental Equipment

A block diagram of the apparatus used is shown in Fig. 1. The majority of the apparatus is fairly standard and calls for little comment but the target presentation apparatus perhaps requires an explanation. Eight tubes each of 2 in. diameter were arranged parallel to one another, their centres forming the corners and side-centres of a square. Each tube had, at one end, a sheet of diffusing perspex and a slide holder to take 2 in. × 2 in. slides and could be illuminated by a lamp located at the opposite end of the tube. An opaque card or slide with the target of necessary size and shape cut on it was put into each slide holder, all lamp units being provided with identical cards.

The random target selector when triggered by key K1 connected one of the tube lamps to a supply so illuminating one of the targets for a fixed time. The subject was provided with an eight-position selector device and having decided at which position he considered the target to exist, depressed the appropriate key. This operation re-set the apparatus for the next trial and at the same time recorded a success on a counter if the subject's decision was correct. A trial counter was operated for each trial and it was thus a simple matter to determine the percentage of successes.

With this apparatus the experimenter can be his own subject, operating the equipment in the above manner since in each trial the position of the target is selected at random and he has no prior knowledge of its position.

A gamma-correcting circuit is included in the monitor and this enabled the law of detection to be varied.

A simple photometer using a 931A photo-multiplier was constructed so that the luminance signal/noise ratio could be obtained directly as well as the average background luminance.

3. The Measurement of Signal/Noise Ratio

The threshold signal/noise ratio is defined in terms of the voltage signal/noise ratio at the input of the demodulator. A "linear" demodulator was used but this was followed by a gamma-correcting circuit and owing to the non-linearity of the tube luminance/grid voltage relationship, the effective law of the demodulator could in fact be altered.

The luminance of the screen may be expressed as a factor of the demodulator output voltage by the equation

$$B = k_1 V^{\nu} \qquad \ldots\ldots(2)$$

where B is the screen luminance and k_1 and ν are constants. V is proportional to the amplitude R of the r.f. signal hence

$$B = k_2 R^{\nu} \qquad \ldots\ldots(3)$$

The luminance contrast is defined as

$$C = \frac{\text{change in mean luminance in the presence of a signal}}{\text{mean background luminance in the absence of a signal}}$$

which can be divided as below,

$$C = \frac{\text{change in mean luminance}}{\text{r.m.s. luminance fluctuation}} \times \frac{\text{r.m.s. luminance fluctuation}}{\text{mean background luminance}}$$

The first factor is the luminance signal/noise ratio (defined in the same manner as R_B) and the second factor (r) depends only on the power law ν. The relation between r and ν is calculated in Appendix 1 but since in fact the c.r.t. luminance/grid voltage curve cannot be represented exactly by a power law it was considered desirable to measure r directly.

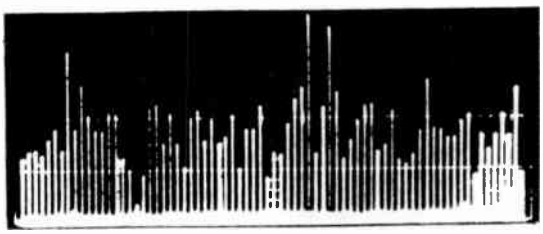


Fig. 2. Photograph of train of pulses.

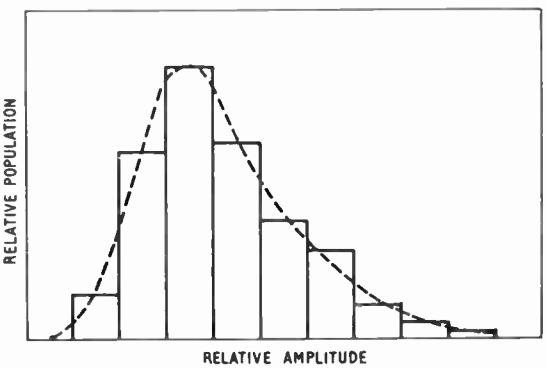


Fig. 3. Histogram of pulse amplitudes.

The method employed is essentially to find the r.m.s. value of the average brightness of a short length of the trace and from this and the type of noise spectrum, to determine the r.m.s. luminance of the spot. The experimental procedure is as follows. Using the frame time-base controls the television raster was distorted so that the first line of the raster appeared isolated from the rest by a distance of a few millimetres. A mask of black binding tape was used to cover the line except for a short length (5–10 mm) which was left exposed and in front of which was placed the photometer. The photometer was used with an integrating circuit whose time-constant was long compared with the time the line takes to cross this small gap, but short

compared to the interval between frames. Hence the magnitude of the photometer output just after the spot has passed the gap was proportional to the integral of the spot luminance over the interval of time which the spot takes to traverse the gap. Thus the output is in the form of a series of pulses at a p.r.f. of 25 c/s. Photographs of these pulses were taken using an oscilloscope with a slow time base and a statistical analysis made of the pulse amplitudes. A typical train of pulses is shown in Fig. 2 and the resulting histogram in Fig. 3.

4. Relation between Target Area and the Number of Samples

When a video signal pulse is long compared with the reciprocal of the receiver bandwidth, more than one independent sample can be taken within the pulse. To find the number of such samples, the following criterion will be used: the equivalent number of samples N is such that the variance of the mean receiver output over a period equal to the signal duration is the same as the variance of the mean of N independent samples. This point will now be elaborated.

Let the receiver have an r.f. bandwidth f_0 and a square-law envelope demodulator, so that the output consists of a voltage V_r proportional to the square of the envelope amplitude R_r. The optimum decision, as shown in Section 6, is based on the computation of the mean value of V_r. The detection of the signal is then governed by changes in and variances of the mean value. Instead of averaging discrete samples, let a continuous average be taken over a period T equal to the duration of the signal. Let

$$X = \frac{1}{T}\int_{t}^{t+T} V\,\mathrm{d}t \qquad \ldots\ldots(4)$$

X is a random variable, which by inspection, has the same mean value as that of independent samples. The variance of X can be calculated and if we equate this to the variance of the average of N independent samples of V, N can be determined.

The variance of X can be expressed either in terms of the auto-correlation function of $V(t)$ (ref. 6) or in terms of the spectrum of $V(t)$ (ref. 10), the two forms being equivalent.

$$V_{ar}(X) = \overline{X^2} - (\overline{X})^2 = \frac{1}{T^2}\int_0^\infty W(f)\frac{\sin^2\pi fT}{\pi^2 f^2}\,\mathrm{d}f$$

$$= \int_{-T}^{+T}\frac{T-|\tau|}{T^2}\phi(\tau)\mathrm{d}\tau \qquad \ldots\ldots(5)$$

where $W(f)$ is the power spectrum of $V(t)$ and $\phi(\tau)$ its auto-correlation function. The first of these expressions will be used, as the power spectrum of the receiver output is known. For a narrow-band receiver of bandwidth f_0 and a white noise input, the law of the frequency spectrum is a triangular one which can be represented as

$$\left.\begin{aligned} W(f) &= A_1\left(1-\frac{f}{f_0}\right) && 0<f<f_0 \\ W(f) &= 0 && \text{elsewhere}\end{aligned}\right\} \qquad \ldots\ldots(6)$$

In the television receiver used, the spectrum is limited by a post-detection filter to half the above width and thus

$$\left.\begin{aligned} W(f) &= A_2\left(1-\frac{f}{f_0}\right) && 0<f<\frac{f_0}{2} \\ &= 0 && \text{elsewhere}\end{aligned}\right\} \qquad \ldots\ldots(7)$$

These will be referred to as Case I and Case II and the variance of X will be evaluated for both.

Case I

$$\mathrm{Var}\,X = \frac{A_1}{T^2}\int_0^{f_0}\left(1-\frac{f}{f_0}\right)\frac{\sin^2\pi fT}{\pi^2 f^2}\,\mathrm{d}f$$

$$= \frac{A_1}{\pi T}\left\{\mathrm{Si}(2\pi f_0 T) + \frac{1}{2\pi f_0 T}\mathrm{Ci}(2\pi f_0 T) - \right.$$

$$\left. - \frac{\sin^2\pi f_0 T}{\pi f_0 T} - \frac{1}{2\pi f_0 T}\log_e(\gamma.2\pi f_0 T)\right\} \qquad \ldots\ldots(8)$$

Case II

$$\mathrm{Var}\,X = \frac{A_2}{\pi T}\left\{\mathrm{Si}(\pi f_0 T) + \frac{1}{\pi f_0 T}\mathrm{Ci}(\pi f_0 T) - \right.$$

$$\left. - \frac{2\sin^2\frac{\pi f_0 T}{2}}{\pi f_0 T} - \frac{1}{\pi f_0 T}\log_e(\gamma.\pi f_0 T)\right\} \qquad \ldots\ldots(9)$$

where $\mathrm{Si}(\theta)$ and $\mathrm{Ci}(\theta)$ are the sine and cosine integrals and $\gamma = 1{\cdot}7810$ (approx) and use has been made of the relation[11]

$$\mathrm{Ci}(x) = \log_e(\gamma x) - \int_0^x \frac{1-\cos t}{t}\,\mathrm{d}t$$

In both the above cases, for large values of T, all terms except the first one become negligibly small so that

$$\mathrm{Var}\,X = \frac{A_1}{\pi T}\times\frac{\pi}{2} = \frac{A_1}{2T} \qquad \ldots\ldots(10)$$

The mean square value of independent samples is the area under the spectral distribution curve and the variance of the sum of N samples is this quantity

divided by N. Equating this to eqn. (10) in the two cases, we get:

Case I

$$\frac{1}{2}\frac{A_1 f_0}{N} = \frac{A_1}{2T}$$

or $N = f_0 T$

Case II

$$\frac{3}{8}\frac{A_2 f_0}{N} = \frac{A_2}{2T}$$

or $N = \frac{3}{4} f_0 T$ .....(11)

The general assumption that the number of samples in time T is $f_0 T$ is confirmed in the first case and a slightly reduced value due to a different spectrum is obtained in the second case. It is interesting to note that in the case of a uniform power spectrum by similar reasoning, one gets

$$N = 2Tf_0 \qquad .....(12)$$

which corresponds to the number of samples which may be expected from considerations of the sampling theorem.

With a receiver bandwidth of 6 Mc/s, the number of samples per second corresponding to Case II is $4{\cdot}5 \times 10^6$, so that in one line scan of 81 ms there are 365 independent samples. The picture area corresponding to one line is 53·9 mm² and therefore the number of samples per mm² is about 6·8. As the complete target area is scanned in two interlaced frames, the number of samples per mm² of target area per unit time is $6{\cdot}8 \times 25 = 170$. Thus to get the number of samples presented in T_0 seconds one uses the equation

$$N = 170 \,.\, A \,.\, T_0 \qquad .....(13)$$

It must be noted, however, that the equations (11) which have been used to obtain this apply only when the duration of individual video pulses is long, due to the approximation made to eqn. (9). However, as the error is not appreciable even for relatively short durations, the approximate formula has been used to calculate the number of independent samples for all the target sizes which will be used in the experimental work. These numbers are given in Table 1.

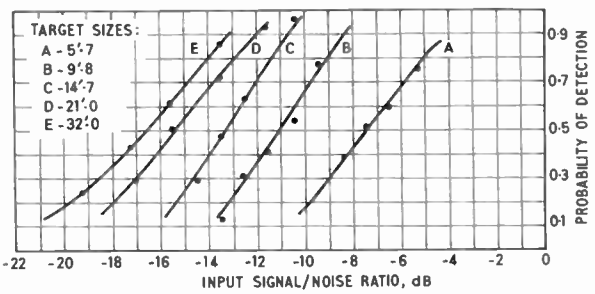
Table 1

Number of independent samples for each target area

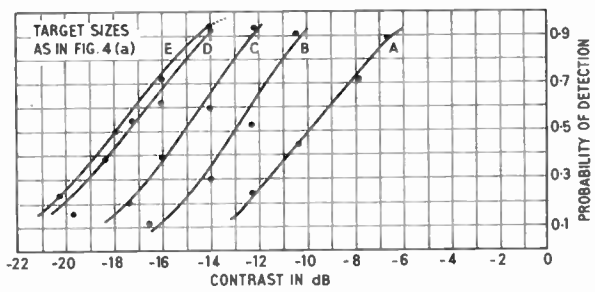
Angular width of target	Area on the screen mm²	Presentation time	Number of samples
5·7	2·2	8 s	3 000
9·8	6·5	8 s	8 850
14·7	14·8	8 s	20 100
21·0	28·5	8 s	39 000
32·0	71·1	8 s	97 000
21·0	28·5	40 ms	195
21·0	28·5	80 ms	390
21·0	28·5	160 ms	780
21·0	28·5	320 ms	1 560
21·0	28·5	640 ms	3 120

5. Experimental Results

The first object of the experiments described here was to find the manner in which the threshold contrast and the threshold signal/noise ratio vary with the area of the target at different luminance levels and to compare the results with those obtained by Blackwell[9] under similar conditions. The second was to investigate the effect of increasing the non-linearity of the relation between the input signal amplitude and the output luminance i.e. of increasing the value of ν. Such a change has two effects on the output luminance. The increase of non-linearity increases the change of mean for a given input signal/noise ratio, i.e. it increases the contrast, tending to make the target more visible. At the same time, it increases the luminance fluctuations, off-setting the improvement in contrast. Under these conditions, an optimum might exist which could be found from these experiments. In the practical operational situation, changes in non-linearity are brought about by the adjustment of the bias of the c.r.t. and the video gain which are



(*a*) Noise background.



(*b*) Uniform background

Fig. 4. Frequency-of-seeing curves for 0·1 ft-lambert (Best visual fit to sigmoid curves.)

manipulated by the radar operator. The study of the effect of non-linearity may be expected to give a logical basis for an optimum adjustment.

A further object of the experiments was to find the effect of the shape of the target on its visibility, but in fact only one set of measurements were done with other than a circular target. In this experiment a slit target was used and the results showed clearly that shape is an important factor in detectability. Work is progressing on this aspect and it is hoped that further results will be available later.

5.1. *Effect of Target Area on Detectability*

Tests were carried out at three background luminance levels of 1·0 ft-lambert (high), 0·1 ft-lambert (medium) and 0·01 ft-lambert (low), with a nearly linear input/brightness law. Circular targets of angular diameters ranging from 5·7 min to 32 min were used. Threshold contrast under noise-free conditions were also determined at the three brightness levels. In all the three cases, the input threshold signal/noise ratio decreased as the areas increased but there are quantitative differences between them.

Tests at 0·1 *ft-lambert:* Five target sizes were used, having angular diameters of 5·7, 9·8, 14·7, 21 and 32 minutes. The frequency-of-seeing curves are given in Fig. 4(*a*) for the noisy background and in Fig. 4(*b*) for the uniform background. The two sets of curves appear similar and when the threshold signal/noise ratios for 50% detection are converted to contrasts, it is found that the difference between them is very small (less than 0·5 dB).

In Fig. 5, the threshold contrasts are plotted as a function of the area of the target and Blackwell's results are also indicated. It is evident from this figure that the nature of the variation of the threshold with area is practically the same as in the case of Blackwell's experiments. The variation of the threshold signal/noise ratio with area is about 2·3 dB per

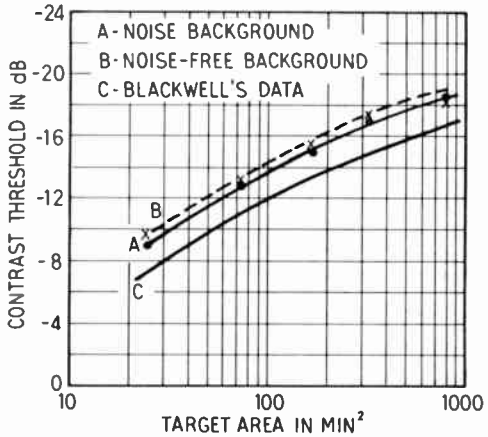
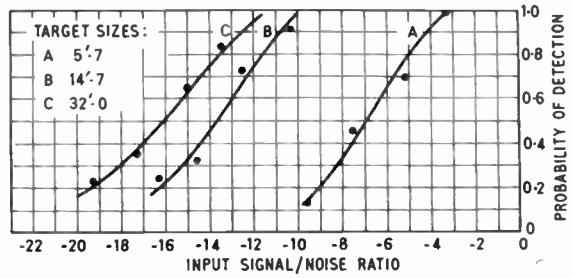
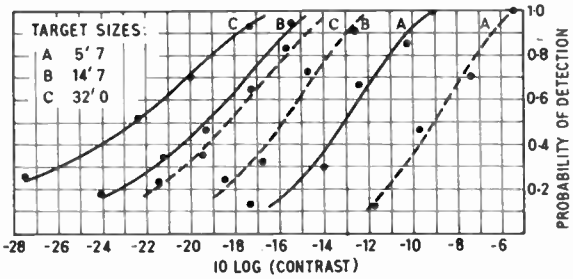


Fig. 5. Threshold data 0·1 ft-lambert.



(*a*) Noise background.



(*b*) Uniform background.

Fig. 6. Frequency-of-seeing curves for 1·0 ft-lambert.

doubling for small areas and about 1·3 dB for large areas. It is to be noted that the increase in the threshold contrast due to the fluctuating background is only about 0·5 dB, though, to the subject, the difference in the nature of the background is clearly visible. A further point which can be observed is that the threshold contrasts found in these experiments are consistently lower than those found by Blackwell. The probable reasons for this will be discussed later.

Tests at 1·0 *ft-lambert:* Tests were carried out with three circular targets of angular diameter 5·7, 14·7 and 32 minutes. In Figs. 6(*a*) and (*b*), the frequency-of-seeing curves are given for the tests with noise background and uniform background respectively. The abscissa in the former case is the input signal/noise ratio. When converted to contrast (by adding −2·3 dB to the signal/noise ratio), it is found that there is a difference of about 4 dB in all the three cases between the thresholds with and without noise. It is to be noted that the frequency-of-seeing curves are similar in shape in the two cases and the effect of noise in the background is to shift the curve towards the higher contrast side. In Fig. 7, the threshold contrasts for 50% detection are plotted along with Blackwell's results for comparison. Though the two curves are similar, they are less so than at 0·1 ft-lambert. In terms of the input signal/noise ratio, the depression of the threshold is about 2·6 dB/doubling for small area and less than 1·0 dB/doubling for large ones.

Experiments were carried out with a background made up partly of noise and partly of uniform

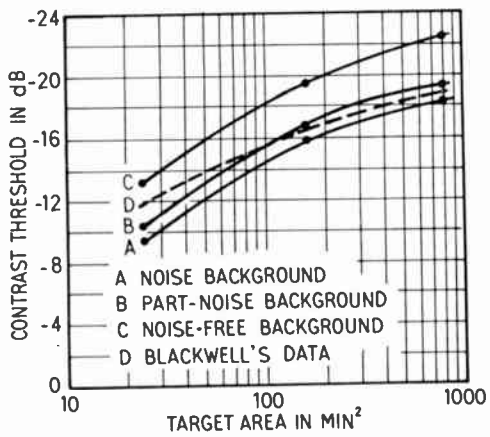


Fig. 7. Threshold data 1·0 ft-lambert.

luminance. The threshold contrasts obtained in these cases were intermediate between the values for noise-free and noise backgrounds dealt with earlier. These data will be discussed separately in reference 13.

The threshold signal/noise ratios at 1·0 ft-lambert are slightly different from those at 0·1 ft-lambert, being about 0·3 dB less in one case, and about 0·5 dB higher in the other two cases.

Tests at 0·01 *ft-lambert:* These tests were carried out with three targets of the same size as in the case of 1·0 ft-lambert. The reduction of luminance was accomplished by interposing a neutral filter before the television screen, the luminance of which was set at 0·1 ft-lambert. At this lower luminance, the noise fluctuations are not visible at the viewing distance of one metre, indicating that the resolution of the subject's eye is below the angular separation between noise peaks. The frequency-of-seeing curves are shown in Fig. 8. Only one test (for the 14·7 min target) was done with a uniform background and the frequency-of-seeing curve obtained is also shown in Fig. 8 curve D. When the signal/noise ratios are converted to contrast (by adding −2·6 dB), it is seen by comparing the curves D and B that the threshold contrast differs by about 0·2 dB from that in the case of uniform background. This is what might be expected as there is no real difference visually between these two cases. The closeness of the two results merely confirms that the two different methods of making measurements of contrast give the same results. In Fig. 9 the threshold contrast is plotted against the target area, with Blackwell's results for comparison. As was observed for the 0·1 ft-lambert results, there is a close correspondence between the shape of the two curves, although there is an appreciable difference between the absolute contrasts in the two cases.

As would be expected, the threshold signal/noise ratios are considerably higher than those at 0·1 ft-lambert and 1·0 ft-lambert. The rate of change with increase in area varies from 2·3 dB/doubling for small areas to 1·5 dB/doubling for large areas.

In the experiments with noise-free background, the threshold contrasts obtained are lower than those of Blackwell by 1 to 2 dB at all the three luminance levels. There may be several reasons for this. The conditions of the test are slightly different—the presentation time is longer and the angular separation of the target positions smaller than in Blackwell's case. The spectrum is also different in the two cases. Further, in Blackwell's tests, the contrast data are the average for a number of subjects whereas in these experiments they pertain to one subject. But perhaps the most important difference is that in Blackwell's experiments, the background illumination is continuously present, whereas in experiments with a c.r.t. it is made up of a moving spot of high intensity which occurs for a short time in each element of area. It is possible that under these conditions the effective average luminance is different from the time average of the spot luminance and the threshold contrasts correspondingly so.

5.2. *Effect of Non-linearity of the C.R.T. Brightness Law*

Threshold contrasts were determined at two non-linearities corresponding approximately to square-law and cubic-law relations between the signal input

Table 2

Summary of results for the three different brightness levels

Target diameter	Brightness level 1·0 ft-lambert		0·1 ft-lambert		0·01 ft-lambert	
min.	S/N ratio dB	Contrast dB	S/N ratio dB	Contrast dB	S/N ratio dB	Contrast dB
5·7	− 6·8	− 9·1	− 7·3	− 8·9	− 2·5	− 5·1
9·8			− 11·1	− 12·7		
14·7	− 15·8	− 15·8	− 13·3	− 14·9	− 9·1	− 11·7
21·0			− 15·3	− 16·9		
32·0	− 16·1	− 18·3	− 16·7	− 18·3	−12·7	− 15·3

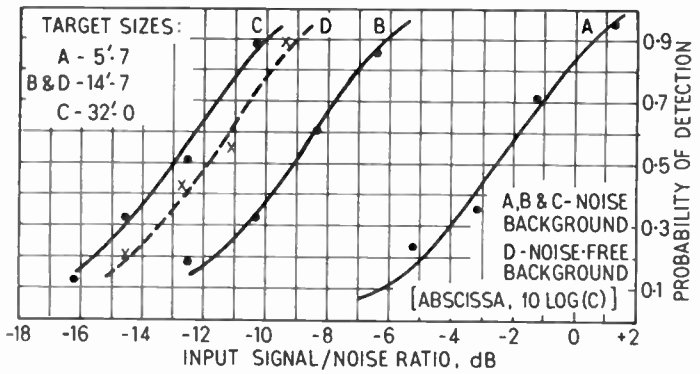


Fig. 8. Frequency-of-seeing curves for 0·01 ft-lambert.

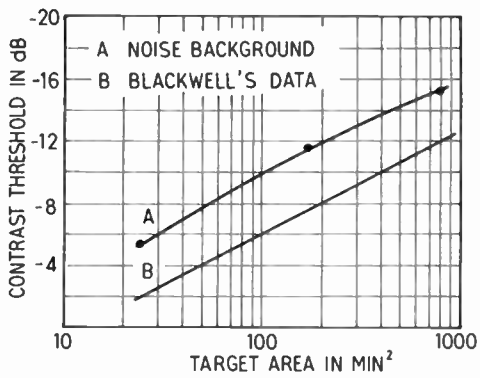


Fig. 9. Threshold data for 0·01 ft-lambert.
(A) Three target sizes at 0·1 ft-lambert
(B) Three luminance levels for 14·7′ target.

and luminance output. The measurements were made for three target sizes at 0·1 ft-lambert and for one target size (14·7 min) at the other two luminance levels.

The effect of increased non-linearity was found to be mainly the shifting of the frequency-of-seeing curve towards higher contrast, without any significant change in its slope. Further, the measurements at 0·1 ft-lambert showed that the shift is nearly the same for the three target sizes with angular diameters in the range 5·7 to 32 min. The changes in the threshold signal/noise ratio are considerably smaller than the changes in contrast and the magnitude and direction depend upon the mean luminance. The results are given below and the actual thresholds are shown in Table 3.

Table 3

The effect of non-linearity on contrast threshold

Luminance ft-lambert	"Square" law		"Cubic" law	
0·1	5·7′	— 8·3 dB	5·7′	— 6·3 dB
0·1	14·7′	— 13·5 dB	14·7′	— 12·0 dB
0·1	32·0′	— 17·8 dB	32·0′	— 15·4 dB
1·0	14·7′	— 12·8 dB	14·7′	— 13·0 dB
0·01	14·7′	— 11·2 dB	14·7′	— 12·2 dB

Mean luminance of 1·0 ft-lambert: Increase of non-linearity to square and cubic worsens the threshold signal.

Mean luminance of 0·1 ft-lambert: With square-law relation the threshold signal/noise ratio is lowered but with a cubic-law relation, it becomes higher even than for the linear-law case.

Mean luminance of 0·01 ft-lambert: Increase of non-linearity improves the threshold even up to the cubic law. It was also noted that at these higher non-linearities the luminance noise fluctuations become visible.

These results are made more evident in the curves drawn in Fig. 10(*a*) and 10(*b*). The abscissa in these graphs is the r.m.s. luminance fluctuations in 1 mm² of the screen (i.e. 11·9 mm² of solid angle) as a ratio of the mean luminance which will be called contrast in noise ($B_{r.m.s.}$). This quantity is related to the input voltage brightness law and has been directly measured (Appendix 1). The ordinate in the above figures is the threshold contrast. A smooth curve has been drawn connecting all the points including the one for uniform background when, of course, $B_{r.m.s.} = 0$. The slope of a line joining any point on the curve to the origin gives the ratio of the change of mean to contrast in noise. This latter quantity is a function of r the ratio of r.m.s. to mean luminance and if the spectrum does not change appreciably with non-linearity, it can be shown that $B_{r.m.s.}$ is proportional to r. Thus the slope of the line is proportional to the output luminance signal/noise ratio, which in turn is equal to the input electronic signal/noise ratio for low input signals. Expressing these relations mathematically we have

$$\text{threshold contrast } C = \frac{\Delta B}{B}$$

$$\text{contrast in noise} = \frac{\text{luminance fluctuation in 1 mm}^2}{B}$$

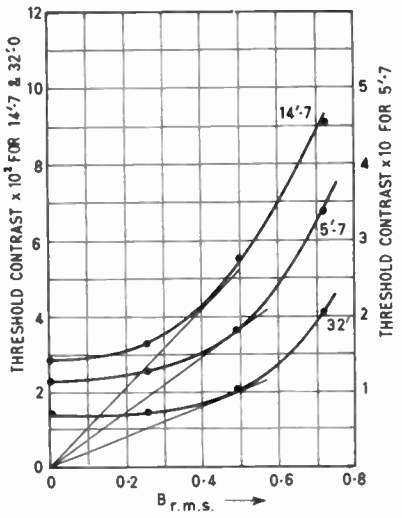
$$\text{slope of line joining point on curve to origin} = \frac{\text{threshold contrast}}{\text{contrast in noise}}$$

$$= \frac{\Delta B}{\text{r.m.s. luminance fluctuation}}$$

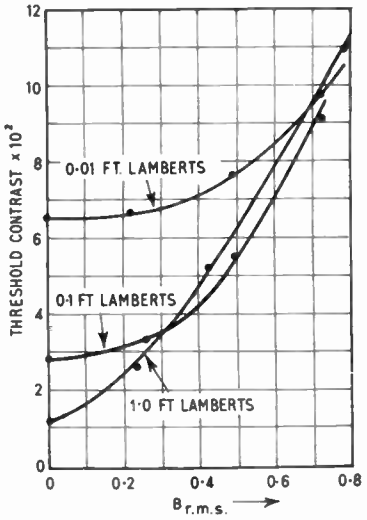
$$= k \text{ (output luminance S/N ratio)}$$

$$= k \text{ (input electrical S/N ratio)}$$

It is seen from Figs. 10(*a*) and (*b*) that at the higher brightness levels there is a minimum value which the



(*a*) Contrast thresholds for three target sizes at 0·1 ft-lambert.



(*b*) Contrast thresholds at three luminance levels for 14·7′ target.

Fig. 10. Contrast thresholds vs $B_{r.m.s.}$.

slope of the line can have for a given area and luminance, corresponding to the slope of the target to the curve passing through the origin. There is therefore a minimum signal/noise ratio which can be attained. In the case of the lowest luminance, the shape of the curve indicates that there may be a similar minimum for a higher non-linearity than the cubic law. The non-linearity governing the conditions corresponding to the point on the curve at the tangent may be regarded as optimum. In Fig. 10(*a*), these tangential lines are drawn for the three curves (for the three target areas) and it can be seen that the optimum non-linearity is nearly the same for the three sizes. On the other hand, the optimum non-linearity is not the same for the three luminance levels, and has a value approximately of $\nu = 1{\cdot}5$ at 1·0 ft-lambert, $\nu = 2{\cdot}1$ at 0·1 ft-lambert and higher than $\nu = 3$ at the lowest luminance level. It should be remarked, however, that these values cannot be taken to be very precise as the threshold contrasts may be in error by $\pm 0{\cdot}3$ dB. Errors of this magnitude will however not make any qualitative changes in the conclusions drawn. It is also to be noted that the optimum non-linearity is not very critical, as the rate of change of input signal/noise ratio with non-linearity is very slow near the optimum point.

Examination of the curves in Fig. 10(*b*) shows that, as the non-linearity is increased, the threshold contrast approaches a nearly constant value for all the three luminance levels, and is roughly proportional to the contrast in noise. In this region, it would appear that the threshold contrast is governed mainly by the external noise (i.e. luminance fluctuations) and this suggests that the external noise is large compared with the internal noise of the human vision channel.

It is also found that the lowest input signal/noise ratio under optimum conditions is obtained at 0·1 ft-lambert. This is a qualitative difference from the behaviour under noise-free conditions, where the threshold contrast always decreases with the increase of background illumination.

6. Comparison of Visual Detection with an Optimum System

In this Section, the threshold signal/noise ratios obtained in the experiments described in the last section are compared with the thresholds predicted for a theoretical optimum model, the "ideal observer", to establish how near the performance of the visual channel is to the ideal. As a preliminary to this, a digression on statistical decision theory as applied to the ideal observer is necessary.

6.1. *The Ideal Observer Criterion*

The detection of an echo-ranging signal in noise is a decision process in which the observer chooses one of two mutually exclusive hypotheses, namely that the given data belong to the class "signal+noise" or to the class "noise alone". The acquired data are generally a set of amplitudes of r.f. envelope at a fixed range, taken in a number of successive traces. The pulse repetition period is invariably very large when compared with the reciprocal of the bandwidth of the radar receiver, so that the samples may be regarded as being statistically independent. The decision made is liable to two types of errors, the type I error of calling noise alone a "signal and noise" (false alarm), and the type II error of deciding "noise only", when a signal as well as noise is actually present (missed signal). Each type of error may be associated with a cost, and on the basis of these and the *a priori* prob-

abilities of signal being present, the average risk associated with the decisions may be computed (ref. 7, p. 804). The optimum decision rule is one which minimizes the average risk, and it can be shown that in all such rules, (called the Baye's decision rules), the decision is based on the "likelihood ratio" Λ, which is defined by the formula

$$\Lambda = \frac{\text{probability of the acquired samples belonging to "noise + signal"}}{\text{probability of the acquired samples belonging to "noise alone"}} \quad \ldots\ldots(14)$$

There are several decision criteria which are optimum in this sense but which differ in respect of cost assumptions. For example, in the "Neyman-Pearson criterion", the probability of the type I error is held constant and the probability of the type II error is minimized. In the "ideal observer criterion", the total probability of the two types of errors is minimized. We are concerned with the latter in all the subsequent discussion, since as indicated later, the conditions of visual test conform closely to this.

A special application of the ideal observer criterion arises in tests of the "forced-choice" method. In these tests, the signal is present in one of a number (say m) of sets of data and the decision to be made is as to which set contains the signal. Lawson and Uhlenbeck (Ref. 1, p. 171) have dealt with this case and have shown that the optimum decision is the selection of the set with the highest value of Λ. To find the probability of correct detection and related quantities, Λ and its probability distribution will now be computed.

An expression for Λ can be developed from the probability distributions of envelope amplitude R of the r.f. signal. Assuming narrow-band Gaussian noise of mean power ψ_0 and a sinusoidal signal of amplitude P, the probability densities of amplitude when the signal is present (p_s) and when only noise is present (p_n) are given by

$$\left.\begin{aligned} p_s(R) &= \frac{R}{\psi_0}\exp\left[-\frac{R^2+P^2}{2\psi_0}\right]\mathrm{I}_0\left(\frac{RP}{\psi_0}\right) \\ \text{and } p_n(R) &= \frac{R}{\psi_0}\exp\left[-\frac{R^2}{2\psi_0}\right] \end{aligned}\right\} \quad \ldots\ldots(15)$$

where $\mathrm{I}_0(z)$ is the modified Bessel function of zero order.

When N samples are taken, these being uncorrelated, the probability of the set belonging to one of the above distributions is given by the product of individual probabilities. Substituting these values in eqn. (14),

$$\Lambda = \frac{\prod_{r=1}^{N} p_s(R_r)\mathrm{d}R_r}{\prod_{r=1}^{N} p_n(R_r)\mathrm{d}R_r} = \exp\left[-\frac{NP^2}{2\psi_0}\right]\prod_{r=1}^{N}\left\{\mathrm{I}_0\left(\frac{R_rP}{\psi_0}\right)\right\} \quad \ldots\ldots(16)$$

It is more convenient to use $\log\Lambda$ and it is possible to do so as $\Lambda > 0$ and $\log\Lambda$ is a monotonic function of Λ. Taking the logarithm on the right-hand side of eqn. (16)

$$\log\Lambda = -\frac{NP^2}{2\psi_0} + \sum_{r=1}^{N}\log\left\{\mathrm{I}_0\left(\frac{R_rP}{\psi_0}\right)\right\} \quad \ldots\ldots(17)$$

The first term on the right-hand side is independent of sample amplitudes. Therefore, the optimum decision reduces to the computation of $\log(\mathrm{I}_0)$ for each of the possible sets of samples, and choosing that set which has the highest value of $\Sigma\log(\mathrm{I}_0)$. The probability of correct detection for a given input signal/noise ratio can be calculated by finding the probability distribution of $\log\Lambda$ when signal is present and when only noise is present.

Some approximations are possible when the number of samples is large and the input signal is small. For small values of ρ

$$\text{i.e. } \frac{P}{\sqrt{2\psi_0}} \ll 1$$

$$\mathrm{I}_0\left(\frac{R_rP}{\psi_0}\right) \simeq 1 + \frac{R_r^2P^2}{4\psi_0^2} \simeq \exp\left(\frac{R_r^2P^2}{4\psi_0^2}\right) \quad \ldots\ldots(18)$$

Substituting in (17),

$$\log\Lambda = -\frac{NP^2}{2\psi_0} + \frac{P^2}{4\psi_0^2}\sum_{r=1}^{N}R_r^2 \quad \ldots\ldots(19)$$

It is convenient to change the variable to

$$\frac{1}{N}\sum R_r^2 = q = 2\psi_0 + \frac{4\psi_0^2}{NP^2}\log\Lambda \quad \ldots\ldots(20)$$

All the terms in eqn. (20) are positive and so the maximum value of $\log\Lambda$ corresponds to the maximum value of q. The optimum decision is thus the choice of the highest value of q, among the sets of amplitude samples available. For large N, q has a Gaussian distribution, the mean values and variances being

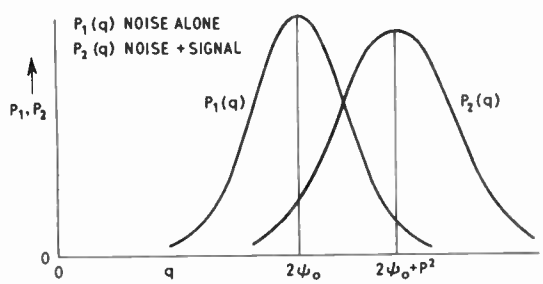


Fig. 11. Probability distributions of q.

different when the set of samples contains signal and when it contains noise only. These values (Lawson and Uhlenbeck, Ref. 1, p. 172) are:

$$\left.\begin{aligned} &\text{A}\quad \text{Noise only (distribution } p_1(q)) \\ &\qquad \text{mean value} = 2\psi_0 \\ &\qquad \text{variance} \quad = \frac{4\psi_0^2}{N} \\ &\text{B}\quad \text{Noise+signal (distribution } p_2(q)) \\ &\qquad \text{mean value} = 2\psi_0 + P^2 \\ &\qquad \text{variance} \quad = \frac{4\psi_0^2}{N}\left(1+\frac{P^2}{\psi_0}\right) \end{aligned}\right\} \quad \text{.....(21)}$$

These distributions are illustrated in Fig. 11. In the m-position experiments, one position (containing signal+noise) has the value of q following the distribution $p_2(q)$ and the remaining m-1 positions have values of q following the distribution $p_1(q)$. For correct detection the latter must be less than the former. For a given value y of the former, the probability that the latter do not exceed it is

$$\left[\int_0^y p_1(q)\,\mathrm{d}q\right]^{m-1}$$

Taking all possible values of y from 0 to ∞, the gross probability of correct detection is given by

$$W_m = \int_0^\infty p_2(y)\left[\int_0^y p_1(q)\,\mathrm{d}q\right]^{n-1}\mathrm{d}y \quad \text{.....(22)}$$

Allowing for chance success, the probability of correct detection is

$$p = \frac{mW_m - 1}{m-1} \quad \text{.....(23)}$$

substituting the expressions for $p_1(q)$ and $p_2(q)$ and making the assumption that the variances of the two distributions are the same, we obtain

$$W_m = \frac{1}{2^{m-1}\sqrt{\pi}}\int_{-\infty}^{+\infty} \mathrm{e}^{-x^2}\left[1+\mathrm{erf}\left(x+\frac{\sqrt{N}P^2}{2\sqrt{2}\psi_0}\right)\right]^{m-1}\mathrm{d}x \quad \text{.....(24)}$$

This integral can be evaluated in a closed form for $m=2$, giving

$$W_2 = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\sqrt{N}P^2}{2\sqrt{2}\psi_0}\right)\right] \quad \text{.....(25)}$$

and

$$p = \mathrm{erf}\left(\frac{\sqrt{N}P^2}{2\sqrt{2}\psi_0}\right) \quad \text{.....(26)}$$

For $m=8$, the integral has been evaluated by numerical methods for 12 values of $\frac{\sqrt{N}P^2}{2\psi_0}$, and "betting curves" have been drawn for the two values of m in Fig. 12. The abscissa in these curves is $10\log\frac{\sqrt{N}P^2}{2\psi_0}$, i.e. the input signal/noise ratio plus a constant (5 log N).

The principal point to be noted here is that the optimum statistic is an average based on squared amplitudes and that the decision process reduces to the choice of one of two possible Gaussian distributions of this average. Further, the detectability depends upon $P^2/2\psi_0$, the output signal/noise ratio with a square-law demodulator. It follows from this that the decision based on other averages (e.g. the average of amplitudes, or of some other power of amplitudes) will be sub-optimum, but to the extent that the output signal/noise ratio is independent of the demodulator law, any demodulator obeying a power-law will be nearly optimum.

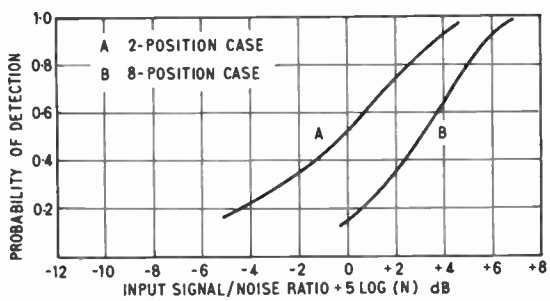


Fig. 12. Theoretical betting curves for the ideal observer.

6.2. *Comparison of Visual Thresholds with the Ideal Observer*

The manner of conducting the subjective tests adopted here conforms to the conditions stipulated for the ideal observer in the above analysis, because the signal is present in each trial with equal *a priori* probability in one of eight possible positions and the decision to be made is the position of the signal. Comparison of the experimental results with the thresholds for the eight-position ideal observer tests is therefore valid. In Section 4, eqn. (13) the number of independent samples in a target of area A mm^2 presented for T_0 seconds was shown to be

$$N = 170\,AT_0$$

Using this value of N, the threshold signal/noise ratio for the ideal observer can be found from Fig. 12. In Fig. 13, the threshold signal/noise ratios determined in the tests for circular targets (v = 1·33) are plotted against the number of samples. It is evident that the visual thresholds fall considerably short of the ideal observer even when the difference is least.

For a target subtending about 20 min the threshold signal/noise ratio approaches nearest to that of the ideal observer and is about 4·5 dB worse. For smaller

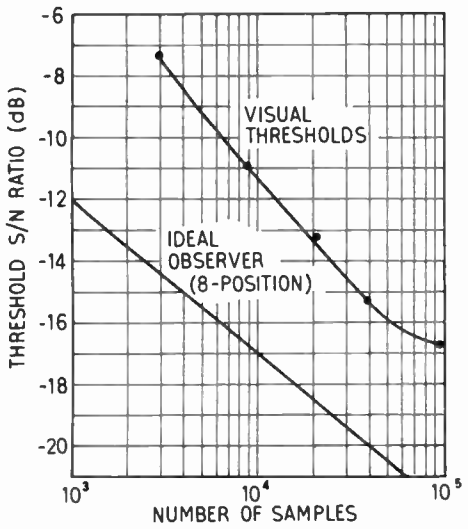


Fig. 13. Comparison of experimental and theoretical results for the 8-position test.

and larger targets, the differences are greater. The relatively high rate of improvement (up to 2·3 dB/doubling would therefore appear to be due to the fact that for small targets, the system is more sub-optimum than for the 20 min one. Under more favourable conditions, (e.g. with a slit target, and optimum non-linearity), the least difference would have been somewhat less (about 3 dB). It should be remembered however that this difference occurs even though the number of samples is very large, i.e. about 40 000.

The reasons for this difference may now be considered. The ideal observer criterion assumes a perfect integrating mechanism. The human eye may be expected to fall short of this on account of at least two factors, firstly the internal noise of the vision channel and secondly the inability to integrate over a large area for a long time. So far as integration over space is concerned, the maximum use has presumably been made of this faculty as we are considering the optimum size. But so far as integration in time is concerned, however, the period of 8 seconds may be expected to be well above the optimum, particularly as the use of eight positions necessitates eye movements. To establish how much these factors did affect the results some experiments were conducted using only two target positions and relatively short presentation times. The two positions were 4·4 min apart and a fixation mark was provided in the centre, in addition to the incomplete crosses at the targets positions. The presentation times used were 40, 80, 160, 320 and 640 ms. About 350 trials were made in each case for each signal/noise ratio. The frequency-of-seeing curves are shown in Fig. 14, and the threshold data are compared with the ideal observer criterion in Fig. 15.

It can be seen that the experimental thresholds approach nearest to the ideal observer performance for a presentation period of about 320 ms, but the difference between the ideal and visual thresholds is still about 4·5 dB. The shorter presentation time has therefore not helped to reduce this difference, and though this may be partly due to the fact that a new viewing technique (i.e. fixation instead of search) was used and partly to inefficient spatial integration, it would appear that there must be other factors than the integration time to account for this difference.

Though the visual system is sub-optimum, it must be noted that the difference of 4·5 dB occurs in spite of the large number of samples. To obtain an idea of the relative magnitude of this difference, one might examine a different type of sub-optimum system. A decision criterion may be formulated thus: a threshold level is set for a fixed false alarm rate, and a decision that a signal is present is made if at least one of the

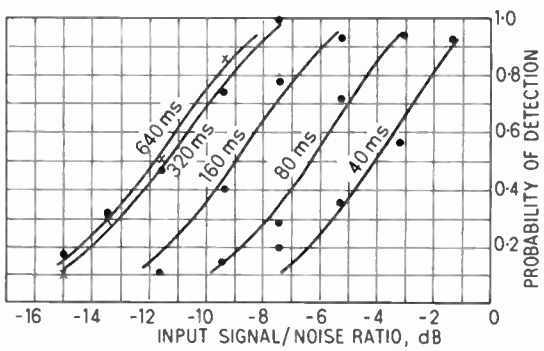


Fig. 14. Frequency-of-seeing curves for 2-position test.

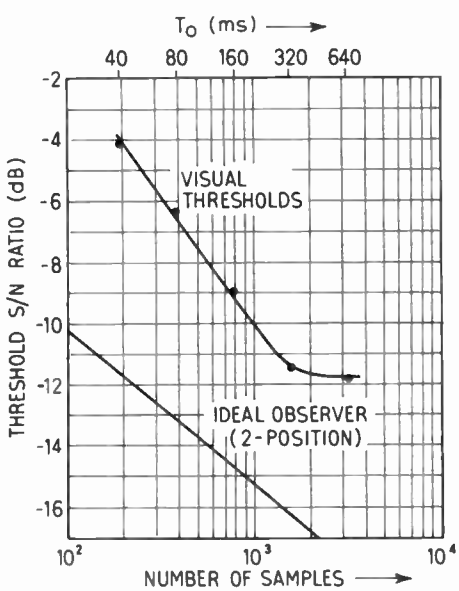


Fig. 15. Comparison of experimental and theoretical results for 2-position test.

group of N samples exceeds the threshold. In this case averaging of the samples is not done, but the decision is based on the individual sample amplitudes. This type of sub-optimum system has been analysed and compared with an optimum (Neyman-Pearson) system by Middleton (Ref. 7, p. 909). For the same false alarm rate, the sub-optimum system is worse by an order of magnitude when compared with the optimum system. For values of $N = 10$, 100 and 1000, the minimum detectable signals are worse by 6·8, 10·2 and 13·6 dB respectively.

It would appear from this that the visual detection system is only moderately sub-optimum and as such may be expected to employ an averaging faculty in the detection of intensity modulated signals.

7. Effect of Weighted Integration

The experiments have shown that visual detection in intensity-modulated displays utilizes an integrating mechanism and consequently, under favourable conditions, can approach the performance of optimum systems. The integrating faculty does not appear to interfere with the faculty of detail discrimination. Since, at the medium luminance, it was found that the contrast required for detection in the presence of noise was not appreciably different from the contrast required under noise-free conditions, though the luminance fluctuations were clearly visible. This suggests that the locus of the integrating mechanism is more central than that of detail discrimination or operates parallel to it.

As remarked earlier, there is a particular area for which integration seems to be most effective. For areas smaller than this, the threshold is relatively high. One possible reason for this is a mismatch between the area over which integration takes place and the target area. If the former is constant and independent of target area, integration is inefficient for targets smaller than this area, because a number of noise pulses are integrated along with those containing the signal. The signal/noise ratio of the integrated output is therefore reduced and the threshold of detection consequently raised. We will examine this possibility quantitatively by calculating the output signal/noise ratio under these conditions, and assuming that the detectability depends only on this quantity. Two different averaging processes will be considered.

Let us suppose that a circular area containing N_1 independent picture elements is averaged irrespective of the size of the target and that the number of elements in the latter is N_2 ($N_2 < N_1$). Let ΔB be the change in the luminance in the target area, contained within the larger averaging area. The change in mean over the whole of the area is then $\Delta BN_2/N_1$. The

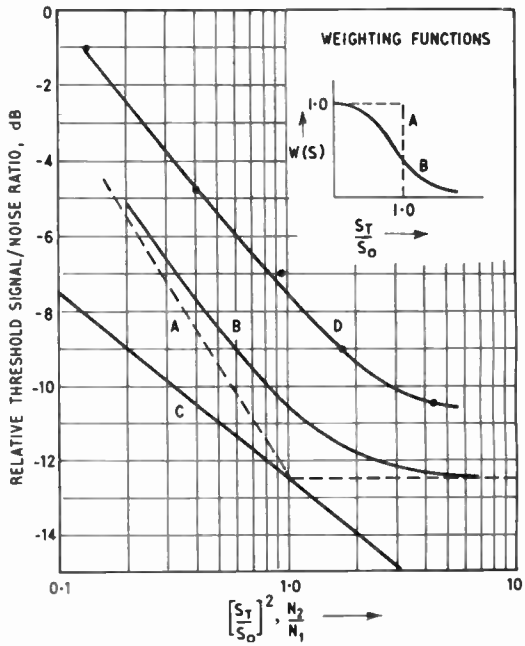


Fig. 16. Effect of mismatch of integration area.

r.m.s. fluctuation of the average of N_1 elements is $\frac{rB}{\sqrt{N_1}}$. Detection takes place when

$$\frac{\Delta BN_2}{N_1} = K\frac{rB}{\sqrt{N_1}}$$

or

$$\Delta B \Big/ \frac{rB}{\sqrt{N_1}} = (KN_1)\frac{1}{N_2} \qquad \text{.....(27)}$$

$\Delta B/rB$ is the output luminance signal/noise ratio and is equal to the input electrical signal/noise ratio. Thus the above relation shows that the threshold input signal/noise ratio varies inversely as N_2, the number of samples in the target area resulting in a 3 dB change per doubling of the number of samples. This rate of improvement continues until $N_2 = N_1$ and with further increase of N_2, the output signal/noise ratio remains constant. It may be noted that for $N_2 = N_1$ there is a perfect match and threshold should equal the theoretically predicted one (Fig. 16).

The above model is a very simple one and rather improbable as it assumes perfect integration over a sharply defined area and no integration outside it. It seems more likely that the effectiveness of integration will change gradually, possibly falling off gradually from some central point (where it is maximum) outwards. We will postulate such an averaging process applicable to circular targets by supposing that the elements at a distance s from the centre of the target area are weighted by a factor $\exp\left(-\left|\frac{s}{s_0}\right|^2\right)$

where s_0 is a constant. We will proceed to find the output signal/noise ratio by taking the weighted sums of the variances and changes of mean for this Gaussian weighting function.

Let σ^2 be the variance of the luminance of an independent element of area and ΔB the change in the mean luminance over the target and let there be D elements per unit area (measured in the same units as s^2). The maximum of the weighting function (i.e. for $s = 0$) will be assumed to be at the centre of the target. The number of elements between s and $s+ds$ is $2\pi s ds \times D$. The contribution to the change in mean from these elements is thus $\Delta B . 2\pi s ds \exp\left[-\left(\frac{s}{s_0}\right)^2\right]$. The contribution of all the elements within the target area is therefore

$$\Delta B' = \int_0^{s_T} \Delta B . 2\pi s D \, e^{-(s/s_0)^2} \, ds$$
$$= \pi D \Delta B s_0^2 [1 - e^{-(s_T/s_0)^2}] \qquad \text{.....(28)}$$

where s_T, the upper limit of integration, is the radius of the target. The weighted sum of the variances of the elements is similarly

$$\sigma_0^2 = \int_0^{\infty} \sigma^2 2\pi s \, e^{-(s/s_0)^2} ds$$
$$= \pi \sigma^2 D s_0^2 \qquad \text{.....(29)}$$

The luminance output signal/noise ratio is therefore

$$\frac{\Delta B'}{\sigma_0} = \frac{\Delta B}{\sigma} \sqrt{\pi D s_0^2} [1 - e^{-(s_T/s_0)^2}] \qquad \text{.....(30)}$$

Let ΔB be such as to give 50% detectability in the above case. Let ΔB_1 be the luminance increment required in the case of ideal matching. There are $D\pi s_T^2$ elements within the target area and if these elements were all uniformly weighted and integrated, the luminance output signal/noise ratio would have been

$$\frac{\Delta B''}{\sigma_0'} = \frac{\Delta B_1 D \pi s_T^2}{\sqrt{\sigma^2 \pi D s_T^2}} = \frac{\Delta B_1}{\sigma} \sqrt{\pi D s_T^2} \qquad \text{.....(31)}$$

The left hand side of eqns. (30) and (31) must be equal. The luminance increments in the two cases therefore bear the ratio

$$\frac{\Delta B}{\Delta B_1} = \frac{\Delta B/B}{\Delta B_1/B} = \frac{s_T/s_0}{1 - e^{-(s_T/s_0)^2}} \qquad \text{.....(32)}$$

The threshold contrast being proportional to the input signal/noise ratio, the difference in the threshold signal/noise ratio between the ideal and non-ideal integrating system is given by

$$10 \log \left\{ \frac{s_T/s_0}{1 - e^{-(s_T/s_0)^2}} \right\} \qquad \text{.....(33)}$$

With the ideal integrating system, the threshold decreases at the rate of 1·5 dB per doubling of the number of samples. The threshold for the non-ideal system may be found by adding the quantity given by eqn. (33) to the threshold for the ideal system. This is illustrated in Fig. 16 curve B. The actual thresholds have been arbitrarily chosen. The abscissa is $(s_T/s_0)^2$ which is proportional to the number of samples. The curve pertaining to the non-ideal integrator is similar to the one experimentally obtained, as may be seen by comparing with curve D, which has been transferred from Fig. 15. It has been positioned to match B, while ensuring at the same time that its displacement from the perfect matching curve is equal to that in Fig. 15. The point of interest is that with non-uniform weighting, there is no target size for which the ideal threshold is attained and that the minimum difference is 2·0 dB.

In the above calculations it has been assumed that the maximum weighting occurs at the centre of the target. This follows from an assumption that the weighting function is independent of the target size, i.e. the observer does not alter his weighting function according to the particular target he knows he is viewing. Since the slope of practical curve obtained (curve D, Fig. 16) was such that the rate of improvement with increase of area was always below 3 dB/doubling it follows that, for a fixed weighting function, the slope of the function must be negative, i.e. the curve falls away from the centre of the target.

However, if we assume that the weighting function is controlled by the observer according to the expected size of target, then it is possible to obtain curves similar to curve B with maxima at other than the centre, in particular with maximum weight given to the periphery of the target. It is quite possible that the maximum information may be obtained from the region where there is a transition from signal-plus-noise to noise alone.

This simplified analysis was merely intended to show the effect of mismatch and because of the many assumptions the similarity of curves D and B cannot be claimed to prove that the physical situation has been exactly represented. However, it lends support to the possibility of mismatch as an explanation of the type of curve obtained.

The analysis given above is equally applicable to integration in time. The shape of the curve in Fig. 15 is similar to that of Fig. 13, and by a suitable choice of the weighting function the variation of the threshold signal/noise ratio with presentation time can be accounted for. The mismatch in time (as a constant factor) can at least partly account for the difference of about 2 dB between the curves B and D in Fig. 16.

The data in Fig. 15 show that the optimum integra-

tion time is about 320 ms. An increase of presentation time by a factor of 2 to 640 ms results in a very small reduction of the threshold. But the results of the experiment with long presentation time (8 seconds) are also sub-optimum to the same extent. In other words, the increase of exposure by a factor of 25 over the optimum period has not impaired the efficiency of temporal integration whereas an increase by a factor of 2 appears to impair it appreciably. This apparent anomaly may be due to different mechanisms being brought into operation by the longer period available in the former case. If this is true the visual detection system can operate with greater efficiency if longer periods of observation are available without a proportional increase in the number of samples. The p.p.i. and the chemical recorder provide these situations in practice.

8. Conclusions

An attempt has been made to compare the performance of a human observer and a theoretical model in the task of the detection of a target area in a uniform background on a background of known graininess. In the experimental work a closed circuit television system was used to simulate an intensity—modulated display and the close comparison of the results for the uniform background with those of other workers in this field give validity to this method. A forced choice method was used for the tests since not only does this give more consistent results but fortunately there is an exact mathematical model for this condition.

The results suggest that in this task the visual detection system is a sub-optimum one and its efficiency is dependent, among other things on the area of the signal. The efficiency is very poor for small areas but its efficiency increases up to an optimum after which it decreases again. At the optimum the human observer is about 4·5 dB below the ideal but it appears that the optimum depends on the experimental conditions. A different optimum being found in the 8-position experiments compared with the 2-position experiments. This improvement of efficiency with area for small areas may go some way towards explaining results obtained by various workers in which the *rate of improvement* has been greater than would be expected by simple theoretical treatments.

The inefficiencies of the visual observer can be simulated by assuming that he weights the information he receives by a Gaussian weighting function with its maximum at the centre of the target. This produces an inefficiency of the sort recorded in practice but it would be pretentious to suggest from this similarity that the human observer makes his decisions on this basis.

It has been shown that an optimum non-linearity exists for the c.r.t. but that this depends on the average background illumination. Effectively this is saying that if the noise on the display is less than the internal noise of the visual system it pays to introduce non-linearity since this increases the mean of the target compared with the mean of the background at the expense of increasing the display noise.

A further paper by one of the authors[13] will deal more thoroughly with the aspects of the results which can be related to the measurement of the internal noise of the human visual system.

9. References

1. J. L. Lawson and G. E. Uhlenbeck, "Threshold Signals", (McGraw Hill, New York, 1950).
2. R. Payne-Scott, "The visibility of small echoes on radar p.p.i. displays", *Proc. Inst. Radio Engrs*, **36**, p. 180, 1948.
3. P. McGregor, "A note on trace-to-trace correlation in visual displays: elementary pattern recognition", *J. Brit.I.R.E.*, **15**, p. 329, 1955.
4. J. W. R. Griffiths and I. G. Morgan, "The chemical recorder and its use in detecting pulse signals in noise", *Trans. Soc. Instrum. Tech.*, **8**, p. 62, 1956.
5. D. G. Tucker, "Detection of pulse signals in noise: trace-to-trace correlation in visual displays", *J. Brit.I.R.E.*, **17**, p. 319, June 1957.
6. Y. W. Lee, "Statistical Theory of Communication", (John Wiley, New York, 1960).
7. D. Middleton, "An Introduction to Statistical Communication Theory", (McGraw Hill, New York, 1960).
8. J. W. R. Griffiths, "Detection of pulse signals in noise: the effect on visual detection of the area of the signal paint", *J. Brit.I.R.E.*, **17**, p. 330, June 1957.
9. H. R. Blackwell, "Contrast thresholds of the human eye", *J. Opt. Soc. Amer.*, **36**, p. 624, 1946.
10. J. S. Bendat, "Principles and Applications of Random Noise Theory", (John Wiley, New York, 1958).
11. E. Jahnke and F. Emde, "Tables of Functions with Formulae and Curves", (Dover, New York, 1962).
12. S. O. Rice, "Mathematical Analysis of Random Noise" in "Selected Papers on Noise and Stochastic Processes", p. 133, (Dover, New York, 1954).
13. N. S. Nagaraja, "The Effect of Luminance Noise on Contrast Thresholds", (To be published).

10. Appendix 1

Calculations Relating to Signal/Noise Ratio

The probability distribution of the envelope amplitude R of narrow-band white noise with a c.w. signal of amplitude at the input is given[12] by the probability density function

$$\frac{R}{\psi_0}\exp\left[-\frac{R^2+P^2}{2\psi_0}\right]I_0\left(\frac{RP}{\psi_0}\right) \qquad \text{.....(34)}$$

where ψ_0 is the mean-square noise at the input and I_0 is the modified Bessel function of zero order.

The average of the vth power of R is given by the integral

$$\overline{R^v} = \int_0^\infty \frac{R^{v+1}}{\psi_0} \exp\left[-\frac{R^2+P^2}{2\psi_0}\right] I_0\left(\frac{RP}{\psi_0}\right) dR \quad \text{.....(35)}$$

$$= (2\psi_0)^{v/2}\,\Gamma\left(\frac{v}{2}+1\right) {}_1F_1\left(-\frac{v}{2};1;-\frac{P^2}{2\psi_0}\right) \quad \text{.....(36)}$$

This expression may be used to obtain the averages associated with the output of a vth law demodulator.

The output signal/noise ratio R_B is obtained by putting the appropriate expressions for the mean values.

$$R_B = \frac{\Gamma\left(\frac{v}{2}+1\right)}{\left\{\Gamma(v+1)-\Gamma\left(\frac{v}{2}+1\right)^2\right\}^{\frac{1}{2}}} \times \times \left[{}_1F_1\left(-\frac{v}{2};1;-\frac{P^2}{2\psi_0}\right)-1\right] \quad \text{.....(37)}$$

The ratio of r.m.s. fluctuation to the mean value of noise output r is similarly obtained

$$r = \frac{\left\{\Gamma(v+1)-\Gamma\left(\frac{v}{2}+1\right)^2\right\}^{\frac{1}{2}}}{\Gamma\left(\frac{v}{2}+1\right)} \quad \text{.....(38)}$$

The change of mean in the presence of the signal when the input signal/noise ratio is unity (0 dB), as a ratio of the mean value of noise is obtained by (36)

$$\frac{\text{change of mean at 0 dB}}{\text{mean of noise alone}} = \left[{}_1F_1\left(-\frac{v}{2};1;-1\right)-1\right]$$

$$C_0 = \left[e^{-1}\,{}_1F_1\left(1+\frac{v}{2};1;1\right)-1\right] \quad \text{.....(39)}$$

This expression can be evaluated from the known values of v, but unfortunately the hypergeometric function is tabulated only for integral values of v and so the above ratio has been obtained by interpolation. In Table 4, the measured and calculated values of C_0 are given. The differences are not appreciable and may be attributed to the fact that the luminance law cannot be exactly represented by a power-law relation.

Table 4

Relation between the measured and calculated values for C_0.

Luminance ft-lambert	v (measured)	C_0 (calculated)	C_0 (experimental)	v (calculated from C_0)
0·1	1·33	0·63	0·66	1·39
1·0	1·15	0·525	0·58	1·25
0·1	2·19	1·06	1·23	2·39
1·0	1·8	0·88	1·01	2·02
0·1	3·0	1·67	1·70	3·01
1·0	3·27	2·04	2·12	3·3

Manuscript first received by the Institution on 22nd June 1962 and in final form on 8th January 1963 (Paper No. 796/SS18)

U.D.C. 536.53: 616

The Electronic Heat-Camera in Medical Research

K. LLOYD WILLIAMS,
M.A., M.Chir., F.R.C.S.,†

C. MAXWELL CADE,
(*Member*)‡

AND

D. W. GOODWIN, B.Sc., Ph.D.§

Presented at a meeting of the Medical and Biological Electronics Group in London on 16th May 1962.

Summary: The first part of the paper describes early work on determining skin temperatures by detecting and measuring infra-red radiation. The results of recent measurements have been related to the presence of growths and other conditions beneath and upon the skin. Possibilities exist for reflection spectroscopy by relating the wavelength of the reflected radiation to the constitution of the surface. Pictures built up by scanning techniques are shown and the second part of the paper deals with the design of such equipment. The scanning detector is known as the pyroscan. Finally in the third part of the paper a review is given of the relevant parameters of fast photoconductive cells suitable for use with the pyroscan. The possible application of other techniques for medical diagnosis is also described.

1. Infra-red Techniques in Medical Research

1.1. *Measurement of Skin Temperature*

Some two years ago, attempts were made at the Middlesex Hospital to determine the site of perforating veins in varicose ulceration. It is known in this condition that the valves in the veins which normally permit blood to pass in one direction only, from superficial to deep veins, are incompetent; and with each contraction of the muscles of the calf, blood is squirted out from these deep veins into the superficial system of veins. It was believed that the blood in the deep veins should be "hotter" than the skin, and that if the calf muscles were made to contract, the site of the incompetent perforaters should be demonstrated as a "hot spot" in the overlying skin. This turned attention to methods of measuring skin temperature. The measurement of surface temperature from the infra-red radiation has great advantages over direct thermometry because:

(i) the rays can be measured at a distance and contact with the skin is unnecessary;

(ii) it is very rapid;

(iii) measurement at a distance yields an average reading for a small area, rather than for a point, which compensates for local variations in capillary tone;

(iv) by measuring temperature at a distance it is possible to use a scanner which will demonstrate the spatial distribution of surface temperature changes over a wide area. This, it was thought, might show several perforating veins at the same time.

It is well established that the skin, whether black or white, is within 3% of being a "black-body" radiator.[1–5] J. D. Hardy was advocating the use of radiation thermometry in 1935, and he showed that the skin's emission of 2 to 20 microns, with a peak intensity at around 9 microns, corresponded fairly well with the calculated emission of a "black-body" at 35° C.

A Schwarz thermopile fitted with a calcium fluoride window and sensitive up to 11 microns was therefore used, and the relationship of emission to temperature was determined by comparison with the emission from the blackened surface of a Leslie cube at known temperature. It was found that the distance of the thermopile from the skin was not critical, and any distance between ½ and 1½ cm produced the same deflection.

With this apparatus it was possible to prove the accuracy of the first premise and we could, in fact, demonstrate an incompetent perforating vein. Unfortunately the technique was too tedious to be of clinical value, because only a very small area could be measured at a time, and it took quarter of an hour for the skin temperature conditions to stabilize after each measurement.

1.2. *Relation of Skin Temperatures to Clinical Details*

Attention was turned next to conditions known to have elevated temperatures such as abscesses and local inflammation, and it was found that they could be of higher temperature than the normal body internal temperature. This in itself was interesting because the usual explanation of the increase in temperature over an acute abscess is increased blood

† Senior Surgical Registrar, The Middlesex Hospital, London, W.1.

‡ S. Smith and Sons (England) Limited, Carlisle Road, London, N.W.9.

§ Royal Radar Establishment, Malvern, Worcestershire.

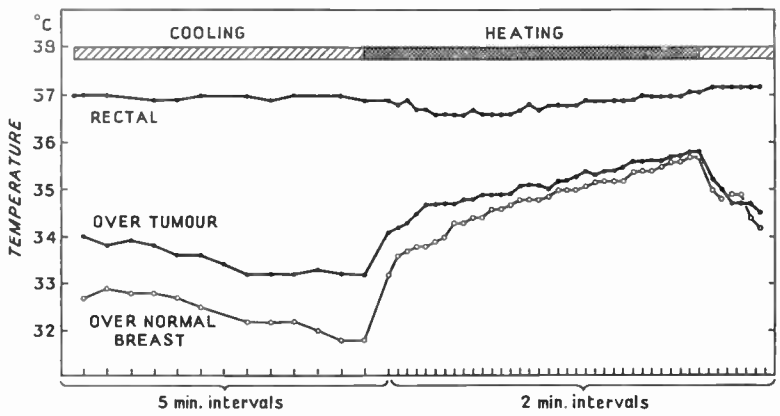


Fig. 1. Effect on skin temperature of warming and cooling a patient with spheroidal-cell carcinoma of breast (room temperature 20·5° C).

supply. Next, the temperature of the skin overlying superficial malignant conditions was measured, and it was found that these appeared to be "hot" in relation to surrounding structures and the contra-lateral normal area. It was therefore decided to investigate a series of lumps in the breast and determine the temperature pattern of the skin in relation to the underlying pathology. It is known that identical symmetrical areas of the body surface are at the same temperature. In fact, the temperature difference between two symmetrical parts under standard environmental conditions does not vary more than 1 deg C unless there is derangement of the vascular supply or some pathological process to explain it.

Early on it was discovered that the surface temperature of the body drops rapidly for the first 10 to 15 minutes after exposure to a lower ambient temperature and thereafter drops slowly for the next hour or so until equilibrium is achieved. Twenty minutes exposure was chosen as an arbitrary time to overcome this profound initial fall, but considerable difficulty was experienced in controlling the ambient temperature. Another discovery was that cooling accentuates the temperature difference between a cancer and normal tissue. This is represented diagrammatically in Fig. 1 for the case of carcinoma of the breast. The temperatures were measured in this instance by thermocouples and it can be seen that the temperature difference is increased by cooling and diminished by heating the patient.

An investigation was carried out on 100 patients admitted to the Middlesex Hospital with a lump in one breast.[6] In all except two cases of advanced malignant disease the diagnosis was confirmed by histology. The method was to expose the patient for 20 minutes on a couch with the arms raised behind the head to prevent cross-radiation from the arms to the chest wall. The infra-red emission was then measured with a Schwarz thermopile held in the hand and the resulting galvanometer deflection was recorded on a diagrammatic chart. Measurement was first of one side and then the directly opposite area on the other side. Figure 2 shows the chart made out in a patient with carcinoma of the breast. The site of the cancer is indicated by cross-hatching.

The results of these investigations are shown in Table 1. All cases showing a rise of more than 1 deg C over the contra-lateral normal area have been designated as "hot", while all cases with less than 1 deg C rise are shown as "cold". Summarizing these findings, the lumps in the breast seem to separate themselves into two groups. The "hot" group consisted of abscesses and cancers and the "cold" group of degenerative lesions, such as cysts and duct stasis. There are four exceptions to this: three cases of carcinoma out of 57 failed to show a rise in temperature, and one cyst out of eighteen did show a rise in temperature. The latter cyst was found histologically to be inflamed, and should therefore be included with the abscesses. However, there was no explanation for the three "cold" carcinoma. They appeared similar in all ways, in histology, size and site in the breast as the other carcinomas. The fibroadenomata or benign

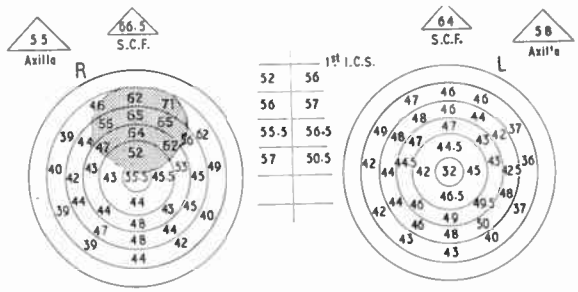


Fig. 2. Diagrammatic chart on which temperature readings were recorded, completed for a case of breast carcinoma.

The figures represent centimetres deflection on a mirror-galvanometer scale. 36° C is represented by reading of 61·6 cm. Rectal temperature was 37·1° C. Calibration against a standard heat source showed that the average temperature over the tumour was 2·5 deg C higher than the corresponding area in the normal breast.

S.C.F. = superclavicular fossa; I.C.S. = intercostal space.

Table 1

Results of infra-red measurement of skin temperature in 100 cases of lump in one breast

	No. of Cases	Rise	No Rise
Abscesses	4	4	0
Carcinomata	57	54	3*
Fibroadenomata	10	6	4
Cysts	18	1*	17
Adenosis	6	0	6
Fat Necrosis	2	0	2
Duct Stasis	3	0	3

"Rise" indicates more than 1 deg C difference between the breasts. The exceptional cases of carcinoma and cyst are indicated thus *.

growths of the breast were sometimes "hot" and sometimes "cold" and the temperature rise appeared to be related to the degree of cellularity; the more cellular tended to be "hotter". These results substantiate the work of Lawson.[7]

Figure 3 shows the degree of temperature rise in the series. The majority of cancers have a rise of 1–2 deg C though some are much "hotter". One was 7½ deg

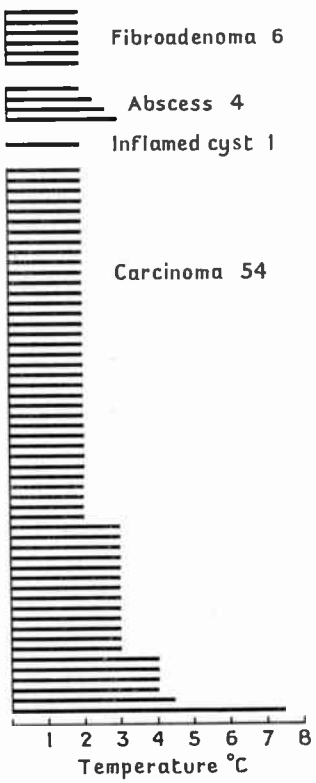


Fig. 3. Temperature range over 65 "hot" lumps.

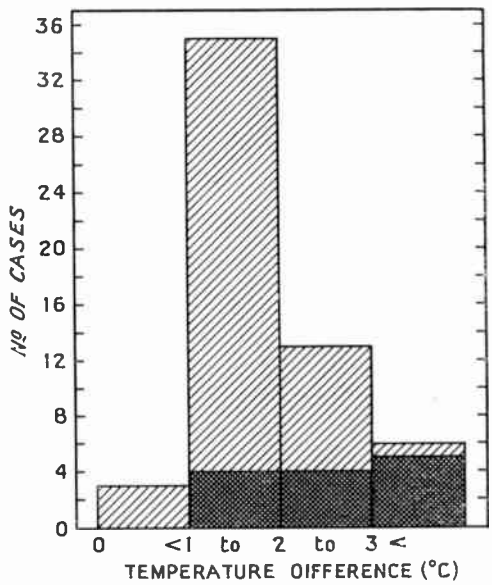


Fig. 4. Skin temperature difference in 57 cases of carcinoma of the breast. Cases in which metastases subsequently developed are stippled.

"hotter" than the other side. This was a large ulcerating tumour. Several of the "hotter" cases were found to have a temperature greater than the patient's rectal temperature.

The question then arose whether the degree of temperature difference in any way reflected the malignancy of the tumour. Figure 4 shows the temperature overlying the cancers in the series plotted against the number of cases. The cross-hatched areas in the graph show the number of cases in the different temperature groups which demonstrated spread of their disease to other organs within 18 months. Thus, of those cases with 3 deg C or more temperature rise, five out of six had spread of their tumours within 18 months. It is obviously too early to do more than speculate on the value of these findings, but it may be that the degree of temperature rise is a reflection of tumour activity. If the temperature rise is a parameter of cellular activity, then it should be of value in diagnosis and prognosis and may facilitate the selection of the correct hormone balance or cytotoxic agent which will affect a particular tumour in a particular patient.[8]

Because infra-red waves can be measured at a distance, the production of a pictorial representation of temperature differences by a scanning mechanism appeared to be possible. This has been described by Lawson[9] and it so happened that at this time (March, 1961) the Kelvin-Hughes Division of S. Smith & Sons (England) Ltd., had produced a prototype scanner, believed to be the only machine of its kind in Europe. It must be emphasized that this apparatus, called the

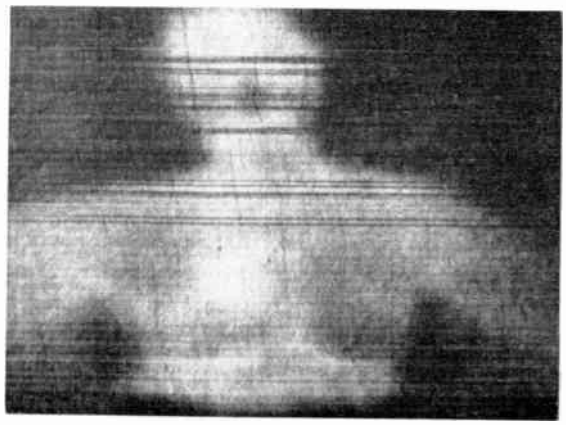


Fig. 5. Heat picture produced by infra-red scanning of a patient with carcinoma of the breast. The light areas are "hot".

Fig. 6. 2 to 5 micron heat scan of tuberculous abscess of the right breast which is 2½ deg C hotter than the opposite side.

Pyroscan, was designed for long-range vision through fog, and much adaptation was required to view a patient at 5 ft.[10] The Pyroscan's photo-sensitive cell was an indium antimonide type ORP 13, kindly loaned by Mullard. The machine scanned the head and thorax in 4 minutes, writing its information on electro-chemical paper with 100 lines to the inch. Its visual discrimination of temperature appeared to be 1 deg C at body temperature. Figure 5 shows a heat picture in a patient with a cancer of the right breast. White is "hot" and black is "cold". The large "hot" area in the right breast was 5 deg C hotter than the left side. It is also noticeable that the area above the right clavicle is "hotter" than the comparable area on the left side. Examination of enlarged glands found here demonstrated secondary deposits of cancer.

Figure 6 shows a tuberculous abscess in the right breast. These are often referred to medically as "cold abscesses". This one was 2½ deg C "hotter" than the opposite side, and note also that the right axilla, which contained palpable inflammatory glands, is "hotter" in the picture than the left.

Although these pictures are crude, which is not very surprising considering that the scanning element was a searchlight mirror, they did show that heat scanning would demonstrate temperature contours.

Attention has so far been confined to the applications of infra-red scanning in the diagnosis of breast disease. However, there are many other applications and a few of these will be mentioned. As the temperature of the limbs at rest is largely maintained by heat transferred to them from other areas by their blood supply, surface temperature can be used as an assessment of the efficiency of that blood supply. It can be used to demonstrate the position of a block in a peripheral vessel by a sharp fall in the limb temperature at the site of the block, and may be of value in determining the optimum site of amputation in the presence of an impaired blood supply.

Heat pictures would be valuable in demonstrating "hot" spots in connection with arteriovenous shunts and fistulae. In plastic surgery temperature measurements may be used to assess the blood supply of pedicle grafts, where skin from one part of the body is raised in the form of a tube and transplanted in stages to another part. The vitality of the tube is dependent on the blood supply entering its ends and this can be assessed by its surface temperature. In operations such as radical mastectomy where large areas of skin are raised it may give some assessment of the subsequent viability of the skin. In vascular or highly metabolic areas, as in the thyroid gland in thyrotoxicosis, heat scanning will demonstrate the lesion. Figure 7 shows a patient with thyrotoxicosis with an enlarged bi-lobed thyroid. The thermal picture shows two rounded "hot" areas in the neck, corresponding to the large thyroid lobes.

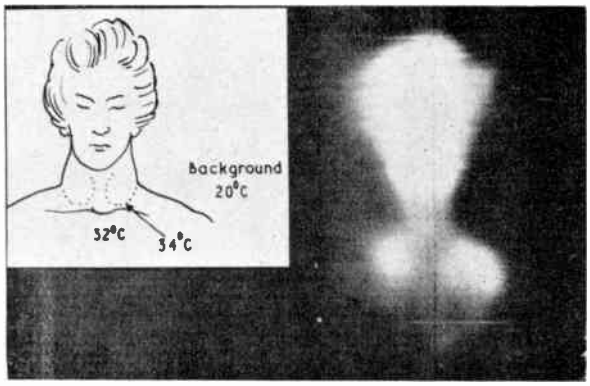


Fig. 7. Heat scan of neck of patient with thyrotoxicosis.

In rheumatoid arthritis, temperature may be used as a measurement of activity of the disease, or alterations in joint temperature may be of value in assessing the effect of therapy.

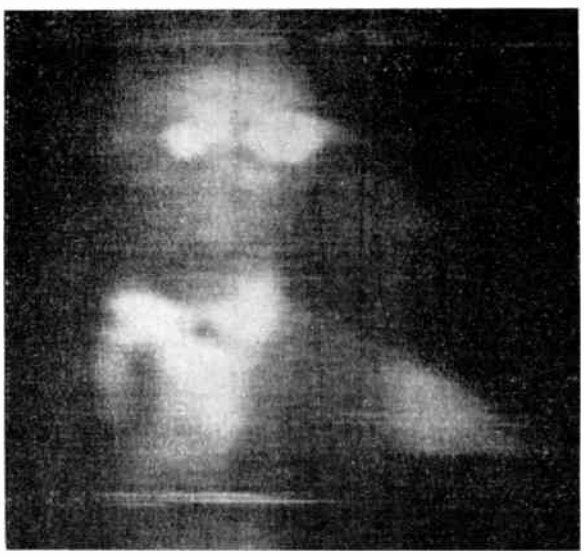
A fast scanner may be able to demonstrate incompetent perforating veins, and Lawson has shown[11] that heat pictures will demonstrate the depth of a burn or frostbite within a short time of the injury because the dead and avascular tissue diminishes the radiation of infra-red. This has now been confirmed with the Pyroscan and should be of great value as in many cases the viability of the skin can only be assessed by clinical methods after 10 to 12 days. Further applications which have yielded interesting and significant results include studies of bone tumours and fractures.

1.3. *Problems in Heat Scanning Diagnoses*

It would be misleading to imply that no problems have been encountered with the technique and interpretation of heat scanning. Firstly, some lumps in the breast, demonstrably "hot" when measured with a thermistor bolometer, thermocouple or thermopile, failed to show up on scanning. Figure 8 shows a 2–5 micron infra-red picture of a lady with a lump in the left breast which was 2 deg C "hotter" than the opposite area; this does not show in the picture. Next, notice the black "cottage loaf" marking on the lower chest. This was a silver sixpence, which appears bi-lobed because of respiratory movement. The sixpence appears very much blacker ("colder") than the surrounding area although its temperature was less than 1 deg C different from this area. This is due to the difference in emissivity between the shiny sixpence and the skin, whose emissivity is closer to that of black-body. The sixpence appears almost the same temperature as the black area in the middle of the face which is the lady's cold nose, which by thermocouple measurement was 10 deg C colder than the sixpence.

Fig. 8. 2 to 5 micron infra-red scan of patient with a lump in the left breast 2 deg C hotter than the right.

(*a*) 5·0–5·5 microns.



(*b*) 3·5–4·0 microns.

Fig. 9. Reflection pictures taken with the Pyroscan in narrow wave-band with half micron band-pass filters.

The second problem also concerns the emissivity and reflectivity of surfaces. When we see things in light, the colours and characteristics of the objects in the particular wavelengths in which we observe them

are determined by the absorption and reflection of the objects at those wavelengths. In fact, it might be said that we see by virtue of reflection spectroscopy. To illustrate this, two photographs (Fig. 9(*a*) and (*b*)) are shown which were taken with the Pyroscan in reflected infra-red light with different half-micron bandpass filters. (These two are selected from a larger series.) The first, Fig. 9(*a*), is a picture of one of the authors (C. M. C.) wearing glasses and a bow tie; the waveband was 5·0 to 5·5 microns; notice that the glasses are black and the bow tie is black also. The next picture, Fig. 9(*b*), was taken at 3·5 to 4·0 microns. The bow tie is now white, and the glasses have changed to white because of a change in the angle of incidence of the infra-red source. The change of the bow tie from black to white is, of course, related to the molecular constitution of the surface observed, and this illustrates the possibility of reflection spectroscopy as a method of qualitative analysis.

The next example concerns the possibility of an optimum band for the measurement of infra-red emission. If one produces a picture of wide bandwidth in infra-red, then this picture is obviously composed of the addition of numerous pictures in narrow bandwidths: these, in turn, are related to the emission from the subject in each particular narrow band. Thus, as in the last two pictures where in one narrow band the bow tie was black and in the other white, these two pictures will be added together to produce a wideband picture, and this may result in the bow tie appearing grey. What is important is the possible loss of contrast which may result.

Similarly, the lady with the "hot" lump in the breast which did not show on scanning in the wide band might have shown in a narrow band. An opportunity of testing this occurred when another of the authors (D. W. G.) developed a vaccination reaction on his left arm during the smallpox epidemic: this was a very hot, tender, localized inflammation with large tender axillary glands. Pictures taken in the bands 2·0–6·0 microns and 5·0–5·5 microns, showed nothing of any real interest. However, a picture taken at 4·0–4·5 microns, showed very clearly the "hot" area of the vaccination in the left arm and also what appeared to be a "hot" area related to the axillary glands. It may be that selection of a particular band may help in differentiating various conditions because of their spectral characteristics in the infra-red.

As has been stated, it is established that normal human skin is very close to black-body in its infra-red characteristics when measured over wide bandwidths. The authors' investigations confirm that beyond about 6 microns this is true; from 3 to 6 microns the departure of skin from black-body appears to be related to the water content of the epidermis.† Below 3 microns there are probably transmission bands because, for example, infra-red photography at 0·9 to 1·2 microns will show up veins which are deep below the skin.

If therefore the energy emitted by the skin by virtue of its temperature is integrated over a wide band as with a Schwarz thermopile, any small departures from black-body conditions are diminished, and hence the energy emitted more accurately reflects the skin temperature, whereas in narrow bands these departures from black-body emission are relatively greater and therefore the energy emitted less accurately reflects the skin temperature.

The present detector, when cooled with liquid nitrogen, will only respond out to 5·5 microns, some way below the peak emission of the skin at 9 microns; and this may possibly explain some of the difficulties, although in theory it should enable smaller temperature differences to be resolved than would otherwise be possible. We have also to consider that the physical properties of the skin overlying pathological conditions may be changed, and hence its emission, particularly for the shorter wavelengths corresponding to molecular resonances, may be altered.

2. Principles of Electronic Heat-Cameras

Electronic heat-cameras, or thermal image-converters as they are also called, require no external source to illuminate the scene (although, as has been noted, there may be times when an external source extends the value of the apparatus). They respond to the radiation which is emitted by every object at a temperature above absolute zero, and which is proportional to the temperature and emissivity of the object. However, the emissivity may be a rapidly-varying function of both temperature and wavelength, as appears to be the case with human skin under certain conditions, and since the energy emitted in a narrow spectral band may be very small at body temperature, we may be forced to a choice between local "thermal illumination", which completely alters the picture contrast ratios, and a long integration time, which improves the sensitivity of the apparatus, but makes things difficult for the patient. The use of a separate source really opens up a new field: by viewing a surface in successive narrow spectral bands it is sometimes possible to learn far more than could be derived from a study of the object's own broad-band thermal radiation.

Within the range of its spectral response, an electronic heat-camera transcribes the information received by its detector into a two-dimensional record, presenting an intensity pattern, and the machine may be calibrated, if so desired, to give an accurate representation of the isothermal contours of the object scanned. The theoretical limit to the performance of

† To be published elsewhere.

such machines is set by the random fluctuations in the radiation received, or by electrical noise in the receiver circuits. In practice the performance is usually circumscribed by optical or electrical defects in the system, or by the ability of the human eye to discriminate intensity differences in the record.

Two principal groups of thermal image-forming devices may be distinguished, based upon the type of radiation detector used:

(1) Temperature-sensitive, or thermal image converters, which employ heat-sensitive detectors (thermocouples, thermopiles, or thermistors) and which have a very wide spectral response;

(2) Photo-sensitive devices, which may be subdivided into: (2a) photo-emissive detectors with a spectral response limited to the visible and the near-infra-red wavelengths; (2b) photo-conductive detectors, with response extending usually to about 6 to 10 microns, and in a few special cases to 20 or 30 microns. These latter usually require refrigeration with liquid helium, and are rather an expensive complication for non-military purposes.

Of the above, the response time of group (1) is too slow, and the wavelength limit of (2a) is too short, for them to be of much use in medical studies, which must therefore use detectors of type (2b) to meet their special requirements. The principal limitation to the extensive work which has been carried out in America and Canada has been due to the use of thermistor devices, which although they are very sensitive, require a time of ten to fifteen minutes to produce a thermal picture of a patient's head and shoulders.

2.1. *General Description of the Heat-Camera*

An electronic heat-camera using a detector of type (2b) can be divided into the following component parts for ease of reference:

(1) *The Detection Unit*, comprising (*a*) the sensitive cell itself, together with its associated refrigeration equipment (where applicable) and any spectral pass filters; (*b*) the pre-amplifier and main amplifier (a pre-amplifier is usually used only when it is necessary in order to match the cell impedance to the input of the main amplifier, or when the cell must be situated at some distance from the main amplifier); (*c*) the radiation chopper and phase-reference circuit. The radiation chopper serves to give a.c. signals of pre-determined frequency and bandwidth, even when the received radiation is of constant intensity; the phase-reference circuit is used to eliminate the thermal radiation of the chopper itself, which otherwise will set the background noise limit to the system.

(2) *The Optical System*, comprising a main mirror (or lens) and subsidiary mirrors (or lenses) to collect radiation from the patient under examination and focus it upon the sensitive cell. Mirrors are preferred, as the cost of achromatic infra-red lenses is extremely high.

(3) *The Scanning System*, which is basically a simple arrangement for mechanically causing different parts of the image to fall upon the cell in sequence. The design is greatly complicated by requirements for variable scan-rate, variable patient-scanner distance, variable depth of focus, and anti-vibration precautions, since the cells are sometimes extremely microphonic.

(4) *The Display Unit*, which may contain a mosaic of neon tubes, a cathode-ray tube or storage tube, a scanning neon lamp and photographic plate, or some type of electrochemical recorder.

2.2. *The Pyroscan*

The Pyroscan uses a detector of type (2b). It was originally developed for research into vision through fog, aerial reconnaissance and industrial pyrometry at low temperatures, and it sacrificed both speed and portability to the requirements of sensitivity and versatility. The optical system employs a 16-in. diameter primary mirror and a 3-in. secondary mirror in a modified Cassegrain configuration. Scanning is carried out mechanically by moving the optical system in two dimensions, using a fast-return mechanism which operates automatically at the end of each frame. The image is printed on an electrochemical paper, and may be either positive or negative as desired. The sensitive cell was at first a Mullard ORP 10 indium antimonide cell operating at room temperature, but this gave rise to considerable drifting in the background level, and it was replaced by an ORP 13, using liquid nitrogen as a refrigerant, with an immense improvement in the signal to noise ratio. The driving stage of the recorder is fed by a d.c. amplifier, which sets the contrast ratio, and this in turn is fed through a rectifier stage from an amplifier with a very low-noise input stage and a total voltage gain of 120 dB.

The Pyroscan has to handle very feeble signals. The resolution element at the patient is about one square centimetre, and 1 cm^2 of black-body surface radiates roughly 50 milliwatts, integrated over all wavelengths, at the normal body temperature range of 32–37° C. The use of $\frac{1}{2}$ micron filters reduces this immediately to 1 milliwatt, which at the normal patient-scanner distance of six feet, falls to $2{\cdot}5 \times 10^{-8}$ watt.cm^{-2}.

There are various reasons why it is desirable to complete a thermal picture of a patient in a shorter time than the two to five minutes required by the present apparatus: movement by the patient can blur the picture, a difficulty which would be reduced by

faster scanning; also, faster scanning would allow some transitory phenomena to be observed. These requirements have to be weighed against the need for better thermal resolution, preferably of the order of 0·1 deg C. Investigations using spectrally-filtered radiation require detectors of very high performance, but unfortunately the requirements of high thermal resolution and high frame speed are incompatible.

A completely re-engineered version of the Pyroscan, the Mark II, is now nearing completion. Tests on the component assemblies show that it will give a picture of almost photographic quality in 30 seconds, with a thermal resolution of about 0·5 deg C. A thermal resolution of better than 0·1 deg C can be obtained at the cost of increasing the scanning time to 3 minutes. The opportunity has been taken to incorporate a number of mechanical refinements which will greatly facilitate the use of the apparatus for clinical purposes. For example, pictures can be taken in very rapid succession in adjacent wavebands, thus facilitating the selection of the optimum waveband for a particular case, without requiring the patient to spend a long time in a chilled room. Facilities are also being designed to give a direct black-body temperature comparison, and read-out in the form of plotted isothermal contours or as printed digits. These additional facilities are expected to take some time to develop.

3. Infra-red Detectors for use with Pyroscan and for Human Thermometry

3.1. *Photo-conductive Detectors*

3.1.1. The limit of detection

The Pyroscan unit utilizes a photo-conductive detector of indium antimonide cooled to 77° K. Although there are many materials exhibiting photo-conductivity yet very few have been developed fully into useful detector systems and consequently the possibility of finding a detector having the correct wavelength response, detectivity and time constant is very remote.

The performance of a detector is described in terms of a normalized detectivity (D^*) which is the signal/noise ratio for an input power of one watt normalized for 1 cm^2 of area and 1 c/s bandwidth. The limit to the detectivity is set usually by the noise voltage. In the case of indium antimonide cells there are several sources of noise.[12]

For low frequencies, below 100 c/s, the noise voltage increases with decreasing frequency, the so-called flicker noise. The mean square noise current $\bar{i}^2$ is given by

$$\bar{i}^2 = \frac{AI^2\Delta f}{f^\alpha} \qquad \ldots.(1)$$

when A and α are constant, I the bias current and f the frequency. It has been shown that A and α are functions of the surface preparation of the detector and that α can be as large as 3. There is obviously a need to operate an InSb photocell at frequencies greater than 100 c/s in order to overcome flicker noise. For frequencies higher than this the noise voltage becomes independent of frequency (generation-recombination noise), and for frequencies comparable with the reciprocal of the carrier lifetime decreases rapidly. A cell having unit quantum efficiency, high responsitivity and low noise is capable of detecting environmental radiation, usually black-body in nature because of the cavity-like shape of the room and high emissivity of walls. This flux is of the order of 10^{16} photon per second, hence the statistical fluctuation ($\sqrt{n}$) will be 10^8 photon/second. Any detector capable of detecting this flux within the relevant waveband will see it as noise, which will constitute the limit of detection, defined[13] as D^*_{Blip}. For a detector capable of detecting all wavelength, such as a thermopile or Golay detector D^*_{Blip} has a value of 2×10^{10} cm/watt for an ambient temperature of 300° K. The use of cooled filters can increase this value. However, if the spectral variation of the source and background are very similar, i.e. similar temperature and emissivity, and if a cell whose D^* approaches D^*_{Blip} is used, then little improvement can be gained by the use of cooled filters. The only slight improvement that can be made is by operating at as short a wavelength as is possible and by integrating signal and noise over long periods.

3.1.2. Indium antimonide cells

Until a few years ago the only available photo-conductive cells were of the layer type of PbS, PbSe and PbTe having sensitivities out to 3, 4 and 6 microns respectively, the first two at room temperature and the latter at 77° K. Although improvements have been made over the years in both PbS and PbSe cells† single crystal photocells are preferred because of their reproducibility and spatial uniformity of response. In Table 2 the properties of InSb photo-conductive detectors at various temperatures are listed.[14]

Table 2

Temperature (°K)	Wavelength (microns)	Time-constant (microseconds)	D^* (cm/watt)
290	7·5	0·03	5×10^8
195	6·7	0·2	5×10^9
77	5·9	3·0	7×10^{10}

The value of D^*_{Blip} at 77° K is $1{\cdot}2 \times 10^{11}$ cm/watt. Thus the best cooled cells approach this limit. The values of D^* in Table 2 are for a chopping frequency

† By Infra-red Industries Ltd. and Eastman Kodak Inc. respectively.

of 800 c/s, a frequency at which $1/f$ noise is not dominant. Thus indium antimonide when operated at 77° K offers high speed of response with high detectivity.

In the past several attempts have been made to formulate cell sensitivity, the Havens' formula for thermal detectors being a typical example. However, for photon detectors in which the responsivity and noise voltage are limited by recombination processes within the material there appears to be no universal rule. However, for some photo-conducting materials at a particular temperature some conclusions can be reached and indium antimonide at 77° K is a typical example. The majority carrier lifetime (τ) is inversely proportional to the carrier density and so proportional to the resistivity (ρ). Because the cell is limited normally by generation-recombination noise, the detectivity can be written in the form

$$D^* = 4 \times 10^{12} \rho \sqrt{\tau} \text{ cm/watt} \quad \ldots.(2)$$

Hence the minimum energy which can be detected is

$$W_{min} = \frac{1}{D^*} \frac{\sqrt{A}}{\sqrt{t}} \quad \ldots.(3)$$

where A is the area of the detector and t the integration time. Obviously the minimum value of W_{min} occurs when $D^* = D^*_{Blip}$. For short wavelengths the total radiation coming from a black-body is given by Wien's law and the rate of change of energy with temperature is given by

$$\frac{du}{dT} \propto \frac{1}{T} \exp\left(\frac{1}{h\nu/kT}\right) \quad \ldots.(4)$$

which approximates to a T^{-25} law.

On the other hand a broad-band detector such as a Golay detector will have $du/dT \propto T^3$. Hence the use of an InSb detector of high responsitivity will give much better temperature discrimination than a black-body detector. Recent measurements indicate that within the waveband of an InSb detector the emissivity of the skin is far from unity and it is to be expected that the utilization of the short wavelength emission together with good filtering will lead to a better understanding of the properties of the human epidermis.

3.2. *Impurity Photo-conductors*

For some medical diagnostic applications there is a need for fast detectors capable of detecting as large an amount of the total radiation as possible, where it is necessary to sacrifice selectivity of wavelength for speed of response. For such applications it is necessary to consider impurity photo-conductors in which photo-excitation occurs between band impurity states and the conduction band. Such detectors need cooling to temperatures of 20° K and below in order to reduce thermal excitation into the conduction band. A summary is given in Table 3 of various dopants in both germanium and indium antimonide and the wavelength response likely to be achieved.

Table 3

Semi-conductor	Dopant	Operating Temperature °K	Wavelength (microns)	D^*
Ge	Cu[6]	5	25	10^{10}
Ge	Zn	4	40	10^{10}
Ge	Hg	35	10	2×10^{10}
InSb	Ag[7]	5	32	—
InSb	Au[7]	5	20	—
InSb	Cu[7]	5	23	—
InSb(n)[8]	—	1·35	8600	—

3.3 *Future Detection Systems*

Recently an attempt has been made to produce a narrow-band detector tunable with the aid of a magnetic field. When the frequency of an electromagnetic field corresponds to the cyclotron resonance frequency of electrons within a solid, power will be absorbed thus raising the electrons out of thermal equilibrium and modulating their mobility. A

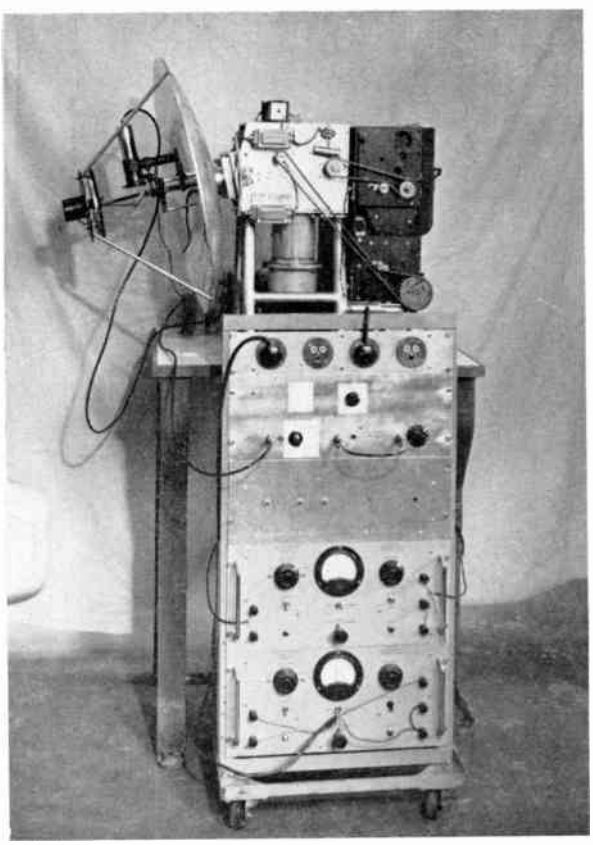


Fig. 10. The Pyroscan Mk. I.

detector utilising electrons in germanium has been operated at microwave frequencies with a magnetic field of 2 kilogauss.[18] The possibilities of extending this type of detector to wavelengths shorter than 40 microns is remote.

Mechanical scanning could be done away with if a suitable image-orthicon could be manufactured. The use of a high resistive layer of copper-doped germanium, though requiring cooling to 20° K, would provide an infra-red picture out to 13 microns with a fast response time.

3.4 *Reflection Spectroscopy*

Measurements of reflection coefficient give valuable information about the emissivity, whilst emission measurements give the product of emissivity and temperature. Thus from a combination of these measurements both temperature and emissivity may be determined independently. Any detailed study of the causes of changes in emissivity with temperature requires these combined studies. Unfortunately the currently-available sources are very weak and it may prove that the use of infra-red masers will be of interest in this field as a source of high powered infra-red radiation.

4. Conclusions

It is clear that many new techniques can have applications to medical electronics. For instance, the need to obtain information on the velocity of temperature changes requires either fast mechanical scanning systems with impurity doped detectors cooled with liquid hydrogen or else an infra-red vidicon system. Any extension of these researches into human thermometry could act as a spur for the development of such techniques.

5. Acknowledgments

Thanks are due to the Clinical Research Committee of the Middlesex Hospital Medical School, and the surgeons of the Middlesex Hospital, for facilities granted in this research.

Acknowledgment is made to the directors of S. Smith & Sons (England) Ltd., for permission to publish the material contained in Section 2 of this paper. Crown Copyright is reserved in respect of Section 3.

Figures 1 to 6 are reproduced by courtesy of the Editor of The *Lancet*.

6. References

1. R. Cobet and F. Bramigk, "Uber Messung der Wärmestrahlung der menschlichen Haut und ihre klinische Bedeutung", *Deutsches Archiv klin. Med.*, **144**, p. 45, 1924.
2. J. D. Hardy, "The human skin as a black-body radiator", *J. Clin. Invest.*, **13**, p. 615, 1934.
3. J. D. Hardy and C. Muschenheim, "The emission, reflexion and transmission of infra-red radiation of human skin", *J. Clin. Invest.*, **13**, p. 817, 1934.
4. J. D. Hardy and C. Muschenheim, "The transmission of infra-red radiation through skin", *J. Clin. Invest.*, **15**, p. 1, 1936.
5. H. Bohnenkamp and H. W. Ernst, "Untersuchungen zur den Grundlagen des Energie und Stoffwechsels. II, Mitteilung über der Strahlungsverluste des Menschen. Der Strahlungsmessung im absolute Energiemass", *Pflügers Archiv gesamte Physiol.*, **228**, p. 63, 1931.
6. K. Lloyd Williams, F. J. Lloyd Williams and R. S. Handley, "Infra-red thermometry in the diagnosis of breast disease", *Lancet*, **1961, ii**, p. 1378.
7. R. N. Lawson, "Implications of surface temperatures in the diagnosis of breast cancer", *Canadian Med. Assoc. J.*, **75**, p. 309, 1956.
8. K. Lloyd Williams, "Infra-red techniques", *Middlesex Hospital J.*, **62**, p. 13, 1962.
9. R. N. Lawson, "Thermography—a new tool in the investigation of breast lesions", *Canadian Med. Serv. J.*, **13**, p. 517, 1957.
10. C. M. Cade, "Seeing by heat waves", *J. Sci. Industr. Res.*, **20A**, p. 624, 1961.
11. R. N. Lawson, G. D. Wlodek and D. R. Webster, "Thermographic assessment of burns and frost-bite", *Canadian Med. Assoc. J.*, **84**, p. 1129, 1961.
12. D. W. Goodwin, "The noise-power spectrum of *p*-type indium antimonide", *J. Phys. Chem. Solids*, **22**, p. 401, 1961.
13. R. L. Petritz, "Fundamentals of infra-red detectors", *Proc. Inst. Radio Engrs*, **47**, p. 1458, 1959.
14. D. W. Goodwin, "Cooled photo-conductive detectors using indium antimonide", *J. Sci. Instrum.*, **34**, p. 367, 1957.
15. H. D. Adams, W. J. Beyen and R. L. Petritz, "Photoconductivity research on Cu-Ge", *J. Phys. Chem. Solids*, **22**, p. 167, 1961.
16. W. Engeler, H. Lerinstein and C. Stannard, "Photoconductivity in *p*-type indium antimonide with deep acceptor impurities", *J. Phys. Chem. Solids*, **22**, p. 249, 1961.
17. E. H. Putley, "Impurity photo-conductivity in *n*-type InSb", *J. Phys. Chem. Solids*, **22**, p. 241, 1961.
18. D. W. Goodwin and R. H. Jones, "Far infra-red and microwave detector", *J. Appl. Phys.*, **32**, p. 2056, 1961.

Manuscript first received by the Institution on 26th September 1962, and in final form on 10th January 1963.

(*Paper No. 797/MBE10*)

U.D.C. 621.382

Teaching Transistor Theory and Applications

By

J. C. CLULEY, M.Sc.†

This paper is sponsored by the Institution's Education Group

Summary: Methods of presenting transistor theory in the final year of an engineering degree course are described. After introducing small-signal theory, temperature and h.f. effects are covered and the treatment of relaxation-type circuits presented in terms of charge-control parameters. Reference is made to suitable supporting experiments.

1. Introduction

Since transistors became available in bulk some six or seven years ago, undergraduate teaching on the subject has passed through three stages. At first, short courses were given, generally in the final year, to introduce the principle of transistors and transistor circuitry to students already familiar with the theory of the thermionic valve. Although acceptable as a temporary measure, this scheme revealed considerable gaps in the background material presented in earlier years, making it unnecessarily difficult for the students to acquire a sound understanding of transistor behaviour.

Accordingly many departments have proceeded to the second stage, by modifying the content of earlier years of the course, particularly the second year, so as to provide an adequate background of physics and network theory. This is followed by a more thorough transistor course, perhaps with other courses on network analysis and materials, in the final year.

The final stage, which we have yet to reach ourselves, is the integrated treatment of both valve and transistor circuits in a combined "active networks" course, preceded by separate introductory courses dealing solely with the physical principles and characteristics of thermionic valves and transistors. Compared with two courses dealing with valves and transistors separately, this method should avoid a considerable amount of duplication, and give a clearer picture of the different applications to which transistors and valves are particularly suited.

The effect of the proposed changes in first and second year syllabuses is that the electronic engineering course will be based much more upon physics, rather than upon mechanical and civil engineering, the usual background of the "heavy" electrical engineer. This move towards more pure science has already occurred in some departments concerned only with electronic engineering.

2. Background Material

The most important preliminary course for the student of transistor theory concerns the physics of the solid state, particularly the conduction processes in pure and impure semiconductors. Whilst a comprehensive treatment can be accommodated only in a post-graduate course, a short series of lectures suffices to deal with the band structure of solids, (with major emphasis upon semiconductors), the effect of impurities, the process of diffusion, carrier mobilities and life times, a simple treatment of the *p-n* junction and an introduction to transistor action. If time permits the influence of the base width on transistor performance could usefully be included, and frequency effects.

In the field of network theory, the student should be introduced to the general theory of two-port networks, and the use of matrices and their manipulation, together with some general feedback theory and stability criteria.

The above material should ideally be given in the second year of a three-year course, but time-tables are frequently too crowded to permit this. If some of it must be deferred until the final year, it should be given as early as possible so that the final year transistor course may build upon it.

Our own students take a second year course in electronics, of which about five hours are devoted to introducing transistors and to giving a simple account of their characteristics. This reminds them that the solid-state physics course is directly related to practical engineering devices, helps to prepare them for the final year course, and also helps their understanding of several electronics laboratory experiments involving transistors. In future the second year course will include about fifteen hours on transistors, and will include some preliminary work, for example an introduction to small signal parameters, now given in the final year.

3. Final Year Course

The following scheme is suggested for a 20-hour final year course dealing with transistor theory and applications. An introductory survey serves to revise the previous years' work on the d.c. characteristics of transistors, and includes some account of their present limitations and future possibilities regarding frequency response, operating voltage, power dissipation and

† Electrical Engineering Department, University of Birmingham.

junction temperature. A cathode-ray oscilloscope display has been found a very useful teaching aid for displaying the static characteristics and load lines, and phenomena such as saturation.

The characteristics of *n-p-n* and *p-n-p* transistors may also be compared, as may those of silicon and germanium transistors. For this, we use the Tektronix 575 transistor curve tracer, which permits the display of a family of input or output curves, for up to 12 different values of base or emitter voltage or current, as well as a rapid change-over from one transistor to another, and the inclusion of a range of collector and input resistors. This is an expensive device, used mainly for research, but a simpler circuit could be devised fairly cheaply, using a standard d.c.-coupled oscilloscope or, better still, a 12- or 17-inch demonstration oscilloscope.

The curve tracer may be used to demonstrate the cut-off, active, and saturation regions of the collector characteristic, the output and input impedances of various connections, and the non-linearity of the input characteristic.

Whereas the second-year treatment of transistor operation is mainly descriptive, the final-year course should include more quantitative information. Thus by differentiating the diode equation

$$I = I_0 \left[\exp\left(qV/kT\right) - 1\right]$$

where q = electronic charge, V = diode voltage, k = Boltzmann's constant, T = junction temperature, the input impedances of the common-base and common-emitter connection may be estimated, and their dependance upon emitter current shown. The magnitude of the current gain can be estimated from the approximate expression for emitter efficiency $(1 - W/L_n.\sigma_B/\sigma_E)$ and transport factor $1 - \frac{1}{2}\,(W/L_p)^2$, where L_n, L_p = diffusion lengths for electrons and holes respectively, σ_B, σ_E = conductivities of base and emitter regions, and W = base width.

The collector multiplication factor may generally be taken as unity, but its increase at high voltages may be demonstrated on the curve tracer, as may the reduction in current gain at high emitter currents. By using a large series resistance it is possible to illustrate the "turn-over" point in the collector characteristics at high collector voltage, but this should be done at low collector current, otherwise the permissible power dissipation of the transistor will be exceeded.

A number of simple transistor models have been proposed, with the object of giving the student a clearer understanding of transistor action. The simplest of these is the two diode model as shown in Fig. 1, but this requires a current generator, αI_E, to represent the transistor action. This circuit gives an

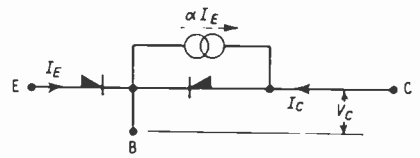


Fig. 1. Two diode model of *p-n-p* transistor.

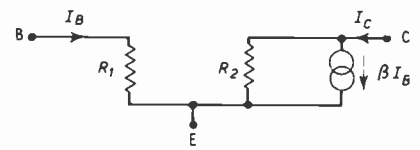


Fig. 2. Simple transistor model neglecting feedback.

adequate model of the collector voltage-current relation, for any specified value of emitter current, and of the emitter input characteristic, but it cannot be applied with much benefit to the common-emitter connection which is most frequently used.

As an intermediate step to a more accurate representation, the simplified circuit of Fig. 2 is useful, the errors involved in neglecting the feedback term being small if the load resistance and so the voltage gain are small. This is generally the case in cascade common-emitter stages. Having introduced this circuit, the simple addition of a voltage generator in the input branch serves to represent the complete small signal behaviour of the transistor in terms of *h*-parameters. Furthermore, a simplified high-frequency model of the transistor can also be produced from this circuit, as shown in Fig. 3. This is suitable only for wide-band aperiodic amplifiers, and by replacing C_c by two capacitors across the B–E and C–E ports, a very simple network is produced which is nevertheless a useful aid to computing the high frequency behaviour of the transistor. Again this serves as an introduction to the complete hybrid-π circuit required in the treatment of tuned amplifiers.

3.1. *Parameters and Small-signal Theory*

The course proper starts with the consideration of various methods of specifying the transistor characteristics in terms of matrix parameters, pointing out the reasons for the general acceptance of the hybrid parameters. This is followed by a review of the possible

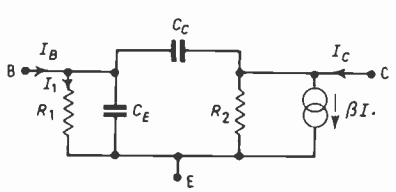


Fig. 3. Simple transistor model valid at high frequencies, neglecting resistive feedback.

low-frequency equivalent circuits, and the characteristics of the common-base, common-emitter, and common-collector connection. Most students find that the analogy between these connections and the use of the thermionic valve in the common-grid, common-cathode, and common anode connection is helpful, but as the teaching of active circuit theory comes to rely more on transistor examples, and less on the use of thermionic valves, such an analogy may become less helpful.

This treatment of parameters and equivalent circuits should also include if possible the conversion of parameters, both from, say, hybrid to admittance parameters, hybrid to T and π parameters, and from common-emitter to common-base and common-collector hybrid parameters. It is essential to avoid over-loading the student with masses of algebra, and the method we have used consists in going over one conversion thoroughly, and providing the student with duplicated sheets giving the remainder. One or two other conversions can be given as tutorial examples. It is important not to devote too much time to these processes, as they are purely manipulative and make little contribution to the student's grasp of transistor principles.

Using the hybrid parameters, the expressions for input and output impedance in terms of load and source resistance are then derived, with typical curves for the three connections plotted on logarithmic scales, to give the student an impression of the orders of magnitude concerned. This is followed by expressions for voltage current and power gain, with numerical examples. The material mentioned in this paragraph is also emphasized by laboratory experiments in which the computed quantities are measured, and curves for example of input resistance as a function of load resistance, plotted and compared with theoretical predictions.

By this time the student should be able to compute the small-signal, low-frequency behaviour of any transistor stage, and also the effect of, for example, simple feedback circuits, such as the use of a resistor between collector and base, or in series with the emitter of a common-emitter stage.

3.2. *Temperature Effects*

Since the design of transistor circuits generally involves problems of thermal stability, the various effects of junction temperature changes should be discussed before dealing with amplifiers and other applications.

The most important factor is the temperature dependance of leakage current I_{co}, which may be derived from the expression $I_{co} = A \exp(-qV_g/kT)$, where A is a constant, V_g is now the energy gap potential (720 mV for germanium), and the other symbols are as before.

$$\text{Thus } \frac{\partial I_{co}}{\partial T} = I_{co} \times \frac{qV_g}{kT^2} = I_{co} \times 0{\cdot}0985/\text{deg C}$$

for germanium at 300° K.

By combining this expression for I_{co} with the diode equation $I = I_0[\exp(qV_j/kT) - 1]$, and using the condition that I is constant, the incremental relation between external junction voltage V_j and temperature T may be calculated, giving an approximate relation between the base-emitter voltage of a transistor and temperature—of the order of 2 mV/deg C—at constant emitter current. This is an important parameter in assessing the zero drift in d.c. amplifiers.

The expressions for stability factor $S = \partial I_c/\partial I_{co}$ for one- and two-battery systems may then be derived, and their consequences in circuit design mentioned. This is followed by a derivation of the thermal stability condition for transformer and resistance-coupled amplifiers, considering the relation between collector power dissipation p_c, junction temperature T_j, I_{co}, and I_c as a closed-loop system with positive feedback. The stability limit is thus the condition of unity loop gain. The simplest case is that of transformer coupling, in which the primary resistance is neglected, so that $p_c = I_c \times V_b$, where V_b is the supply voltage. The resistance-coupled case is treated by replacing the relation $\partial p_c/\partial I_c = V_b$ by $\partial p_c/\partial I_c = V_b - 2I_cR_1$, where R_1 is the resistance of the load. The loop gain is clearly reduced, and if $I_c.R_1 = \frac{1}{2}V_b$, the "half-power" condition, it falls to zero thus preventing thermal instability.

The final temperature effect is that of a change in β, and the three effects may be compared on the basis of equivalent input signals to an amplifier, as a function of the source resistance. The relative magnitude of the three components should also be compared for germanium and silicon transistors.

Some practical data may be given at this stage, such as the thermal resistance of typical transistors and heat sinks, and the recommended maximum junction temperatures. This information not only indicates the magnitudes involved, and present limitation, but in addition is a useful basis for tutorial design problems.

3.3. *Linear Amplifiers*

At this stage the material previously presented is used to analyse and design a.c. and d.c. voltage amplifiers, including the bias networks. Temperature effects are particularly involved in d.c. amplifier performance, and the various methods used to minimize zero drift should be mentioned. The equivalent circuit for the transistor may be used, for example to discuss the effect on the frequency response of a typical emitter bias circuit, at the lower end of the spectrum, but no treatment of the high frequency response is given at this stage.

This is followed by a treatment of Class A and Class B power amplifiers, with particular emphasis on the circuits such as the single-ended push-pull stage, and complementary transistor stages which are inconvenient or impossible with thermionic valves. Other topics which may be included are the use of temperature-dependent resistors to improve thermal stability, the various composite circuits and some special techniques for increasing amplifier input impedance. The difficulties caused by transistor frequency effects in conventional feedback circuits may also be mentioned.

3.4. *High Frequency Effects*

A full treatment of the high-frequency behaviour of the transistor is possible only at postgraduate level —in undergraduate courses there is time only for an introduction to the subject in terms of the simpler equivalent circuits, and an indication of the properties needed—e.g. thin base region—for high-frequency operation.

The average student easily becomes confused by the large number of equivalent circuits given in various text-books, many of them complicated, and some containing frequency-dependent elements. He needs to be reminded that they are all lumped-circuit approximations to a distributed parameter situation, and are to this extent inexact. The best equivalent circuit from the teaching aspect is the simplest which fits the observed data reasonably well, and which contains no frequency-dependent elements. Thus the simple circuit of Fig. 3 forms a useful introduction to the more complicated circuits, and can be used for amplifiers where the voltage gain is low. For other cases, the full hybrid-π circuit can be used, following on directly from Fig. 3.

Given R_1, the value of C_E can be estimated in terms of the α cut-off frequency by considering that only the current I_1 flowing in R_1 contributes to transistor action, and the input capacitance can be calculated, and its dependence upon the voltage gain, and the possibility of oscillation with an inductive load can be determined. The bandwidth of an aperiodic amplifier may be estimated by reducing either the hybrid-π or the simpler circuit of Fig. 3 to a simple input branch, by replacing C_c by two equivalent capacitors across the input and output ports.

It is important that the student should distinguish between those features of the equivalent circuit, such as C_E and g_m, which represent the behaviour of an idealized transistor, and the components such as $r_{bb'}$ which are parasitic and represent the practical imperfections of a real transistor. Although the initial description of transistor action can be given in terms of an ideal transistor, it is essential to include $r_{bb'}$ in any realistic discussion of high-frequency performance, since together with C_E it ultimately limits the high-frequency gain.

Additional topics which can be included in this section are the unilateralization of tuned amplifiers, and a simple treatment of oscillators.

Suitable laboratory experiments to accompany this part of the course would be the measurement of common-base and common-emitter cut-off frequencies, the measurement of transistor parameters at high frequencies using an r.f. bridge, the bandwidth of wide-band amplifiers, and, using analogue methods at low frequencies, the behaviour of the more complicated equivalent circuits.

Switching and pulse circuits constitute a field in which the transistor is generally superior to the valve, and in which transistor circuit techniques are markedly different from those used with valves. It has consequently been our policy to devote adequate time to this section of the course to give the student at least a grounding in the important multivibrator circuits, if necessary by reducing the time spent on topics such as sinusoidal oscillators, in which transistor and valve circuits are largely similar.

The introduction is made by considering the steady-state transistor behaviour in the cut-off and saturation regions, estimating its impedance, and comparing it with alternative devices such as the mechanical switch and the thermionic valve. The two-diode equivalent of Fig. 1 is useful in the study of the cut-off region, with the inclusion of the two depletion capacitances, but it is of little value in the saturation region, as the transistor collector impedance is much lower than that of a diode. It may be estimated by using the equation for I_c as a function of V_c and differentiating. This should be illustrated by substituting typical values in the equation, and the effects of the additional series resistance in a practical transistor pointed out.

Having established the behaviour of the transistor in the saturation and the cut-off modes, some simple multivibrator circuits can be discussed, and the important design criteria established. The consequences of saturation can be pointed out, and attention drawn to methods of avoiding it in high-speed circuits.

Important features of relaxation circuits are switching speed, and the magnitude of any delay effects. These are generally governed by transistor performance rather than external circuit capacitance, and are best estimated using charge-control parameters. A full treatment of this method of analysing switching performance must generally be deferred to post-graduate courses, but a simple introduction has been given at undergraduate level.

This method of analysis of switching behaviour avoids the use of small-signal parameters, which

change with collector current, and deals with the charge which must be injected into the base region to produce a certain collector current. Starting from the threshold of conduction, this may be divided into three categories, Q_B, the charge required to establish the collector current I_c at constant collector voltage, Q_V the charge necessary to change the voltage across the depletion layer capacitance C_d, and, if the transistor saturates, the extra charge Q_{BS} caused thereby.

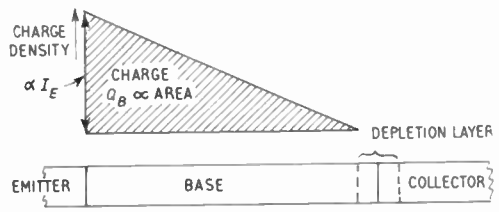


Fig. 4. Normal charge distribution in base region.

By reference to a charge distribution such as that shown in Fig. 4 for the active mode, it may be shown that the charge Q_B is proportional to emitter current, and thus, approximately, to collector current. Students with an adequate background of semiconductor theory will be familiar with the approximation

$$Q_B = \frac{I_c \times W}{2D_p},$$

where D_p is the diffusion constant (in this case for holes) and W the base width. The expression for T_c, the collector time-constant, is thus derived as

$$T_c = \frac{B}{I_c} = \frac{W^2}{2D_p},$$

showing the importance of a thin base region in minimizing Q_B and thus obtaining rapid switching.

The charge Q_V may be evaluated as

$$\int_{V_1}^{V_2} C_d \,.\, \mathrm{d}V$$

(since C_d is voltage-dependent) if the collector voltage excursion is known, while Q_{BS} may be shown to be proportional to the "surplus" base current $(I_B - I_c/\beta)$. β is strictly the large-signal current gain in this expression.

In terms of these charges the stages involved in switching transistors on and off may be described, and a simple explanation given for the delay time when a saturated transistor is being switched off.

This leads to the design of input circuits for switching transistors, and to a simple treatment of switching times. A more rigorous solution, for constant base current switching, yields exponential wave-forms for collector current and voltage (with a resistive load), similar to those derived by Moll for a simple model such as that given in Fig. 2, with a frequency-dependent current gain of the form

$$\alpha = \frac{\alpha_o}{1 + \mathrm{j}\left(\frac{\omega}{\omega_\alpha}\right)}$$

The charge-control analysis is more fundamental, however, and takes account of Q_V, which Moll's analysis neglects.

The full treatment of switching times, and related topics such as the "on-demand" current gain, are generally left until the postgraduate course, otherwise the undergraduate course will be overcrowded.

It is considered that this part of the course should be given at some length, as the material is not contained in the more common general text-books, and has appeared only in individual papers and in a few books for the specialist. At the same time it is directly related to the fundamental behaviour of the transistor, and uses a considerable amount of the theory established in the earlier courses on semiconductors and *p-n* junctions.

A considerable amount of practical work may be used to illustrate this part of the course; the saturation region of a transistor may be measured in the steady state, and the collector resistance derived; the switching parameters may be measured (preferably using a low-frequency transistor unless fast oscillographs and pulse generators are available); the effect of base drive on switch-on, switch-off and delay time investigated; and the relation between the collector time constant and the current-gain cut-off frequency determined experimentally.

A number of complete circuits may also be investigated, such as the symmetrical multivibrator as a counter and shifting register, and the relation between base resistor and pulse duration in monostable and a stable multivibrator. For demonstrating circuit "building" we have a transistor counter which has four binary stages which may be arranged as a binary counter, or as a decimal counter in several ways. The various gates, the pulse generator, the inverters and the time delay units are interconnected by plugs and sockets, and are provided with lamps to indicate the signal magnitude they emit. This device is used for demonstration with large undergraduate classes, and for students' laboratory work at postgraduate level.

3.5. *Other Semi-conductor Devices*

It is generally convenient to use transistors in conjunction with other solid-state devices such as Zener diodes and silicon controlled rectifiers, and the properties of these devices and their capabilities and fields of application should be mentioned. In particular, the application of the silicon controlled rectifier to high

power rectifiers and inverters is of particular interest to students who also take courses in power utilization and control systems, and both the teaching and experimental work on these courses can profitably be linked with parts of the transistor course. The tunnel diode is also coming into production, and is of interest in high-speed pulse circuits and in logical and storage units in digital computers.

4. Conclusion

Most students appear to find transistor theory and applications more difficult than a corresponding course based upon thermionic valves. This is partly a question of familiarity—all introductory physics courses mention the valve, as do most school text-books, and most laboratory courses deal with valves before introducing transistor experiments. Nevertheless there are two main difficulties for the student; the more complicated series of physical processes involved in transistor action, and the more complicated network theory inherent in the bilateral nature of the transistor. Thus whereas a simple derivation of the three-halves power law for a planar triode can be given in the first or second year, the corresponding material concerning transistor characteristics is much more involved and must be deferred until the final or the postgraduate year.

Also the need for a more complex matrix or equivalent circuit to represent transistor behaviour makes calculations of amplifier performance more difficult. In many cases simpler models such as those in Fig. 2 or Fig. 3, give a sufficiently accurate picture, and they should be used if possible.

A further minor complication is the need to consider, and design for, adequate thermal stability in transistor circuits.

On all these topics the transistor course involves more work than the corresponding thermionic valve course, and it is consequently important to begin the course at a gentle pace, and to expect the student to cover appreciably less ground than he would in a course of comparable length dealing with thermionic valves.

The introduction of the transistor is only part of the present tendency to use more solid-state components in all branches of electrical engineering, and this trend must be reflected in teaching syllabuses. The transistor and its associated devices are thus bound to encroach more and more upon the time available for electronics in University teaching, but by unifying as far as possible the treatment of all active circuits, both sinusoidal and relaxation, it should be possible to introduce a useful and stimulating transistor course without excising too much other material.

5. Appendix

The choice of a suitable text-book for a transistor course will depend upon the time available in the final year and upon the amount of background material given in earlier years. The undergraduate will usually find a general text-book most convenient, whereas at postgraduate level a number of more specialized works are required, in addition to individual papers and articles.

The following list (by no means a comprehensive one) is suggested as a basis for further study to amplify the lecture course.

5.1. *Undergraduate Course*

S. W. Amos, "Principles of Transistor Circuits". (Iliffe, London, 1959.)

F. C. Fitchen, "Transistor Circuit Analysis and Design". (Van Nostrand, New York, 1960.)

M. V. Joyce and K. C. Clarke, "Transistor Circuit Analysis". (Addison-Wesley, Reading, Mass., 1961.)

A. W. Lo, "Transistor Electronics". (Prentice-Hall, Englewood Cliffs, N.J., 1955.)

R. F. Shea (Ed.), "Transistor Circuit Engineering". (Wiley, New York, 1957.)

E. Wolfendale (Ed.), "The Junction Transistor and its Applications". (Heywood, London, 1958.)

J. R. Tillman and F. F. Roberts, "Theory and Practice of Transistors". (Pitman, London, 1960.)

5.2. *Postgraduate Course*

W. W. Gartner, "Transistors, Principles, Design and Applications". (Van Nostrand, New York, 1960.)

R. A. Greiner, "Semi-conductor Devices and Applications". (McGraw-Hill, New York, 1961.)

R. D. Middlebrook, "An Introduction to Junction Transistor Theory". (Wiley, New York, 1957.)

P. A. Neeteson, "Junction Transistors in Pulse Circuits". (Cleaver-Hume, London, 1959.)

A. I. Pressman, "Design of Transistorized Circuits for Digital Computers". (Rider, New York, 1959.)

The following may be useful if an integrated treatment of valve and transistor circuit is given:

A. J. Cote and J. B. Oakes, "Linear Vacuum-Tube and Transistor Circuits". (McGraw-Hill, New York, 1961.)

J. G. Linvill and I. F. Gibbons, "Transistors and Active Circuits". (McGraw-Hill, New York, 1961.)

J. Millman and H. Taub, "Pulse and Digital Circuits". (McGraw-Hill, New York, 1956.)

J. M. Pettit and M. M. McWhorter, "Electronic Amplifier Circuits". (McGraw-Hill, New York, 1961.)

Manuscript received by the Institution on 24th May 1962. (Paper No. 798/Ed7)

U.D.C. 621.382.3

Transistor and Semiconductor Teaching

By

Professor M. R. GAVIN,
M.B.E., M.A., D.Sc.(*Member*)†

Presented at a meeting of the Education Group in London on 1st November 1961.

Summary: The starting point for teaching transistors must depend on the type of student. For all students, including technicians, who have to study transistors it is desirable that some kind of physical picture, however simple, be given. One way of doing this, based on a two-diode concept of the transistor, is suggested in the first part of the paper. For the degree or diploma student who will ultimately become a professional radio engineer the author believes that a good background of solid state physics should accompany any instruction on transistors. The education of this type of student in general and the need for an understanding of semiconductors in particular is dealt with in the latter part of the paper.

1. Introducing Transistors

There are many ways in which the subject of transistors can be introduced. As an example of one extreme method there is a recent American textbook on transistors of some 500 pages. This is divided into three parts. The first, of 200 pages, is on the physics of semiconductors and aims at explaining the nature and properties of various types of transistors. Part 2 has about 150 pages and deals with the theory of four-terminal networks. Finally, Part 3 deals with transistor circuits and we have reached page 371 before there is any sign of a circuit using a transistor. (I may say in passing that in many respects this is a good book.) At the other extreme we may introduce the transistor as a device with three terminals which has responses which can be measured in terms of characteristic curves. Certain parameters can be defined and evaluated from these curves and the parameters can be used to predict the performance of the transistor at low frequencies when it is connected in various circuits. This is the "black-box" approach and quite a lot of successful work can be done using this method. Everyone uses it to some extent. Most of us would favour an approach to the subject of transistors somewhere between these two extremes, with some attempt to provide a physical picture of the operation. This can be done thoroughly only when the student has a good appreciation of the physics of semiconductors and such an appreciation is most desirable in the professional radio engineer. However, even when there is no background of semiconductor physics it is preferable to give some acceptable picture of transistor operation. One method of doing this is suggested in this paper.

† Department of Electronic Engineering, University College of North Wales, Bangor, Caernarvonshire.

A start may be made by assuming a simple knowledge of the *p–n* junction diode. The $I-V$ characteristic of such a diode can be measured (Fig. 1). The transistor consists of two such diodes joined back-to-back with a common electrode, known as the base (Fig. 2(*a*)). The characteristics of the two diodes can be confirmed by measurement. In normal use one of these diodes is biased in the forward

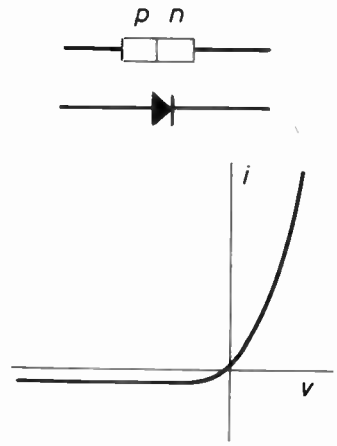


Fig. 1. *p–n* junction diode and its characteristic.

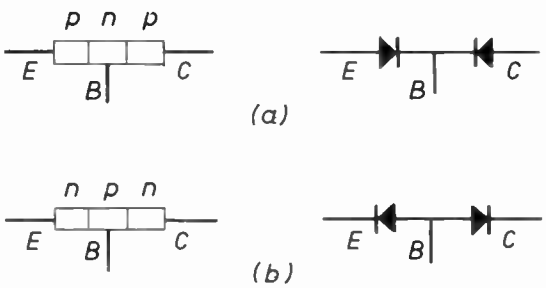


Fig. 2. *p–n–p* and *n–p–n* transistors.

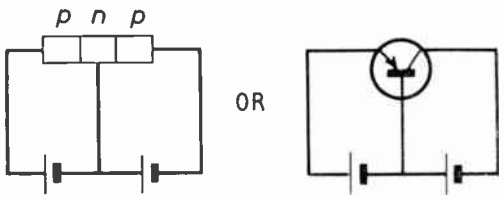


Fig. 3. Bias circuit for *p–n–p* transistors.

direction and the other in the reverse direction (Fig. 3). When the base is the *n*-section of the two diodes, the electrode to which the positive bias is connected is the emitter and the third electrode is the collector. (This is the *p–n–p* arrangement; the *n–p–n* transistor, Fig. 2(*b*) may also be mentioned at this stage.) When current flows in through the emitter and through the forward-biased diode it has two alternative paths. It may either flow out through the base or through the other diode. An outstanding property of the transistor is that practically all of the current flows through the second diode to the collector. Note that this statement involves the major part of transistor physics. It is essential to point out at this stage that two separate diodes could not give the same characteristics. The fraction, α, of the emitter current that reaches the collector is an important parameter of the transistor. The total collector current includes the small current i_{CO} through the reverse-bias diode.

Then $$i_C = \alpha i_E + i_{CO}$$

The base current is given by

$$i_B = (1-\alpha)i_E - i_{CO}$$

When there is no emitter current the collector-base diode is operating on the reverse part of its characteristics (curve $i_E = 0$ in Fig. 4). When some emitter current is introduced most of it flows to the collector and the new curve is almost the same as before but is displaced by the amount αi_E and so the output characteristics may be established. The nature of these characteristics can be confirmed by measurement

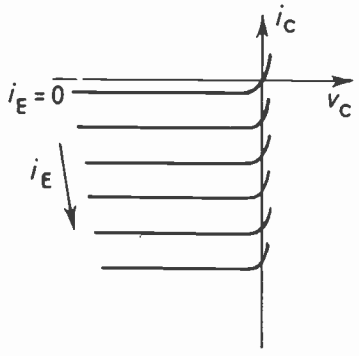


Fig. 4. Output characteristics for *p–n–p* transistor.

and it is found that over an appreciable range of i_E, α remains practically constant.

In actual use the main interest is in changes in currents and voltages, and it is more usual to define α in terms of changes in i_C and i_E, that is,

$$\alpha = i_c/i_e$$

where the lower case suffixes indicate small changes.

Then $$i_b = (1-\alpha)i_e$$

and $$i_c = \frac{\alpha}{1-\alpha}\, i_b$$

The quantity $\alpha/(1-\alpha)$ is called α_{cb} or β. Since α is nearly equal to unity, β is usually large (10 to 200).

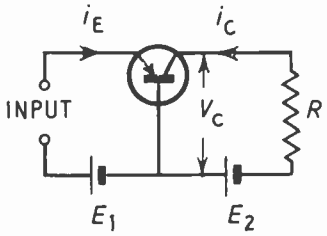


Fig. 5. Simple transistor amplifier.

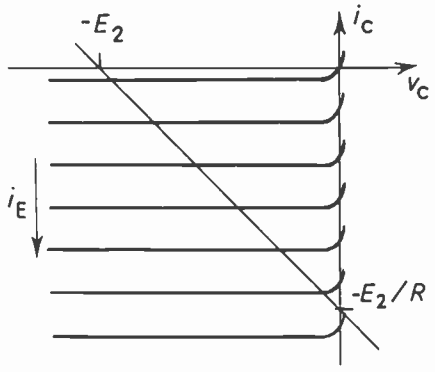


Fig. 6. Output characteristics and load line for transistor amplifier.

When a signal is applied to the input terminals the current i_E changes. As a result there is a variation in the output current i_C which may flow in a load R (Fig. 5). The conditions in the output circuit can then be determined from the equation

$$v_C = -E_2 - Ri_C$$

or from a load line drawn on the collector characteristics (Fig. 6).

When the signal is applied to the input terminals it must be remembered that the emitter-base portion of the transistor is operating as a forward biased diode and so will act as a low resistance to the signal, i.e. the input resistance of the transistor is of the order of

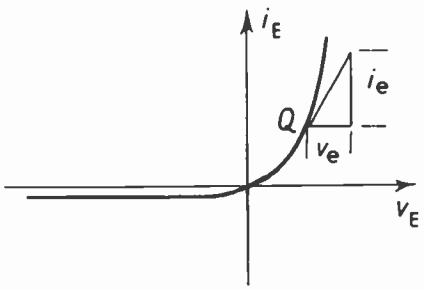


Fig. 7. Input characteristic of transistor.

the slope resistance of the emitter-base diode at the operating point, Q, in Fig. 7. This slope resistance varies appreciably with the emitter current and there is therefore a non-linear input circuit. Unless the signal is very small there can be considerable distortion. The input diode-slope resistance is

$$r_e = v_e/i_e$$

measured at constant v_C.

An approximate equivalent circuit for small changes is shown in Fig. 8. This may be used to determine the performance of an amplifier with a load resistance R across the output terminals. If the input voltage changes by a small amount v_i then the change in emitter current i_e is given by

$$i_e = v_i/r_e$$

The output voltage, v_o, is given by

$$v_o = i_c R = \alpha i_e R$$

Thus the voltage gain A_v is

$$A_v = v_o/v_i = \alpha R/r_e \simeq R/r_e$$

This equivalent circuit ignores the output resistance, r_c, which is the slope resistance of the reverse-bias characteristic

$$r_c = v_c/i_c$$

measured at constant i_E.

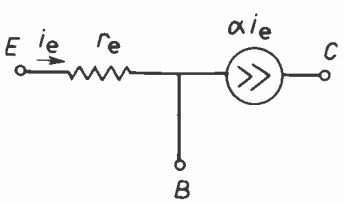


Fig. 8. Simple equivalent circuit of transistor.

This has a much higher value than r_e and may frequently be neglected. To give a more accurate equivalent circuit based on our physical picture of the transistor it is necessary to include a resistance r_b as shown in Fig. 9. This depends to some extent on the resistance of the base material between the diode junctions and the external base connection. The values of r_e and r_b in this equivalent circuit depend on the value of the collector voltage on account of the Early Effect. It is not feasible to attempt any complete explanation at this stage but it is desirable that the students should realize that r_e may be appreciably less than the forward diode slope resistance and that r_b may be quite considerable. Typical values of the quantities shown in Fig. 9 are

$$r_e = 20\ \Omega, \qquad r_c = 1\ \text{M}\ \Omega,$$
$$r_b = 1000\ \Omega, \qquad \alpha = 0{\cdot}98$$

This circuit can be used to determine the behaviour of the transistor when connected to a low-frequency signal source and a load, provided of course that the operation is limited to small signals. This means restricting v_e to a few millivolts.

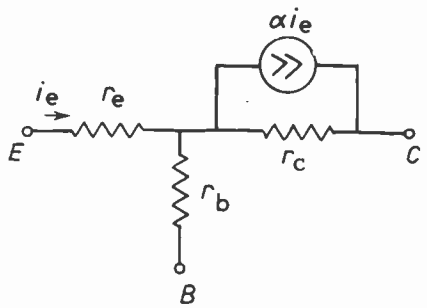


Fig. 9. Transistor equivalent circuit including output resistance and base resistance.

So far the development has been given in some detail to try to suggest how a reasonable picture of transistor operation may be given to students with little physical background. No doubt there are many other ways in which the approach could be made. The subsequent development of the subject is also capable of much variety and below only a very brief outline is given of one possible method.

No mention has been made of common-base, common-emitter or common-collector operation. All of these could be studied in terms of the equivalent circuit of Fig. 9 by appropriate connection of the external circuits. In studying actual circuits greater emphasis should be laid on the common-emitter. In establishing the values of the quantities, such as input resistance, the simple two-diode picture already described should be used to explain the qualitative difference of the various circuit connections whenever possible.

2. Relation to Triodes and Pentodes

It is assumed that some similar introduction will have been given to valves so that their characteristics

and simple properties are known. It is worth determining and discussing comparable values for transistors, triodes and pentodes of r_e, r_c, r_b, α, g_m, μ, and r_a.

Subsequent study of amplifiers (current, voltage and power), frequency response, feedback, switching, etc. should, to a large extent, proceed concurrently with valves and transistors.

3. Four-terminal Network Theory

It is usual to use linear four-terminal network theory in studying transistors in amplifying and other circuits. The generalized ideas of network parameters giving input and output impedance, transfer and feedback impedance, and corresponding admittances are all very useful and instructive. However, at all stages it is desirable to relate these to a physical picture of the transistor and its circuits. An excess of linear network analysis at the early stages of a course may be harmful unless the limitations of the linear theory are constantly stressed. This type of analysis may well be more enlightening to network theory than to transistor engineering.

4. Semiconductors

In considering above a possible introductory course on transistors no knowledge of the physics of semiconductors has been assumed on the part of the student. Students who will ultimately be research and development engineers in industry or government departments should, however, have a good background of the theory of semiconductors and solid-state physics. All radio engineers may not agree with this opinion.

Many people who received their basic college education more than 10 years ago will have experienced the problem of getting used to completely new developments, new materials and even new ways of thinking. The introduction of the waveguide just before the war is one example. Many electrical engineers, who had been brought up strictly on circuit concepts, and looked on flow and return conductors as essential in power distribution, found the waveguide a great mystery. Those who were familiar with the ideas of the electromagnetic field, with or without Maxwell's Equations, found it easier to accept the waveguide. Although they were confronted with quite a new development they had the background of basic principles which enabled them to understand and accept it. To-day the electronic engineer is being inundated with new developments. Semiconductors are revolutionizing the whole of electronics and they are also making a big impact on power generation and utilization. Silicon and germanium rectifiers, transistors, tunnel diodes, parametric amplifiers, photoelectricity, thermoelectric generators, dielectrics, ferrites, ferroelectrics—these are a few of the fields and devices which all depend on new materials, many of them semiconductors. How many of us have the necessary background to enable us to accept these inovations willingly, let alone to enable us to understand them and perhaps help to develop them? One of the essential requirements of a first degree course should be to equip the graduate with the broad background of basic principles which will serve him in good stead for his working life. This should come before the special techniques of the power engineer, the communications engineer, the radar engineer, the control engineer or the device engineer. Special techniques are more appropriate to the postgraduate stage and should be the responsibility of industry, supplemented by intensive postgraduate courses.

5. The Value of Courses in Materials Technology

In order to achieve the essential background it is suggested that the undergraduate course should consist of three main parts—electrical science, mechanical science and materials technology. The elaboration of this theme would not be appropriate or possible in the course of this paper. However, all engineers or applied scientists (and pure scientists for that matter) use materials in great variety—metals, liquids, gases, conductors, insulators, semiconductors, magnetic materials, glasses, plastics, fibres, rubbers, phosphors, etc., etc. Whatever new developments take place the engineer will always have to use materials and it would seem that some knowledge of the nature and properties of materials should constitute an essential part of his education. Materials already play a large part in the courses of physicists, chemists and metallurgists. In considering the course for engineers we should draw on the disciplines of physics, chemistry and metallurgy whenever necessary, but the whole should be integrated in the light of the engineer's special needs as an applied scientist. In such a course for electrical engineers conductors, insulators, semiconductors and magnetic materials would obviously play a large part and would necessitate a good background of the physics of solids.

Here perhaps it should be repeated that the foregoing applies to graduates who will ultimately be research and development engineers. Real understanding of transistor operation based on energy-band theory demands some knowledge of quantum mechanics of solids. High frequency or pulse performance of transistors is all the better appreciated for a knowledge of such physical phenomena as diffusion and carrier storage. There is still a need for a great army of well-educated technicians and these are vitally important and essential people. It is doubtful if they can have much in the way of semiconductor theory. For them perhaps some approach along the

lines suggested in the first part of this paper would be appropriate. At the same time some physical picture is essential to help to illustrate the principles of operation at all stages and all levels. However, there is considerable danger in partial explanations. An advantage of the two-diode picture of the transistor is that it is reasonably correct. As an example of a partial explanation that is often misleading one may take the simple explanation of the rectification properties of *p–n* junction diodes based on the potential barrier at the junction. The same statements could often be applied to the contact potential barrier between two metals. Unless the explanation is sufficiently complete to distinguish between rectification and non-rectification in the two cases it would be much better omitted.

A study of materials such as is advocated here requires a considerable amount of time. Many may say that the time is not available or it is not justified. However, time given to this study would be more profitably spent than if it were given to technical drawing, heat engines, design of electrical machines, television, or some of the many other subjects that are involved largely in techniques and empiricisms, which happen to be in vogue at the moment. Study of such subjects can be made all the more effectively if and when required, if the student has an adequate background of basic principles.

6. Bibliography

"Principles of Transistor Circuits", S. W. Amos, 2nd Edition (Iliffe, London, 1961) (Elementary).

"Introduction of Transistor Circuits", E. H. Cooke-Yarborough (Oliver and Boyd, London, 1957) (Elementary).

"Transistor Electronics", D. De Witt and A. L. Rossoff (McGraw-Hill, New York, 1957) (More advanced).

Manuscript received by the Institution on 26th January 1962 (Paper No. 799/Ed8).

INSTITUTION NOTICES

1963 Convention Banquet

Members are reminded of the Convention Banquet, to be held in the Guildhall, Southampton, on Friday evening, 19th April. Members may also be accompanied by their Ladies.

The cost of tickets, obtainable from the Institution, is £2 10s.

Symposium on Processing and Display of Radar Data

The Radar and Navigational Aids Group Committee is sponsoring a one-day Symposium on "Processing and Display of Radar Data" at the London School of Hygiene and Tropical Medicine on **Thursday, 16th May.** The following papers will be presented:

Morning Session (11–12.30)

"Man and the Machine in the Extraction and Use of Radar Information"—R. Benjamin (*A.S.W.E.*).

"Design and Application of Computers for Radar Data Processing"—R. A. Ballard and L. Moore (*A.S.W.E.*).

Afternoon Session (2–5.30)

"Automatic Radar Data Extraction by Storage Tube and Delay Line Techniques"—J. C. Plowman (*A.S.W.E.*).

"Digital Automatic Data Extraction Equipment"—J. V. Hubbard (*A.S.W.E.*).

"Digital Data Processing Considerations in Radar"—P. J. Child (*Birmingham University*).

"Display of Automatically Processed Radar Information"—D. R. Jarman (*A.S.W.E.*).

"High-reliability Display Systems for use with a Digital Complex"—R. F. Hansford (*Decca Radar*).

"Recent Developments and Future Trends in Radar Display"—D. W. G. Byatt (*Marconi's*).

Reprints of the papers will be prepared and advance registration will be necessary.

The charges are: Members of the Brit.I.R.E., £1; Student members of the Brit.I.R.E., 10s.; Non-members, £1 10s. Symposium and registration forms will be available from the Institution after 1st April.

Annual General Meeting of the B.C.A.C.

In the Annual Report of the Council of the British Conference on Automation and Computation, presented at the Second Annual General Meeting of the reconstituted body held on 2nd October, an encouraging review was given of the Conference's activities during the past year and its proposals for the future. The Annual Accounts were presented and approved and adopted; the subscriptions of the members of the Society for the year commencing 1st January 1963 were fixed at £20 as before.

The meeting elected Honorary Officers and Executive Committee for the coming year as follows:

Sir Walter Puckey, Chairman; Professor G. D. S. MacLellan, Vice-Chairman; Mr. C. Mead, Vice-Chairman; Sir Steuart Mitchell, Vice-Chairman; Mr. S. M. Rix, Honorary Treasurer; Mr. F. Jervis Smith, Honorary Secretary.

Executive Committee:

The Honorary Officers and Mr. S. W. Adey, Dr. E. H. Bateman, Mr. E. C. Clear Hill, Mr. J. F. Coales, Mr. J. Cooper, Mr. D. du Pre, Mr. W. M. Larke, Sir Charles Norris, K.B.E., C.B., D.S.O., Dr. J. M. S. Risk, Mr. T. G. P. Rogers, Mr. G. M. E. Williams, Mr. W. F. S. Woodford.

The Institution was represented at the meeting by Mr. W. Renwick and Mr. A. St. Johnston (Members).

Symposium on Electronics, Instrumentation and Production

The outline programme of the above Symposium, announced in the December *Journal*, which is being sponsored by the South Western Section of the Institution and other engineering Institutions, is now completed. The Symposium will be held at the Bristol College of Science and Technology on 12th and 13th June, 1963. The papers to be presented include the following:

Wednesday, 12th June

"The Implications for Management of Automative Change dependent on Electronic Devices"—Stafford Beer (*Sigma*).

"P.E.R.T. Procedures for Project Organization"—G. P. Tonkin (*Bristol Aircraft*).

"Production Control by Computer"—H. S. Woodgate (*International Computers and Tabulators*).

"Solid-state Control Systems"—G. B. Kent (*Newman Electronics*).

Thursday, 13th June

"Future Possible Developments of Automation"—A. Battersby (*Work Study College, Cranfield*).

"Electronic Aids to Materials Handling"—D. E. Tyzack (*E.M.I. Electronics*).

"Electronics in the Modern Steelworks"—J. K. Edwards (*Steel Company of Wales*).

"Automatic Positioning Devices as Aids to Production"—K. J. Coppin (*Ekco Electronics*).

"Automative Systems in Production Control"—I. A. Dempster (*City of Plymouth College of Technology*).

"Ultrasonic Cleaning as an Aid to Productivity"—A. E. Crawford (*Sonics Division, Elliott Brothers*).

Further details, including registration forms, may be obtained from the Honorary Local Secretary of the South Western Section, W. C. Henshaw, M.Sc., 3 Northwick Road, Bristol 7.

U.D.C. 621.375/6

Circuits with Time-Varying Parameters
(Modulators, Frequency-changers and Parametric Amplifiers)

By

Professor
D. G. TUCKER, D.Sc.
(*Member*) †

This paper is sponsored by the Institution's Education Committee.

Summary: It is shown that the theory of circuits with time-varying parameters may be presented in a manner sufficiently simple for undergraduate instruction, and that the basic conceptions of modulators with complex impedances and of parametric amplifiers may thus be made clear. Moreover, the treatment is general in nature, and so brings out the similarities and differences between circuits with time-varying resistance and those with time-varying inductance and capacitance. Non-linear effects are also touched upon.

1. Introduction

There is no doubt that, as with so many branches of engineering nowadays, the subject of circuits with time-varying parameters is often treated in such an advanced mathematical manner (see, e.g., the series of papers in the *I.R.E. Transactions on Circuit Theory*, March 1955) that it seems to be quite beyond the scope of undergraduate work. It is possible, however, to present the basic concepts very simply, and this simple treatment is adequate for the solution of many important practical circuits in the steady state. Examples of such circuits are amplitude modulators and parametric amplifiers.

There is nothing in Kirchhoff's Laws to prevent their application to time-varying circuits, and we therefore start with a statement of these laws for single loop and single node-pair circuits containing, in addition to ordinary impedances, elements of resistance, inductance or capacitance which are caused to vary with time.

2. Kirchhoff's Laws for Time-varying Circuits

The basic circuits are shown in Fig. 1. The single node-pair circuits can be re-arranged, as will soon become evident, as exact duals of the single loop circuits. Thus the equations may be written in dual pairs, where

$$Y = 1/Z, \qquad g(t) = 1/r(t).$$

Thus from Fig. 1(*a*), (i) and (ii) respectively,

$$E\cos\omega_q t = Z.i(t) + r(t).i(t) \qquad \text{.....(1)}$$

$$I\cos\omega_q t = Y.v(t) + g(t).v(t) \qquad \text{.....(2)}$$

From Fig. 1(*b*) (i) and Fig. 1(*c*) (ii) respectively,

$$E\cos\omega_q t = Z.i(t) + \frac{\mathrm{d}}{\mathrm{d}t}[L(t).i(t)] \qquad \text{.....(3)}$$

$$I\cos\omega_q t = Y.v(t) + \frac{\mathrm{d}}{\mathrm{d}t}[C(t).v(t)] \qquad \text{.....(4)}$$

From Fig. 1(*c*) (i) and Fig. 1(*b*) (ii) respectively,

$$E\cos\omega_q t = Z.i(t) + \frac{1}{C(t)}\int i(t).\mathrm{d}t \qquad \text{.....(5)}$$

$$I\cos\omega_q t = Y.v(t) + \frac{1}{L(t)}\int v(t).\mathrm{d}t \qquad \text{.....(6)}$$

These equations are true for every instant of time.

The practical applications of these circuits are mainly to rectifier modulators for time-varying resistance, and to parametric amplifiers for time-varying inductance and capacitance.

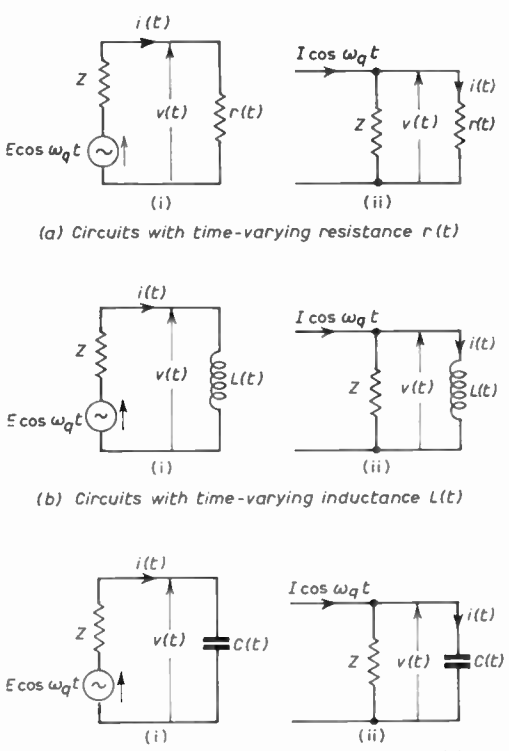


Fig. 1. Basic circuits.

† Electrical Engineering Department, University of Birmingham.

Now, in practice, the time-varying element will vary in a periodic manner, and we call the fundamental angular frequency of this variation ω_p. Although it is not necessary, it is convenient in order to simplify the mathematical working (and also reasonably representative of practice) to restrict the Fourier series representing the time variation to cosine terms.

Thus

$$r(t) = \sum_{n=0}^{\infty} r_n \cos n\omega_p t \qquad \text{.....(7)}$$

$$g(t) = \sum_{n=0}^{\infty} g_n \cos n\omega_p t \qquad \text{.....(8)}$$

$$L(t) = \sum_{n=0}^{\infty} L_n \cos n\omega_p t \qquad \text{.....(9)}$$

$$\frac{1}{L(t)} = \sum_{n=0}^{\infty} \left(\frac{1}{L}\right)_n \cos n\omega_p t \qquad \text{.....(10)}$$

$$C(t) = \sum_{n=0}^{\infty} C_n \cos n\omega_p t \qquad \text{.....(11)}$$

$$\frac{1}{C(t)} = \sum_{n=0}^{\infty} \left(\frac{1}{C}\right)_n \cos n\omega_p t \qquad \text{.....(12)}$$

Note that, except for square-wave variations of $r(t)$, the g_n are not explicitly related to the r_n; similarly for $(1/L)$ and L and for $(1/C)$ and C. Although the term "elastance" with symbol S is often used for $(1/C)$, there seems to be no corresponding term and symbol for $(1/L)$ in common use; we shall therefore continue to use the expressions $(1/C)$ and $(1/L)$.

Evidently the current $i(t)$ and the voltage $v(t)$ must, in general, contain components of all possible modulation frequencies $n\omega_p \pm \omega_q$. They may therefore be written

$$i(t) = \sum_{m=-\infty}^{\infty} i_m \cos[(\omega_q + m\omega_p)t + \theta_m] \qquad \text{.....(13)}$$

$$v(t) = \sum_{m=-\infty}^{\infty} v_m \cos[(\omega_q + m\omega_p)t + \phi_m] \qquad \text{.....(14)}$$

In these expressions the frequency has been written $\omega_q + m\omega_p$, although in practice ω_p often exceeds ω_q, so that for m negative this frequency is then negative. It is important, however, to express the frequency in this way, as it avoids the difficulty of reversal of phase angle which arises if the frequency is taken as $m\omega_p \pm \omega_q$ (see Ref. 1). If, in a practical problem, Z is specified at frequencies $m\omega_p - \omega_q$, then in the equations used here, the conjugate of the specified value must be used to give the appropriate value for frequencies $\omega_q - m\omega_p$.

It is possible, and some would say preferable, to use complex symbolism, i.e. $\exp(j\omega t)$, in place of the cosine notation.[2] The present author finds it hard to see any real advantage in this, however, and as students probably think more easily in terms of cosine waves it seems preferable to use the cosine notation. Moreover, the complex symbolism using a single exponential for each frequency relies on the superposition theorem for its justification, and clearly applies only to linear circuits. It is surely very wrong, when it is avoidable, to teach students to think in a system which is restricted to linear circuits when so many real problems involve non-linearity. Often, too, the working of a problem is greatly complicated by the use of exponentials.

One other point requires noting before proceeding to the expansion of eqns. (1)–(6). The term $Z.i(t)$ is handled by using at each frequency $(\omega_q + m\omega_p)$ the value Z_m which Z has at that frequency, thus

$$Z.i(t) = \sum_{-\infty}^{\infty} Z_m i_m \cos[(\omega_q + m\omega_p)t + \theta_m] \qquad \text{.....(15)}$$

and correspondingly for $Y.v(t)$.

3. Expansion for Time-varying Resistance

We shall deal with eqn. (1). Obviously eqn. (2) is handled in an exactly similar way.

The product $r(t).i(t)$ gives rise to a new series of terms, but without introducing any frequencies not already in $i(t)$. Thus

$$r_n \cos n\omega_p t \,.\, i_m \cos[(\omega_q + m\omega_p)t + \theta_m]$$
$$= \tfrac{1}{2} r_n i_m \cos\{[\omega_q + (m+n)\omega_p]t + \theta_m\} +$$
$$+ \tfrac{1}{2} r_n i_m \cos\{[\omega_q + (m-n)\omega_p]t + \theta_m\} \qquad \text{.....(16)}$$

It is important to note that the phase angle still carries the same subscript as the current magnitude, i.e. although the frequency has been changed, the phase angle and magnitude still correspond. This means that it is quite unnecessary to retain the phase angles if the currents i_m are regarded as vector quantities.

Since the eqn. (1) is true at all instants of time, it must hold for each individual frequency taken separately. In the statement of equilibrium for each frequency, the $\cos(\omega_q + m\omega_p)t$ will appear in every term and may be cancelled out. The expansion of eqn. (1) thus appears:

at frequency ω_q:

$$E = \ldots + \tfrac{1}{2} r_1 i_{-1} + (Z_0 + r_0) i_0 + \tfrac{1}{2} r_1 i_{+1} + \tfrac{1}{2} r_2 i_{+2} + \ldots \qquad \text{.....(17)}$$

at frequency $\omega_q + \omega_p$:

$$0 = \ldots + \tfrac{1}{2} r_2 i_{-1} + \tfrac{1}{2} r_1 i_0 + (Z_{+1} + r_0) i_{+1} + \tfrac{1}{2} r_1 i_{+2} + \ldots \qquad \text{.....(18)}$$

at frequency $\omega_q - \omega_p$:

$$0 = \ldots + (Z_{-1} + r_0) i_{-1} + \tfrac{1}{2} r_1 i_0 + \tfrac{1}{2} r_2 i_{+1} + \tfrac{1}{2} r_3 i_{+2} + \ldots \qquad \text{.....(19)}$$

and at frequency $\omega_q + m\omega_p (m \neq 0)$:

$$0 = \ldots + \tfrac{1}{2} r_{|m|} i_0 + \tfrac{1}{2} r_{|m-1|} i_{+1} + \ldots$$
$$+ \tfrac{1}{2} r_1 i_{m-1} + (Z_m + r_0) i_m + \ldots \qquad \text{.....(20)}$$

We thus have an infinite number of equations, each with an infinite number of terms. Explicit solution is not generally possible. But as soon as we can make some restrictions, solution becomes possible.

A very simple example of such restrictions is to specify that Z is infinite except at the two frequencies ω_q and $(\omega_q+\omega_p)$. This might, in practice, be a rectifier modulator of the "series" type with filters for source and load terminations, although it happens to be an inefficient arrangement and not recommended for general use. It is chosen here only for its analytical simplicity. Then all currents are zero except i_0 and i_{+1}. Equations (17) and (19) then become

$$E = (Z_0+r_0)i_0+\tfrac{1}{2}r_1 i_{+1} \qquad \text{.....(21)}$$

$$0 = \tfrac{1}{2}r_1 i_0+(Z_{+1}+r_0)i_{+1} \qquad \text{.....(22)}$$

giving the solution

$$i_{+1} = \frac{\tfrac{1}{2}Er_1}{\tfrac{1}{4}r_1^2-(Z_0+r_0)(Z_{+1}+r_0)} \qquad \text{.....(23)}$$

It should be noted that although the two eqns. (17) and (19) are sufficient for solving for i_{+1} (and i_0), the other equations do not disappear, as some finite terms (including $Z_m i_m$) remain in each.

An example of a somewhat different kind is to specify that Z is a constant pure resistance (R) at all relevant frequencies, and that $r(t)$ is a square-wave. Then it can be shown by various methods[1,3] that no even-order products exist other than i_0—i.e. $i_m = 0$ when m is even and not zero. Equation (17) then becomes

$$E = (R+r_0)i_0+\tfrac{1}{2}\sum_{m=1,3,5,\cdots}^{\infty} r_m(i_{+m}+i_{-m}) \qquad \text{.....(24)}$$

and eqn. (20) becomes, for m odd,

$$0 = (R+r_0)i_m+\tfrac{1}{2}r_m i_0 \qquad \text{.....(25)}$$

These give immediately a solution for any particular current, e.g.

$$i_{+1} = \frac{-\tfrac{1}{2}Er_1}{(R+r_0)^2-\tfrac{1}{2}\sum_{m=1,3,5,\cdots}^{\infty} r_m^2} \qquad \text{.....(26)}$$

Now as $r(t)$ is a square-wave,

$$\sum_{m=1,3,5,\cdots}^{\infty} r_m^2 = \frac{\pi^2}{8}r_1^2 \qquad \text{.....(27)}$$

so that the solution is explicit.

It is clear that knowledge of the absence of even-order products is a great help in this problem, as it is indeed in many others.

4. Expansion for Time-varying Inductance or Capacitance

4.1. *Solution of Equations* (3) *and* (4)

Here we shall deal with eqn. (3). It is clear that the same working applies also to eqn. (4).

Now

$$\frac{\mathrm{d}}{\mathrm{d}t}[L(t).i(t)] = \frac{\mathrm{d}L(t)}{\mathrm{d}t}.i(t)+L(t).\frac{\mathrm{d}i(t)}{\mathrm{d}t} \qquad \text{.....(28)}$$

From eqn. (9),

$$\frac{\mathrm{d}L(t)}{\mathrm{d}t} = -\omega_p\sum_{n=1}^{\infty} nL_n \sin n\omega_p t \qquad \text{.....(29)}$$

and from eqn. (11),

$$\frac{\mathrm{d}i(t)}{\mathrm{d}t} = -\sum_{m=-\infty}^{\infty}(\omega_q+m\omega_p)i_m \sin[(\omega_q+m\omega_p)t+\theta_m] \qquad \text{.....(30)}$$

Thus eqn. (3) becomes

$$\begin{aligned} E\cos\omega_q t = {} & \sum_{m=-\infty}^{\infty} Z_m i_m \cos[(\omega_q+m\omega_p)t+\theta_m]- \\ & -\omega_p\sum_{n=1}^{\infty} nL_n \sin n\omega_p t\times \\ & \times\sum_{m=-\infty}^{\infty} i_m \cos[(\omega_q+m\omega_p)t+\theta_m]- \\ & -\sum_{n=0}^{\infty} L_n \cos n\omega_p t\times \\ & \times\sum_{m=-\infty}^{\infty}(\omega_q+m\omega_p)i_m \sin[(\omega_q+m\omega_p)t+\theta_m] \end{aligned} \qquad \text{.....(31)}$$

It will be observed that in all resultant terms, θ_m is always associated with i_m as in Section 3, but there are now sine as well as cosine terms. This can be taken care of by the use of the "j" operator with the sine terms, and a set of equations, in which only the vector coefficients appear, can then be formed, on the lines of, but more complicated than, eqns. (17)–(20).

If eqn. (31) is expanded, and the terms at each separate frequency are grouped together, it is readily seen that the following system of equations is obtained:

at frequency ω_q:

$$E = \ldots+\tfrac{1}{2}\mathrm{j}\omega_q L_1 i_{-1}+(Z_0+\mathrm{j}\omega_q L_0)i_0+\tfrac{1}{2}\mathrm{j}\omega_q L_1 i_{+1}+\tfrac{1}{2}\mathrm{j}\omega_q L_2 i_{+2}+\ldots \qquad \text{.....(32)}$$

at frequency $\omega_q+\omega_p$:

$$0 = \ldots+\tfrac{1}{2}\mathrm{j}(\omega_q+\omega_p)L_2 i_{-1}+\tfrac{1}{2}\mathrm{j}(\omega_q+\omega_p)L_1 i_0+[Z_{+1}+\mathrm{j}(\omega_q+\omega_p)L_0]i_{+1}+\tfrac{1}{2}\mathrm{j}(\omega_q+\omega_p)L_1 i_{+2}+\ldots \qquad \text{.....(33)}$$

and at frequency $\omega_q+m\omega_p$ $(m \neq 0)$:

$$0 = \ldots+\tfrac{1}{2}\mathrm{j}(\omega_q+m\omega_p)L_{|m|}i_0+\tfrac{1}{2}\mathrm{j}(\omega_q+m\omega_p)L_{|m-1|}i_{+1}+\ldots+\tfrac{1}{2}\mathrm{j}(\omega_q+m\omega_p)L_1 i_{m-1}+[Z_m+\mathrm{j}(\omega_q+m\omega_p)L_0]i_m+ \qquad \text{.....(34)}$$

As with the time-varying resistance, no explicit solution is possible for the general case. Particular cases in which specified restrictions are imposed are, however, soluble. A most important such case, which forms the basis of the parametric amplifier, is that in which Z, in Fig. 1(*b*) (i), is infinite at all frequencies except ω_q and $(\omega_q+\omega_p)$—(or $(\omega_q-\omega_p)$ as will be discussed later). Then i_0 and i_{+1} are the only currents which are not zero. Using eqns. (32) and (33) we obtain

$$E = (Z_0+\mathrm{j}\omega_q L_0)i_0+\tfrac{1}{2}\mathrm{j}\omega_q L_1 i_{+1} \quad \text{.....(35)}$$

and

$$0 = \tfrac{1}{2}\mathrm{j}(\omega_q+\omega_p)L_1 i_0+[Z_{+1}+\mathrm{j}(\omega_q+\omega_p)L_0]i_{+1} \quad \text{.....(36)}$$

whence

$$i_{+1} = \frac{-\tfrac{1}{2}\mathrm{j}(\omega_q+\omega_p)L_1 E}{\tfrac{1}{4}\omega_q(\omega_q+\omega_p)L_1^2+(Z_0+\mathrm{j}\omega_q L_0)[Z_{+1}+\mathrm{j}(\omega_q+\omega_p)L_0]} \quad \text{.....(37)}$$

4.2. *Solution of Equations* (5) *and* (6)

Here, in order to retain the working for a time-varying inductance, we shall consider eqn. (6). Clearly eqn. (5) is worked out exactly similarly. It is also very important to work as often as possible in terms of admittance rather than impedance, as students seem to find this difficult, and eqn. (6) gives practice in this.

Using the expressions given in eqns. (10) and (14), eqn. (6) may be written

$$I\cos\omega_q t = \sum_{m=-\infty}^{\infty} Y_m v_m \cos[(\omega_q+m\omega_p)t+\phi_m] + \sum_{n=0}^{\infty}\left(\frac{1}{L}\right)_n \cos n\omega_p t \times \times \sum_{m=-\infty}^{\infty}\frac{v_m}{\omega_q+m\omega_p}\sin[(\omega_q+m\omega_p)t+\phi_m] \quad \text{.....(38)}$$

Expanding this into separate equations for each frequency, and regarding the coefficients v_m as vectors, we obtain:

at frequency ω_q:

$$I = \ldots+\frac{1}{2}\cdot\frac{1}{\mathrm{j}(\omega_q-\omega_p)}\left(\frac{1}{L}\right)_1 v_{-1}+\left[Y_0+\frac{1}{\mathrm{j}\omega_q}\left(\frac{1}{L}\right)_0\right]v_0+\frac{1}{2}\cdot\frac{1}{\mathrm{j}(\omega_q+\omega_p)}\left(\frac{1}{L}\right)_1 v_{+1}+\ldots \quad \text{.....(39)}$$

and at frequency $\omega_q+m\omega_p$ $(m\neq 0)$:

$$0 = \ldots+\frac{1}{2}\cdot\frac{1}{\mathrm{j}\omega_q}\left(\frac{1}{L}\right)_{|m|}v_0+\frac{1}{2}\cdot\frac{1}{\mathrm{j}(\omega_q+\omega_p)}\left(\frac{1}{L}\right)_{|m-1|}v_{+1}+\ldots+\frac{1}{2}\cdot\frac{1}{\mathrm{j}[\omega_q+(m-1)\omega_p]}\left(\frac{1}{L}\right)_1 v_{m-1}+ +\left[Y_m+\frac{1}{\mathrm{j}(\omega_q+m\omega_p)}\left(\frac{1}{L}\right)_0\right]v_m+\ldots \quad \text{.....(40)}$$

It can be seen at once that this is quite different from eqns. (32)–(34); in particular, *different* frequencies occur in the coefficients of the various terms of one equation, instead of just the one frequency. Nevertheless, this set of equations is used in just the same way as before.

A particular case, which can also be used as the basis of a parametric amplifier, is that in which Y is infinite at all frequencies except ω_q and $(\omega_q+\omega_p)$, so that only v_0 and v_{+1} are finite, and only two equations are required to solve for v_{+1}.

This case gives

$$I = \left[Y_0+\frac{1}{\mathrm{j}\omega_q}\left(\frac{1}{L}\right)_0\right]v_0+\frac{1}{2}\cdot\frac{1}{\mathrm{j}(\omega_q+\omega_p)}\left(\frac{1}{L}\right)_1 v_{+1} \quad \text{.....(41)}$$

and

$$0 = \frac{1}{2}\cdot\frac{1}{\mathrm{j}\omega_q}\left(\frac{1}{L}\right)_1 v_0+\left[Y_{+1}+\frac{1}{\mathrm{j}(\omega_q+\omega_p)}\left(\frac{1}{L}\right)_0\right]v_{+1} \quad \text{.....(42)}$$

whence

$$v_{+1} = \frac{-\frac{1}{2}\cdot\frac{1}{\mathrm{j}\omega_q}\left(\frac{1}{L}\right)_1 I}{\frac{1}{4}\cdot\frac{1}{\omega_q(\omega_q+\omega_p)}\left(\frac{1}{L}\right)_1^2+\left[Y_0+\frac{1}{\mathrm{j}\omega_q}\left(\frac{1}{L}\right)_0\right]\left[Y_{+1}+\frac{1}{\mathrm{j}(\omega_q+\omega_p)}\left(\frac{1}{L}\right)_0\right]} \quad \text{.....(43)}$$

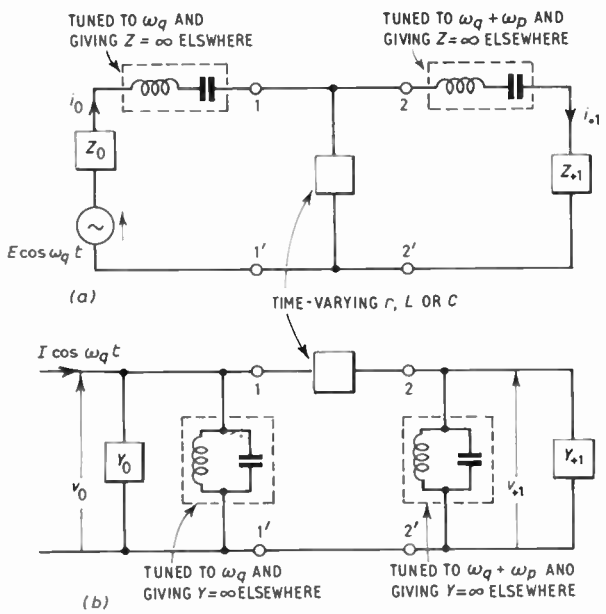


Fig. 2. Time-varying element in a circuit with
(*a*) only two non-zero currents
(*b*) only two non-zero voltages.

5. Further Consideration of Circuits Restricted to Two Non-zero Currents or Voltages

The examples of specific solutions which we have taken above have restricted the circuit to have either (a) non-zero currents at only two frequencies, or (b) non-zero voltages at only two frequencies. It is thus convenient to consider the circuit in its practical form as having separate parts for the input and output signals respectively, as shown in Fig. 2(*a*) and (*b*). It is clear that from the point of view of the previous analysis, since the two parts function at different frequencies, they may be superposed to give the single loops or single node-pairs of Fig. 1. Then Z_0 and Z_{+1} are merely the values of Z at frequencies ω_q and $(\omega_q+\omega_p)$; and similarly for Y. The tuned circuits shown are symbolic only, to indicate a practical way of obtaining an approximation to the required conditions.

We shall now show how the matching conditions and conversion loss or gain may be determined, and draw some general conclusions regarding these circuits.

5.1. *Time-varying Resistance*

Let us assume first of all that $Z_0 = R_0+\mathrm{j}X_0$ is given, and that we require to find the value of $Z_{+1} = R_{+1}+\mathrm{j}X_{+1}$ needed to give a conjugate match (i.e. maximum power transfer) at the terminals 2,2′ in Fig. 2(*a*). The circuit to the left of 2,2′ may be represented as a generator of e.m.f. E_{+1} and internal impedance $Z_{i,+1}$ at the frequency $\omega_q+\omega_p$. Obviously then, if we find $Z_{i,+1}$, we have to make $Z_{+1} = Z^*_{i,+1}$.

Now eqn. (21) gives us

$$i_0 = \frac{E-\frac{1}{2}r_1 i_{+1}}{Z_0+r_0} \qquad \text{.....(44)}$$

and since $E_{+1}-Z_{i,+1}i_{+1} = Z_{+1}i_{+1}$, then, from eqn. (22),

$$E_{+1}-Z_{i,+1}i_{+1} = -\tfrac{1}{2}r_1 i_0 - r_0 i_{+1} = \frac{-\frac{1}{2}r_1 E}{Z_0+r_0} + \left[\frac{\frac{1}{4}r_1^2}{Z_0+r_0} - r_0\right] i_{+1} \qquad \text{.....(45)}$$

so that

$$E_{+1} = \frac{-\frac{1}{2}r_1 E}{Z_0+r_0} \qquad \text{.....(46)}$$

and

$$Z_{i,+1} = r_0 - \frac{\frac{1}{4}r_1^2}{Z_0+r_0} \qquad \text{.....(47)}$$

$$= r_0 - \frac{\frac{1}{4}r_1^2(R_0+r_0)}{(R_0+r_0)^2+X_0^2} + \mathrm{j}\cdot\frac{\frac{1}{4}r_1^2 X_0}{(R_0+r_0)^2+X_0^2} \qquad \text{.....(48)}$$

Therefore for a conjugate match at 2,2′,

$$R_{+1} = r_0 - \frac{\frac{1}{4}r_1^2(R_0+r_0)}{(R_0+r_0)^2+X_0^2} \qquad \text{.....(49)}$$

and

$$X_{+1} = -\frac{\frac{1}{4}r_1^2 X_0}{(R_0+r_0)^2+X_0^2} \qquad \text{.....(50)}$$

When this match is obtained, the power in the output load

$$= |E_{+1}|^2/8R_{+1} = \frac{\frac{1}{4}r_1^2E^2}{(R_0+r_0)^2+X_0^2}\Bigg/\frac{8\{r_0[(R_0+r_0)^2+X_0^2]-\frac{1}{4}r_1^2(R_0+r_0)\}}{(R_0+r_0)^2+X_0^2} \qquad \text{.....(51)}$$

and for the given Z_0 and $r(t)$, the optimum conversion power-loss ratio

$$= \frac{\text{power available from signal source at } \omega_q}{\text{power in load at } \omega_q+\omega_p} = \frac{r_0[(R_0+r_0)^2+X_0^2]-\frac{1}{4}r_1^2(R_0+r_0)}{\frac{1}{4}r_1^2R_0} \qquad \text{.....(52)}$$

If the load Z_{+1} were given, and we had to find the value of Z_0 required for maximum power transfer, then we would calculate the input impedance at frequency ω_q looking from the left into terminals 1,1′ as

$$Z_{i0} = \frac{E-Z_0 i_0}{i_0} = r_0 - \frac{\frac{1}{4}r_1^2}{Z_{+1}+r_0} \qquad \text{.....(53)}$$

We would then choose Z_0 to be the conjugate of Z_{i0}.

Equations (47) and (53) make it clear that the circuit is symmetrical, so that if Z_0 and Z_{+1} are both adjustable, the conversion loss will clearly be a minimum when $X_0 = X_{+1} = 0$ and

$$R_0 = R_{+1} = \sqrt{r_0^2 - \tfrac{1}{4}r_1^2}$$

The minimum conversion power-loss ratio is then

$$[r_0 + \sqrt{r_0^2 - \tfrac{1}{4}r_1^2}]^2 / \tfrac{1}{4}r_1^2 \qquad \text{.....(54)}$$

It happens that the value of optimum circuit resistance thus obtained is usually inconveniently high; if $r(t)$ is a square-wave variation between values r_f (e.g. forward resistance of a rectifier) and r_b (e.g. backward resistance of a rectifier), then $R_0 = R_{+1} \simeq 0{\cdot}39\ r_b$ if it is assumed that $r_b \gg r_f$. The conversion loss may then be as low as 8·9 dB, but if the circuit resistance has to be reduced to some small fraction of r_b for practical reasons, the loss becomes high. If the time-varying resistance were, in fact, the opening and closing of a perfect switch, then it could be shown that the loss becomes infinite.

The method can, of course, be applied (but with more complexity) to other frequency-changer arrangements where more than two currents or voltages exist, and where lower losses may be obtained. For a tabulation of such arrangements and their minimum losses, see reference 1.

The minimum conversion power-loss ratio it is possible to have in any circuit with time-varying resistance is unity—i.e. no loss at all—and this occurs in a single-loop circuit when the following conditions are met:

(a) $r(t)$ is a square-wave function switching between values of zero and infinity,

(b) $Z = R_0$ at frequency ω_q, R_{+1} at frequency $(\omega_q + \omega_p)$, zero at all other odd-order product frequencies, and infinity at all other even-order product frequencies,

(c) $R_0 = (4/\pi^2)R_{+1}$.

There is, of course, a corresponding dual circuit.

This zero-loss condition is hardly a practical one, as the impedance requirements are almost impracticable; nor is it easily deducible from the general theory given here. But it is of basic theoretical importance in showing that it is, in principle, possible to convert all available signal power to another frequency; and it can be readily realized in practice with a ring modulator—i.e. with a 3-loop circuit. The matter is fully discussed by Belevitch.[4]

5.2. *Time-varying Inductance or Capacitance*

5.2.1. Use as a frequency-converter

The same method is used as that discussed in the previous section.

Working with the circuit of Fig. 2(*a*), and using time-varying inductance, we have, from eqns. (35) and (36), the following values for the effective generator at frequency $(\omega_q + \omega_p)$ seen to the left of terminals 2,2′:

$$E_{+1} = \frac{-\tfrac{1}{2}\mathrm{j}(\omega_q + \omega_p)L_1 E}{Z_0 + \mathrm{j}\omega_q L_0} \qquad \text{.....(55)}$$

$$Z_{i,+1} = \mathrm{j}(\omega_q + \omega_p)L_0 + \frac{\tfrac{1}{4}\omega_q(\omega_q + \omega_p)L_1^2}{Z_0 + \mathrm{j}\omega_q L_0} \qquad \text{.....(56)}$$

For a given Z_0, therefore, the value of

$$Z_{+1} = R_{+1} + \mathrm{j}X_{+1}$$

to give a conjugate match at 2,2′ (and hence maximum power transfer) is $Z_{i,+1}^*$ so that

$$R_{+1} = \frac{\tfrac{1}{4}\omega_q(\omega_q + \omega_p)L_1^2 R_0}{R_0^2 + (X_0 + \omega_q L_0)^2} \qquad \text{.....(57)}$$

and

$$X_{+1} = -(\omega_a + \omega_b)\left\{L_0 - \frac{\tfrac{1}{4}\omega_q L_1^2 (X_0 + \omega_q L_0)}{R_0^2 + (X_0 + \omega_q L_0)^2}\right\} \qquad \text{.....(58)}$$

With this match, the power in the load

$$= |E_{+1}|^2 / 8R_{+1}$$

$$= \frac{\omega_q + \omega_p}{\omega_q} \frac{E^2}{8R_0} \qquad \text{.....(59)}$$

so that the power *gain*

$$= \frac{\omega_q + \omega_p}{\omega_q} \qquad \text{.....(60)}$$

Evidently, therefore, if $\omega_p \gg \omega_q$, this gain is very large. This is the basis of the parametric amplifier of the so-called upper-sideband "up-converter" type. It is an amplifier so long as the wanted output signal is of higher frequency than the input signal. The power gain is derived, of course, from the work done in varying L—commonly called the "pumping" action.

The gain given by eqn. (60) is obtained on making Z_{+1} a matched value for any given Z_0. It is interesting that no optimization of Z_0 and Z_{+1} in relation to $L(t)$ is required. In other words, if Z_0 is given, then a choice of Z_{+1} to give a conjugate match according to eqns. (57) and (58) will automatically give the maximum possible power gain of $(\omega_q + \omega_p)/\omega_q$. This is quite different from the behaviour of the circuit with time-varying resistance discussed earlier, since in that circuit the minimum conversion loss was obtained only when Z_0 and Z_{+1} were both chosen to have a particular, unique pair of values in relation to the Fourier coefficients of $r(t)$; mere conjugate matching with an arbitrary choice of Z_0 or Z_{+1} gave, in general, greater loss.

For two extreme cases, interesting results are obtained:

(a) $X_0 = -\omega_q L_0$ and $X_{+1} = -(\omega_q+\omega_p)L_0$

(This is the condition usually assumed in the literature.)

Then $$R_0 R_{+1} = \tfrac{1}{4}\omega_q(\omega_q+\omega_p)L_1^2 \qquad \text{.....(61)}$$

is the condition for maximum gain.

(b) $X_0 = 0$. Then for maximum gain,

$$\frac{R_{+1}}{R_0} = \frac{\omega_q+\omega_p}{\omega_q} \cdot \frac{\tfrac{1}{4}\omega_q^2 L_1^2}{R_0^2+\omega_q^2 L_0^2} \qquad \text{.....(62)}$$

and $$\frac{X_{+1}}{\omega_q L_0} = \frac{\omega_q+\omega_p}{\omega_q}\left[1 - \frac{\tfrac{1}{4}\omega_q^2 L_1^2}{R_0^2+\omega_q^2 L_0^2}\right] \qquad \text{.....(63)}$$

Therefore $$\frac{R_{+1}}{R_0} + \frac{X_{+1}}{\omega_q L_0} = \frac{\omega_q+\omega_p}{\omega_q} \qquad \text{.....(64)}$$

Taking case (a), as it represents a usual practical arrangement, we can examine the effect of changing the signal frequency by a small amount. Since in this case $X_0+\omega_q L_0 \simeq 0$ for small changes in ω_q, we can write eqn. (58) as

$$X_{+1} \simeq -(\omega_q+\omega_p)\left\{L_0 - \frac{1}{4R_0^2}\omega_q L_1^2(X_0+\omega_q L_0)\right\} \qquad \text{.....(65)}$$

Since X_0 is due to a capacitance (or, at any rate, its circuit is dominantly a capacitance), a small increase in ω_q will cause the expression between the large brackets to diminish in magnitude, so that the value of X_{+1} required for conjugate matching becomes less negative. But since X_{+1} is also necessarily due to a capacitance, this is just the way the value of X_{+1} would tend to alter due to the change of frequency. Thus an approximation to conjugate matching is obtained over a relatively wide frequency band. This means, therefore, that the up-converter is inherently a wide-band device.

Manley and Rowe[5] have shown that some general power relations exist in circuits with time-varying capacitance, and one of their results is that in this particular circuit the ratio of power absorbed in the load at frequency $(\omega_q+\omega_p)$ is always $(\omega_q+\omega_p)/\omega_q$ times the power *absorbed* from the signal source at frequency ω_q, irrespective of matching. The power gain, however, relates the load power to the *available* signal power, not to the signal power actually absorbed, and so to obtain the gain given by eqn. (60), matching is required.

If the same process (but using dual parameters) is applied to the circuit of Fig. 2(*b*) with time-varying inductance, working with eqns. (41) and (42), and putting $Y_0 = G_0+jB_0$, then we find that for a conjugate match at 2,2′, we require $Y_{+1} = G_{+1}+jB_{+1}$ where

$$G_{+1} = \frac{\tfrac{1}{4}\left(\frac{1}{L}\right)_1^2 G_0}{\omega_q(\omega_q+\omega_p)\left\{G_0^2+\left[B_0+\frac{1}{\omega_q}\left(\frac{1}{L}\right)_0\right]^2\right\}} \qquad \text{.....(66)}$$

and

$$B_{+1} = \frac{1}{\omega_q+\omega_p}\left\{\left(\frac{1}{L}\right)_0 - \frac{\tfrac{1}{4}\left(\frac{1}{L}\right)_1^2\left[B_0+\frac{1}{\omega_q}\left(\frac{1}{L}\right)_0\right]}{\omega_q\left\{G_0^2+\left[B_0+\frac{1}{\omega_q}\left(\frac{1}{L}\right)_0\right]^2\right\}}\right\} \qquad \text{.....(67)}$$

This match gives a power gain of $(\omega_q+\omega_p)/\omega_q$ as for the previous circuit.

It will be found, indeed, that all the four circuits of Fig. 1(*b*) and (*c*), give the same performance, and expressions for matching, when proper allowance is made for the dual relationships. This means that an up-converter (and indeed, the negative-resistance parametric amplifier discussed below) can be made with time-varying inductance *or* capacitance, and with the circuit impedance made infinite *or* zero at the unwanted frequencies.

It is extremely interesting that the maximum power gain, as given by eqn. (60), is independent of the amount of inductance variation, as represented by L_1 or $(1/L)_1$. One would at first have supposed that as the variation was reduced, so also would the power delivered into the output circuit be reduced. But, taking the circuit of Fig. 2(*a*), the optimum resistance termination (R_{+1}) is seen from eqn. (57) to be a function of L_1, so that as L_1 is reduced, so also is R_{+1}. Thus, as we approach the limit, variations in the inductance (although small) are opposed (according to Lenz's law) by increased forces due to the almost short-circuited condition; this makes it possible for the power absorbed in producing the variations to remain constant in spite of the reduced variation. There is, of course, a discontinuity at the limit where L_1 actually becomes zero, since no gain can then be produced.

It can be shown that if the inductance is associated with a resistance representing its losses—as in practice it must be—the gain reduces continuously to zero as L_1 is reduced.

5.2.2. Negative-resistance parametric amplifier

It is evident that if the output frequency were $\omega_q-\omega_p$ instead of $\omega_q+\omega_p$, there would be less gain, and there could indeed be a power loss instead of a gain. But this case can be exploited, nevertheless, to give another kind of parametric amplifier. Suppose then that the system has Z infinite except at frequencies ω_q and $\omega_q-\omega_p$, according to Fig. 2(*a*), with time-varying inductance, and that we calculate the input

impedance (Z_{i0}) seen from the signal source terminals 1,1′. Equations (35) and (36), adapted for frequency $\omega_q-\omega_p$, give

$$Z_{i0}=\frac{E-i_0Z_0}{i_0}=\mathrm{j}\omega_qL_0+\frac{\frac{1}{4}\omega_q(\omega_q-\omega_p)L_1^2}{R_{-1}+\mathrm{j}[X_{-1}+(\omega_q-\omega_p)L_0]}$$
$$=\frac{\frac{1}{4}\omega_q(\omega_q-\omega_p)L_1^2R_{-1}}{R_{-1}^2+[X_{-1}+(\omega_q-\omega_p)L_0]^2}+$$
$$+\mathrm{j}\left\{\omega_qL_0-\frac{\frac{1}{4}\omega_q(\omega_q-\omega_p)L_1^2[X_{-1}+(\omega_q-\omega_p)L_0]}{R_{-1}^2+[X_{-1}+(\omega_q-\omega_p)L_0]^2}\right\} \quad \ldots\ldots(68)$$

It is thus immediately clear that if $\omega_p \gg \omega_q$, then the real part of this input impedance is a *negative* resistance:

$$R_{i0}=-\frac{\frac{1}{4}\omega_q(\omega_p-\omega_q)L_1^2R_{-1}}{R_{-1}^2+[X_{-1}-(\omega_p-\omega_q)L_0]^2} \quad \ldots\ldots(69)$$

If then a load circuit operating at frequency ω_q is connected across terminals 1,1′, its resistive component may be arranged to be very nearly cancelled out by the negative resistance R_{i0}. A large output may then be maintained by applying only a small power from the signal source, and so a power amplifier has been obtained. The power gain can clearly be made as high as desired; if it is made infinite by making R_{i0} completely cancel out the load resistance, then clearly a self-oscillator is obtained.

It is evident that the negative resistance amplifier, by removing most of the resistance component of any tuned circuit in the signal path, is inherently a narrow-bandwidth device; in this respect it contrasts markedly with the up-converter, which is inherently a wide-band device, as previously explained.

Exactly corresponding results are obtained with the circuit of Fig. 2(*b*), and with a time-varying capacitance. For high-frequency use, the capacitance. amplifier is the more suitable in practice.[6,7]

An account of the history of parametric amplifiers, with a very large bibliography, is given by Mumford.[8]

6. More Complicated Circuit Configurations

In the case of the time-varying resistance, which represents the important practical circuits known as rectifier modulators, circuits comprising more than one loop or one node-pair are common[9,10]—e.g. the ring (or lattice) modulator. However, with due care, all the usual circuit theorems and transformations can be applied to them,[11] and in some circumstances very simple equivalent circuits can be found. For instance, with the assumption of a local carrier (or switching) oscillation having "half-cycle" symmetry—i.e. containing no even-order harmonics—the ring modulator can be shown to be equivalent to a single-loop or single-node-pair modulator.[1] Conditions for modulators to have an input impedance which is not time-varying can be specified[10] on the analogy of Zobel's constant-resistance networks. It is always important to remember, however, when making circuit transformations, that products such as $Z.r(t)$, which frequently arise, are not straight products; since Z can be regarded as a function of $(\mathrm{d}/\mathrm{d}t)$, it is an operator, operating on $r(t)$. This frequently makes it difficult to make any use of the transformation.

7. Non-linear Effects in Circuits with Time-varying Parameters

The treatment so far given of time-varying parameters has merely assumed that the variation $r(t)$, $L(t)$ or $C(t)$ has been produced by some unspecified external agency. It is necessary, therefore, to give students some idea of how the variation is produced in practice, this being generally by the use of large-amplitude local signals which produce a time-varying bias in a non-linear element of r, L or C. The question will then inevitably be asked as to whether the information-signal voltages or currents, which also appear across or in the non-linear element, have any effect on the time-variation of the element, and whether they are subject to non-linear distortion due to the non-linearity of the element. The answer must be given that both effects do occur, and give rise to non-linear distortion.

A general treatment of this matter is prohibitively complicated (even if possible) for undergraduate courses. But with some special simplifying assumptions, an insight into the nature of the effects may be given in terms which are acceptable to undergraduates who have had a course in communication systems, and some previous contact with non-linearity.[12] One group of assumptions is as follows:

(a) the circuits are purely-resistive with time-varying resistance

(b) signal voltage or current small compared with bias wave

(c) nominally square-wave variation of $r(t)$ produced either

(i) by cosine bias wave applied to bilinear rectifiers (i.e. rectifiers with a constant forward resistance and a constant back resistance, switching at zero applied voltage).

or (ii) by square-wave bias applied to any kind of non-linear resistance.

Assumption (c) (i) permits the development of the idea of a square-wave time-variation, phase-modulated by the difference-frequency between signal and bias wave, so that either $r(t)$, or the corresponding time-varying transfer function $\phi(t)$, may be written approximately as

$$h_0+h_1\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{2n-1}\cos[(2n-1)\{\omega_p t-x_1\sin(\omega_p-\omega_{q1})t-x_2\sin(\omega_p-\omega_{q2})t-\ldots\}] \quad \ldots\ldots(70)$$

where ω_{q1}, ω_{q2}, etc. are various frequency-components of the information-signal of amplitudes x_1, x_2, etc. relative to the bias wave. The output signal consists basically of the product of this function and the input signal

$$e_1\cos\omega_{q1}t+e_2\cos\omega_{q2}t+\ldots \quad \ldots\ldots(71)$$

and it is clear that an infinite range of harmonics and intermodulation products is produced, with amplitudes dependent on both the relative signal amplitudes (x) and the order of modulation involved ($2n-1$). A full account of this method is available elsewhere.[13]

Assumption (c) (ii) enables the lattice time-varying network (e.g. the ring modulator) to be regarded[13] as a non-linear but constant lattice followed by a reversing switch operating at frequency $\omega_p/2\pi$. The non-linear lattice can have a transfer function represented by a power-series, and the non-linear distortion products produced by this are easily calculated. They are then all multiplied by the switching function

$$\phi(t)=h_1\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{2n-1}\cos(2n-1)\omega_p t \quad \ldots\ldots(72)$$

to give the output spectrum.

It is clear these two processes (i) and (ii) give different kinds of result, and give a good indication of the complexity of non-linear analysis without introducing any processes beyond undergraduate level.

8. References

1. D. P. Howson and D. G. Tucker, "Rectifier modulators with frequency-selective terminations", *Proc. Instn Elect. Engrs*, **107B**, p. 269, May 1960.
2. A. P. Bolle, "Application of complex symbolism to linear variable networks", *Trans. Inst. Radio Engrs* (*Circuit Theory*), **CT-2**, No. 1, p. 32, March 1955.
3. D. G. Tucker, "Elimination of even-order modulation in rectifier modulators", *J. Brit.I.R.E.*, **21**, p. 161, 1961.
4. V. Belevitch, "Théorie des Circuits Non-linéaires en Regime Alternatif", (Librairie Universitaire, Louvain, 1959).
5. J. M. Manley and H. E. Rowe, "Some general properties of non-linear elements", *Proc. Inst. Radio Engrs*, **44**, p. 904, 1956.
6. G. D. Sims and I. M. Stephenson, "Parametric amplifiers", *Discovery*, December 1960, p. 528.
7. L. A. Blackwell and K. L. Kotzebue, "Semiconductor-diode Parametric Amplifiers", (Prentice-Hall, Englewood Cliffs, N.J., 1961).
8. W. W. Mumford, "Some notes on the history of parametric transducers", *Proc. Inst. Radio Engrs*, **48**, p. 848, 1960.
9. D. G. Tucker, "Modulators and Frequency-changers", (Macdonald, London, 1953).
10. D. G. Tucker, "Constant-resistance modulators", *J. Brit. I.R.E.*, **21**, p. 161, 1961.
11. D. P. Howson, "Some Applications of Network Theorems to Linear Circuits with Time-varying Resistance", Electrical Engineering Dept., University of Birmingham, Memorandum No. 40, 1959.
12. D. G. Tucker, "Non-linear circuits: a course for undergraduates", *Bull. Elect. Engng Educ.*, **26**, p. 62, June 1961.
13. D. G. Tucker, "Intermodulation distortion in rectifier modulators", *Wireless Engineer*, **31**, p. 145, 1954.

Manuscript first received by the Institution on 2nd March 1962 and in final form on 14th December 1962 (Paper No. 800/Ed9).

Commonwealth Broadcasting

Broadcasting in Rhodesia and Nyasaland

The Annual Report of the Federal Broadcasting Corporation for Rhodesia and Nyasaland for 1961–2 describes progress which has been made in the engineering services of the Corporation. The main engineering projects were the completion of the third television transmitting station at Kitwe, improvements at some studio centres, planning of new transmitter installations and better standby facilities at the more important stations.

To give better General Service daytime coverage of the more distant parts of the Federation, particularly Nyasaland, a 2½ kW short wave transmitter was brought into use at Gwelo in the 31 metre band, with marked success. The same transmitter was effectively used in the 120 metre band to improve coverage during the difficult winter months.

From the middle of May, Kitwe was the scene of major activity in preparing for the installation of the television transmitter to serve the Copperbelt. The work at site included additions to the existing building, the erection of a temporary medium wave aerial, the dismantling of an existing 310 ft. mast radiator and its replacement by a 500 ft. stayed lattice mast designed both to support the television aerial and to act as a medium wave radiator.

The television transmitter was put into service on 15th December 1961, twelve months ahead of the date originally planned. It operates in C.C.I.R. channel E4 of Band I, and by means of a directional aerial system provides satisfactory coverage of all Copperbelt townships. There are already television transmitters within the Federation at Salisbury and Bulawayo.

Television for Singapore

The Broadcasting Division of the Singapore Government's Ministry of Culture ("Radio Singapore") have decided to provide a comprehensive television service for the island, which is to be supplied and installed by the Marconi Company. The completed station will carry programmes in four languages, Malay, Mandarin, Tamil and English.

Initially there will be one programme channel, radiated via a pair of vision and sound transmitters (5 kW vision, 1 kW sound), with a second pair acting as standby. At this stage, one studio and an announcer's booth, together with control rooms, come into service, the cameras employed for the main studio being two Marconi Mark IV image orthicons, whilst for telecine work, two Mark IV vidicon camera channels will be used.

The second phase calls for a considerable extension of facilities including the provision of a much more comprehensive system with four studios, two announcer's booths, associated control rooms and ancillary equipment. As at present planned, two of the four main studios will each be equipped with three Mark IV 4½-in. image orthicon camera channels, and two with one Mark IV 4½-in. image orthicon each. There will be a control room for every studio. Other facilities will include the provision of a second announcer's booth, an extensive master control room, a second continuity supervisor's position and a third Mark IV vidicon telecine channel.

Television survey and planning teams frequently have to overcome major topographical problems, but Singapore appears to be a gratifying exception, as it is almost ideally placed for a television service. The transmitters are sited at Bukit Batok, situated in the middle of the area, with the surrounding terrain flat (except for Bukit Tima, another hill to the west of Bukit Batok) and carrying a high concentration of population. The studios are at Caldicott Hill, alongside the existing Radio Singapore sound studios.

The aerial systems consist of two 4-stack quadrant aerials mounted on a common 350 ft tower. A microwave link carries the signals from the studios to the transmitting site and v.h.f./f.m. links are provided for supervisory circuits between various units of the broadcasting organization.

A pilot programme service came into operation on 15th February. The station operates in Band III to 625-line C.C.I.R. standards, with horizontal polarization of the signals.

Sound System for Parliament Buildings in Malaya

A complete sound amplification system is to be installed in the new Parliament buildings in Kuala Lumpur for the Federation of Malaya by Trix Electronics Limited, of London. The system includes more than 200 microphones, together with low-level speakers. The necessary amplifier and control systems cover the installations in the two Houses, as well as an extensive distribution network in the adjoining offices and other rooms. In addition, a 4-channel radiated translation system is to be installed, providing language channels in the two Houses.

Jamaican Television Expansion

The newly-built Studio Centre of the Jamaican Broadcasting Corporation at Kingston will comprise two new studios equipped with four 4½-in image-orthicon camera systems with all ancillary equipment. This equipment, as well as a ten-channel sound mixer and a five-channel sound mixer is being supplied by E.M.I. Electronics.

U.D.C. 621.867

Driverless Tractors for Materials Handling

By

F. G. HELPS, B.Sc.†

Presented at the Symposium on "Recent Developments in Industrial Electronics" in London on 2nd-4th April 1962.

Summary: Tractors have been developed which are capable of carrying goods or towing trailers from one location to another without employing a driver. A buried conductor carrying a low frequency current provides the means of guidance. The tractors are able to make the right decisions at points where alternative routes may be followed in order to arrive at their correct destinations. The problems of designing a system and the economic advantages in employing such a system are discussed.

1. Introduction

Both in this country and in America work has been carried out on guidance systems for use with road vehicles, tractors, and even mowing machines. One such system, known as the Robotug system, will be described.

This principle has so far been applied to materials handling in which battery-powered electric tractors either carry the materials themselves or tow a train of trailers. Load carrying Robotugs have included fork-lift trucks, small vehicles for threading their way through congested warehouses, larger vehicles with roofs to keep the goods dry when they go in the open, and vehicles which can operate in either direction. With the present range of tractors it is possible to carry a load of one ton on the tug itself or to pull two trailers carrying three to ten tons of materials. The Robotug is designed to steer itself automatically from place to place in accordance with instructions given to it. If a number of destinations are given to the tug it will proceed to each in turn. The tug finds its route by sensing the magnetic field produced by 2 kc/s alternating current flowing in a wire buried in the ground, about half an inch below the surface.

Tugs have been designed which can carry a driver when it is desired to use them for manual operations as well as under automatic control. Such occasions are rare as a well-designed system will use them in their automatic rôle to their full capacity.

2. Robotug Control System

The Robotug steering is servo-driven by means of a geared split field d.c. servo motor. The steering bogey, in addition to carrying the front wheel, also carries a plate on which are mounted a number of coils. Two of these coils are in front of the steering wheel and straddle the buried guidance conductor, one slightly to the left and one slightly to the right. These coils detect if the steering wheel is aligned with the conductor in the ground. If the magnetic field, produced by this conductor, induces a larger field in one coil than in the other then a error signal is produced, which is amplified and drives the servo motor to position the coils symmetrically with respect to the track conductor. In this way the tug steers so that it follows exactly the buried conductor.

The other coils on the plate are used to signal the tug to start and stop, to signal the position of the tug in the system, and to pass instructions to the controlling system when a decision is required where there is a choice of routes, in order that the tug shall reach its programmed destination. The location of a tug is signalled to the track control system by a 400 c/s oscillator on the tug which energizes a coil underneath the tug. Coils buried in the ground pick up the signal when the tug passes.

In addition to the steering motor there is a traction motor to propel the tug, which is switched on or off in accordance with instructions from the tug control system. When the tug stops, brakes are automatically applied by means of an electro-magnetic solenoid. The brakes are interlocked with the traction motor so that the brakes must be off before the tug starts again. The speed of tugs is limited by safety requirements and speeds between 2 and 6 miles/hour are the most common.

The tug knows where it is in the system by counting from a fixed reference point. A count coil located in the ground is detected by a sensing coil in the tug. A count is associated with every point where the tug may be required to stop or where a decision is required between alternative routes. In the case of stopping points the tug will only stop at a stopping point if instructions have been given to it to do so, otherwise it will continue uninterruptedly to follow its route. The decision to choose between one or other route at a point of divergence in the track is made automatically by the tug in accordance with the route necessary to reach the programmed stopping points. Some routes may be shorter and require less counts than others. In this case dummy counts are fitted so that by the time the routes converge again the counting system will have clocked up exactly the same counts, regardless of the route followed.

In simple systems the tug is programmed by depressing toggle switches according to the decisions

† E.M.I. Electronics Ltd., Hayes, Middlesex.

required of the tug. Up to 25 such stopping points can be selected. In the case of a more complicated system a computer type patch board is employed and up to 90 stopping points may be programmed.

When a stopping point is selected, by depressing a switch or plugging in a patch board, and the tug is started, it will find its way by the most direct route to the programmed point in the system. It will wait at this point until re-started when it will proceed to the next programmed point and wait again.

3. Track Safety

When a number of tugs are running on a complicated system precautions must be taken to prevent them running into each other.

The well-established practice of the railways of dividing the track into blocks is employed. When a section of the track is occupied by a tug it is not possible for the section immediately astern to be entered by another tug. When a tug reaches the end of a block it will stop unless it is possible to proceed into the next block. It is only possible if the next but one block ahead is free also.

If the tug cannot proceed immediately into the next block section the track memory system records that the tug is waiting and releases it as soon as the track is clear. The position of every tug on the system is automatically monitored and if desired can be indicated on a mimic diagram by means of small electric lamps.

The stopping and starting of the tugs at the end of the block section is controlled automatically and no manual control is necessary.

There may be reasons other than the requirement for one clear block between adjacent tugs in the system for halting a tug at the end of a block. For example, a converging track may have a tug on it, and until such time as this tug has moved one clear block ahead of the halted tug, it cannot proceed. Similarly the tug may be on a track which crosses the block ahead of the halted tug. The length of the block is determined by operational requirements, but cannot be less than the length of a tug trailer train.

There are many built-in safety features in the system, for example:

(1) If there is a failure of the track current due to a mains power failure the tug will stop.

(2) If, due to a steering failure, the tug steers away from the track it will stop.

(3) If the tug runs into something a light bumper will deflect and stop the tug; since this bumper projects some way in front of the tug, it will come to rest before it hits the obstruction.

(4) Should the light bumper fail there is a second, more robust bumper which will stop the tug if depressed.

(5) If the accumulator in the tug is discharged below a safe operating level the tug will stop.

(6) There are a number of fuses and overload trips in the control system which will stop the tug in the event of overloading of any part of the equipment.

Warning devices can be fitted to indicate the approach of a tug. These can be flashing lights or klaxon horns. They can be mounted either on the tug or at specified points in the system. In the latter case the warning devices will only operate when a tug is approaching.

When routes are shared with other driver-controlled traffic this traffic can be stopped to permit the passing of a Robotug by automatically controlled traffic lights. When a Robotug has to leave a building, doors can be arranged to open and close automatically; this helps to conserve heating in the winter when the track goes into the open. (Fig. 1.)

4. Installation Design

When it is proposed to install a Robotug system in a factory warehouse etc., the best possible information of the flow of goods must be obtained. Essentially this problem resolves itself into determining: (1) where goods arrive in the system; (2) where goods leave the system; (3) where goods are stored within

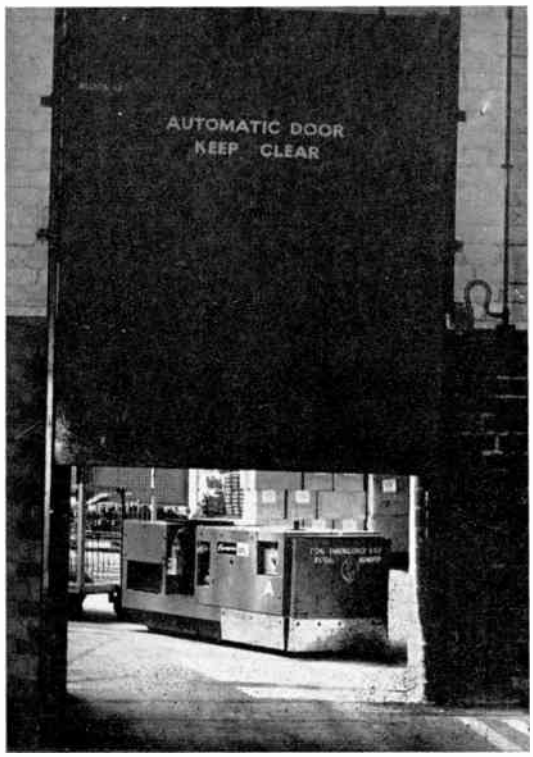


Fig. 1. Automatic doors operated by the approach of a Robotug.

Fig. 2. Robotug tractor pulling a string of trailers on a railway goods platform. The charging points for the tugs can be seen mounted on the wall.

the system. The routes connecting these various points can usually be determined fairly easily. It is obviously desirable to maintain traffic circulating in one direction only and to avoid too many places where tracks cross each other. Where tracks converge or cross it is necessary to see that an adequate length of clear track is available before this point in order that tugs can be stacked should there be any congestion at the point of convergence. In the more complicated installations it has been necessary to employ a computer to analyse the step-by-step movement of each tug in the system in order to determine if any appreciable hold-ups occur at any point in the system. Special computer programs have been prepared for this purpose. One point that emerges as a result of the block method of control is that no closed loop of track should contain an even number of blocks because when every alternate block is occupied the system seizes up and no tug can move.

Other factors that influence the design of the system are the methods employed for handling the goods. For example, are goods carried on the tugs or on the trailers; are the goods loaded on the trailers in bulk, on pallets, or separately; are the trailers left attached to the tug during loading and unloading or are they detached; do tugs pull into lay-bys for loading and unloading or stop on the tracks?

Decisions on these matters are very much influenced by the customer's requirements, although the particular requirements of Robotugs must be taken into account. For example, if trailers are uncoupled from tugs this must be done in such a way that they are not left on the track where they can be hit by another tug.

The length of the working shift and the number of shifts determines the number of tugs and the capacity of the batteries employed. Arrangements must be made for charging the batteries, chargers can be fitted on the tugs, scattered throughout the system, or mounted in special charging areas which may or may not be on the Robotug track system. The whole object of the study is to arrive at the simplest handling system which employs the minimum amount of equipment.

In considering the economic advantages of employing a Robotug system compared with other handling systems, one must consider not only the differences in capital costs involved, but the operating costs as well. Due to the move towards higher labour rates a system which employs smaller manpower is likely to become increasingly more attractive as time goes on.

5. Conclusion

The main justification for the employment of Robotugs is the economies obtained by the reduction of man power in materials handling. There are a number of incidental, but by no means insignificant, advantages in the reduction of the human element. For example, the route followed is accurately predictable and accidents caused when corners are cut with long trailer trains are avoided. The time taken for each journey is accurately known. Tractors cannot be driven over the edge of platforms, over weak floors or in any prohibited areas.

The analysis of the problem can in complicated cases involve appreciable work and a modern digital computer may be the only way to handle the problem. A detailed knowledge of the fundamentals of the tug system and an appreciation of the problems of mechanical handling are an essential basis to deal with the analysis.

Manuscript received by the Institution on 29th May 1962.
(Contribution No. 60.)

U.D.C. 621.375

Stability and Power Gain of Linear Two-Port Active Networks

By S. VENKATESWARAN, B.Sc., M.A., Ph.D., D.I.C.†

My attention has been drawn to a contribution by Cripps and Slatter.[1] It creates an impression that the "invariant stability factor", S, proposed previously[2,3] is only applicable for the conjugate matched situation and is inapplicable where there is mismatch in port terminations. This is not so. S is applicable for both "potentially unstable" and "absolutely stable" linear two-port networks. Where S is greater than unity the network is absolutely stable; where it is complex or less than unity, the network is potentially unstable. However, S equals the modulus of "internal loop loss" (inverse of "internal loop gain") for the terminations associated with maximum power gain, provided S as mathematically defined is equal to or greater than unity. The two-port terminations for a given "device network" can be artificial and even mismatched. This will be illustrated by a simple example.

Figure 1 shows a "device network" with an admittance matrix ($^d y$), whose "inherent stability factor",[2] S_i, may be less than unity, complex or greater than unity. This device network may now be modified by adding a passive "device source conductance", ^dG_S, and a passive "device load conductance", ^dG_L, such that the "modified network" has a stability factor, S, greater than both S_i and unity.[4,5]

The maximum available power gain of the modified network is now given by

$$G_{\max c} = \frac{1}{S}\left|\frac{^d y_{21}}{^d y_{12}}\right| = |G_{lc}|\left|\frac{^d y_{21}}{^d y_{12}}\right|$$

where G_{lc} is the "internal loop gain"[2,3] of the network with its terminations; the terminations y_{Sc}, y_{Lc} provide conjugate match for the "modified network", whereas the terminations $(y_{Sc}+{}^dG_S)$, $(y_{Lc}+{}^dG_L)$ provide mismatch for the "device network". Both for the "device network" and the "modified network", the susceptances at the ports ensure tuning, i.e. "total port susceptance" vanishes at either port. But the conductances at the ports ensure matching for the "modified network" but mismatching for the "device network" unless ^dG_S and ^dG_L both equal zero with $S_i \geq 1$. Any power gain $\leqslant |^d y_{21}/^d y_{12}|$ (if S_i is complex or $\leqslant 1$) or $\leqslant 1/S_i\,|^d y_{21}/^d y_{12}|$ (if $S_i > 1$) may be obtained by a suitable choice of the values of ^dG_S and ^dG_L. This clearly shows that use of S is not confined to conjugate matching of "device network"; it is equally applicable for mismatching.

I have devoted a considerable time to the study of factors based on stability[6] of linear two-port networks and their effect on power gains of both mismatched—

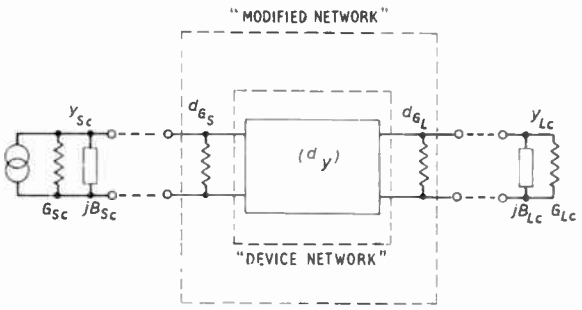


Fig. 1. "Device network" and "modified network". y_{Sc} and y_{Lc} are conjugate matched admittances for the " modified network"'

for any given real parts of passive terminations—and optimum unilateralized amplifier stages. These factors are all related to the maximum moduli of internal loop gains of the network with and without its terminations.[3,4] Based on the stability factor, S, I have proposed a general stability, power gain and bandwidth theory[4,5] of synchronously-tuned cascaded linear two-port networks; herein the individual stages may be mismatched and/or unilateralized. It results in a powerful, simple and closely accurate design procedure[4,7] for such amplifiers; this has been verified experimentally.[4,7]

References

1. L. G. Cripps and J. A. G. Slatter, "Amplifier gain and stability", *J. Brit.I.R.E.*, **22**, p. 417, November 1961.
2. A. R. Boothroyd, "The transistor as an active two-port network", *Scientia Electrica*, **7**, No. 1, pp. 3–15, March 1961.
3. S. Venkateswaran, "An invariant stability factor and its physical significance", *Proc. Instn Elect. Engrs*, **109**, C, pp. 98–102, March 1962. (I.E.E. Monograph No. 468E.)
4. S. Venkateswaran, "Stability, power gain and bandwidth of linear active four-pole networks, with particular reference to transistor amplifiers at higher frequencies", London University Ph.D. thesis, June 1961.
5. S. Venkateswaran, "Stability, power gain and bandwidth of synchronously tuned, cascaded linear active two-port networks", *Int. Conv. Rec. I.R.E.*, Part 2, pp. 49–60, 1962.
6. S. Venkateswaran and A. R. Boothroyd, "Power gain and bandwidth of tuned transistor amplifier stages", *Proc. Instn Elect. Engrs*, **106**, B, Suppl. 15, pp. 518–29, January 1960.
7. S. Venkateswaran, "A generalized design basis for synchronously tuned multi-stage linear amplifiers". *To be published.*

Manuscript received by the Institution on 4th September 1962. (*Contribution No. 61.*)

Messrs. Cripps and Slatter have agreed that there is an alternative interpretation to their contribution and thank Dr. Venkateswaran for clarifying the points in question.

† Department of Engineering, University College of Swansea, University of Wales.

U.D.C. 621.396.963

Visual Displays of Integrated Video Waveforms

By

D. C. COOPER, Ph.D. †

Presented at the Symposium on Sonar Systems in Birmingham on 9th–11th July 1962.

Summary: The operation of single loop and double loop video integrators is discussed briefly and a qualitative measure of integrator performance is given in terms of a simple video signal/noise ratio.

A description is given of a set of pulse detectability experiments which have been conducted using a single trace display of the integrator output. A "peak level" criterion of detection was used in these experiments and the results obtained show that the simple video signal/noise ratio does not give a direct measure of integrator performance. It is concluded that the performance of a video integrator is somewhat better than would be predicted by the use of the simple theory, providing that the pre-detector signal/noise ratio is greater than —5 dB.

Finally a set of theoretical results are given for comparison with the experimental ones.

List of Symbols

β	loop gain, voltage or current.
F_1	video signal-to-noise improvement factor for single loop integration.
F_2	video signal-to-noise improvement factor for double loop integration.
N	number of waveforms integrated or number of pulses in a signal.
N_e	effective number of waveforms integrated.
n	number of repetition periods for which a waveform sample is stored.
$P(y)$	probability density distribution for y.
R	envelope of r.f. waveform.
r	video signal-to-noise ratio $= s/\sigma$.
s	video signal measured as change in mean level.
σ	r.m.s. value of video noise.
ψ_0	mean square value of r.f. noise.
x	r.f. signal-to-noise power ratio.
Y	video waveform amplitude when signal is present.
y	video waveform amplitude due to noise alone.
y_v	video waveform amplitude before integration.

1. Introduction

The detection of recurrent "radio" frequency pulses in noise is usually accomplished by an observer using a visual display of the video waveforms obtained by envelope demodulation.

It is widely known that optimum detectability is obtained by the use of a pre-demodulator filter with a bandwidth approximately equal to the reciprocal of the pulse length. The experimental work of Lawson and Uhlenbeck[1] using the deflection modulated or type A display, and of Payne-Scott[2] using the intensity modulated p.p.i. display, suggests an optimum bandwidth of 1·2 times the reciprocal of the pulse length. In addition the results obtained by the above workers give an integration exchange rate which corresponds to a 1·5 dB reduction in threshold r.f. signal/noise ratio for a doubling of the number of video traces integrated on the display.

If the detectability of a pulse on a video display is assumed to be determined by the video signal/noise ratio r defined as

$$r = \frac{\text{change in mean level of waveform due to presence of signal}}{\text{r.m.s. of noise in the absence of signal}}$$

and the integration process is taken to be ideal, the above exchange rate can be justified[3] providing the r.f. signal/noise ratio x is less than unity. However, for large values of x this simple approach will predict an exchange rate of 3 dB per doubling.

The fact that the exchange rate of 1·5 dB per doubling appears to apply in the case of the type A and p.p.i. displays when x is considerably greater than unity, indicates that the above simple approach is not realistic, and some experimental work described

† Electrical Engineering Department, University of Birmingham.

by McGregor,[4] and discussed in detail by Tucker[5] and Griffiths,[3] shows that an improved exchange rate can be obtained by the use of an intensity modulated display which presents traces side-by-side. An example of this type of presentation is that produced by the chemical recorder.

It is the purpose of this paper to describe a set of experiments which have been performed to determine whether the performance of the deflection-modulated display can be improved by the use of suitable video processing. An electronic integration system was adopted and the processed video information was presented to an observer as a single deflection-modulated trace.

It is realized that deflection-modulated displays are now rarely used in practical situations, but a single deflection-modulated trace is much more easily interpreted than an intensity-modulated one and the benefit, or otherwise, resulting from the use of an electronic integrator should be obtained with either type of display.

2. Integration

2.1. *Storage*

Essentially all integrators incorporate some form of storage which permits the video traces which are to be integrated, or added, to be brought into time coincidence. The integration effects obtained by the use of cathode-ray tube displays are undoubtedly due to screen persistence and human visual persistence, and storage by these means will involve some loss of "memory" as the storage time increases.

2.2. *Electronic Integrators*

2.2.1. The use of simple systems

Electronic integrators can, in principle, incorporate storage which involves negligible loss of "memory" or information as storage time is increased, but such systems tend to be complicated and costly when a large number of traces are to be integrated. However, if some loss of information can be tolerated a simple and relatively inexpensive integrator can be constructed by using a small number of delay loops in cascade.

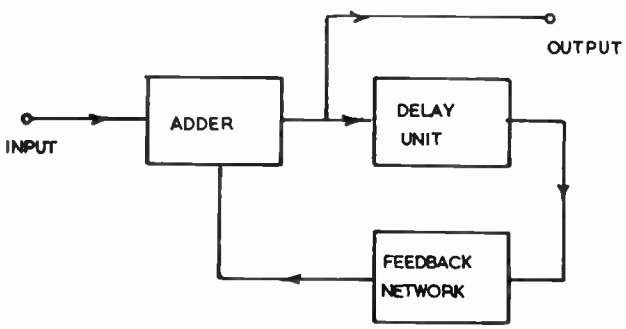


Fig. 1. Block diagram of loop integrator.

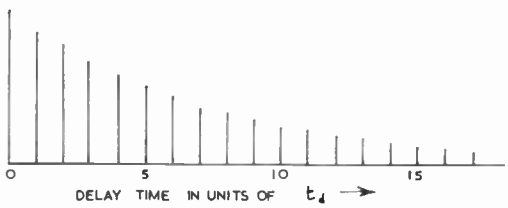


Fig. 2(*a*). Impulse response of single loop integrator.

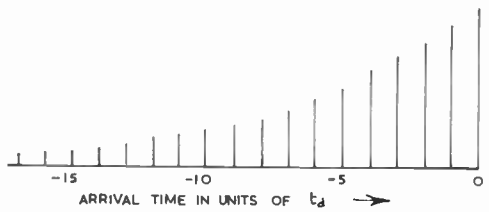


Fig. 2(*b*). Weighting function of single loop integrator.

The operation of such systems is described in some detail in reference 6 and, therefore, only a brief account will be given here.

2.2.2. The single loop integrator

The basic single loop integrator is shown diagramatically in Fig. 1. A delay unit is used in a feedback loop with overall loop gain β.

The integration performance of this arrangement is best considered by first determining its impulse response. A unit impulse applied to the integrator at time $t = 0$ will result in a unit output impulse at the same time followed, after a time equal to the loop delay, by an impulse of magnitude β, then a further impulse of magnitude β^2 after another delay equal to that produced by the loop, and so on indefinitely with each additional journey around the loop modifying the impulse magnitude by a factor β.

Since the loop is only stable when β is less than unity the impulse response will have the form shown in Fig. 2(*a*), and the output arising due to any input waveform will be the weighted sum of the input amplitudes at the observation instant and all times which are earlier by integral multiples of the loop delay. The weighting function for the integration is given by the reversed impulse response shown in Fig. 2(*b*).

If, as in the experimental tests, the single loop integrator is used to integrate a video waveform containing a continuously repeated pulse, and the loop delay is matched to the repetition period of the pulses, the output waveform will have a video signal/noise ratio given by

$$r_{1\infty} = \frac{s(1+\beta+\beta^2+\ldots)}{\sqrt{\sigma^2(1+\beta^2+\beta^4+\ldots)}} \qquad \ldots\ldots(1)$$

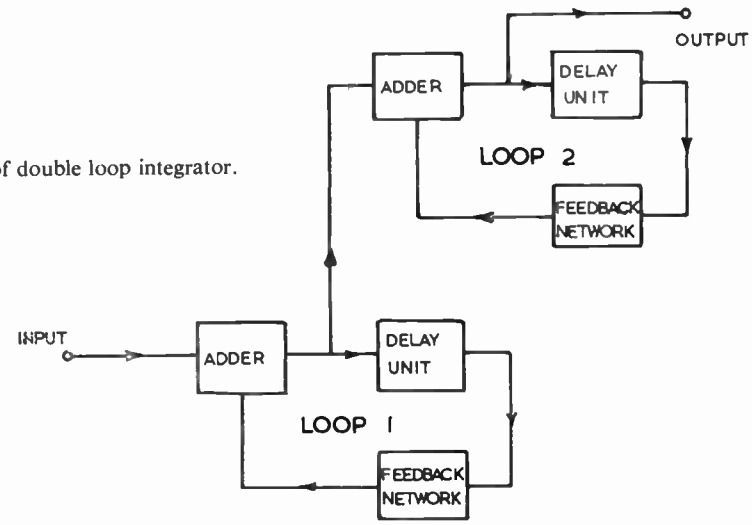


Fig. 3. Block diagram of double loop integrator.

where s = change in mean level of input due to presence of pulse and σ = r.m.s. of noise at input. Thus

$$r_{1\infty} = \frac{s}{\sigma}\sqrt{\frac{1+\beta}{1-\beta}} = r\sqrt{\frac{1+\beta}{1-\beta}} \qquad \text{.....(2)}$$

where r = input video signal/noise ratio, and the single loop integrator operating on the continuously-repeated input pulse has improved the video signal/noise ratio by the factor

$$F_{1\infty} = \sqrt{\frac{1+\beta}{1-\beta}} \qquad \text{.....(3)}$$

2.2.3. The double loop integrator

Two single loop integrators in cascade form the double loop system which is illustrated in Fig. 3, and this arrangement has been shown[6] to have a number of virtues when it is used with narrow beam search radar or sonar systems.

It is easily shown[6] that the impulse response of the double loop integrator, with each loop of gain β, consists of output impulses of magnitude $(n+1)\,\beta^n$ at times nt_d, where t_d is the loop delay and n is any positive integer or zero. The weighting function for the integration performed by the double loop arrangement is again the reversed impulse response, and the resulting function is illustrated in Fig. 4.

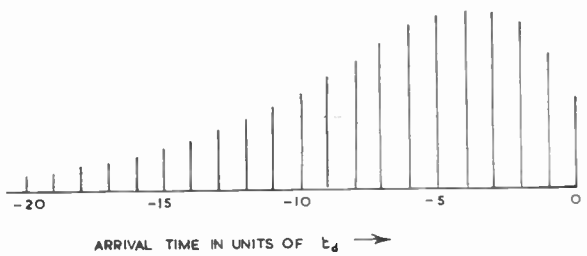


Fig. 4. Weighting function of double loop integrator.

The output signal/noise ratio due to the double loop integration of a continuously repeated pulse will be

$$r_{2\infty} = \frac{s(1+2\beta+3\beta^2+4\beta^3+\ldots)}{\sqrt{\sigma^2(1+4\beta^2+9\beta^4+16\beta^6+\ldots)}} \qquad \text{.....(4)}$$

$$= \frac{s}{\sigma}\sqrt{\frac{1+\beta}{1-\beta}}\frac{1+\beta}{\sqrt{1+\beta^2}} \qquad \text{.....(5)}$$

and the video signal/noise ratio is therefore improved by the factor

$$F_{2\infty} = \sqrt{\frac{1+\beta}{1-\beta}}\frac{1+\beta}{\sqrt{1+\beta^2}} \qquad \text{.....(6)}$$

It is worth noting that the addition of the second loop never increases the signal/noise ratio by a factor greater than $\sqrt{2}$ since the input waveforms to this loop have been partially correlated by the action of the first loop.

3. The Experimental Apparatus

3.1. *The Integrator*

3.1.1. Mode of operation

Single and double loop integrator operation was obtained by the use of one or both of the loops of a double loop system which utilized a single mercury acoustic delay line giving a delay of approximately 750 μs. The general arrangement of the system is shown in Fig. 5.

Separate carrier frequencies of 6 Mc/s and 8 Mc/s were used in the delay unit for each of the loops, and each video waveform was transmitted as an amplitude modulation of the appropriate carrier. Suitable filtering arrangements at the output of the delay unit ensured that negligible interaction resulted from the use of a common delay medium, and by careful

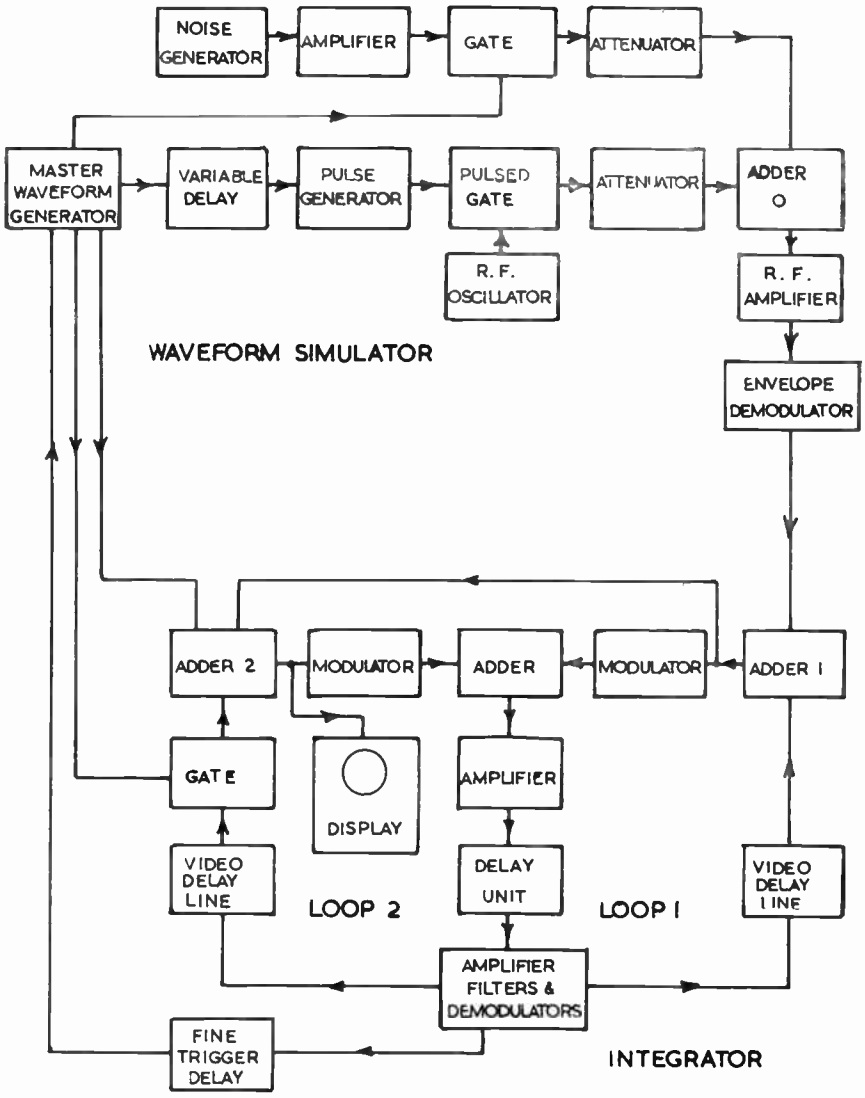


Fig. 5. Block diagram of waveform simulator and double loop integrator.

adjustment of the filter responses two carrier channels each of 1 Mc/s bandwidth were obtained.

3.1.2. Setting of loop gain

The gain setting was made with no noise present. Initially the gain of each loop was reduced to zero and the amplitude of the input pulse was set to a value at which the output pulse could readily be measured with a calibrated oscilloscope. The gain of one loop was then increased until the output amplitude had increased by the factor $1/(1-\beta)$ and when this was done the loop gain had been set to the required value β.

If it was desired to use both loops as a double integrator system the gain of the second loop was finally adjusted to obtain a further increase in output amplitude by the factor $1/(1-\beta)$. Thus both loops could be set to have the same gain β and this condition was always used in double loop experiments.

3.2. *The Waveform Simulator*

3.2.1. Mode of operation

The arrangement of the waveform simulator is shown as part of Fig. 5.

A trigger pulse circulating through the delay unit channel in one integrating loop was used to control the repetition period of the generated r.f. pulses, and a fine synchronization control was provided by using an adjustable delay of the trigger pulse to compensate for a fixed delay of approximately 4 μs introduced into each integrating loop. The fixed delay lines also provided a means for correcting any difference between the delays of the channels in the mercury delay unit.

Each trigger pulse was circulated at a time which

did not fall in the observed interval of the video trace, and the gain of the integrating loop which carried the trigger pulse was suppressed for a period of time embracing its position. This was done in order to prevent the amplitude of the trigger pulse, and therefore the accuracy of synchronization, being affected by the loop gain.

At a selected position in each observation interval a 12 μs pulse was generated at a carrier frequency of 4·5 Mc/s. These pulses were passed through a calibrated attenuator and added to r.f. noise which had also passed through a second calibrated attenuator. The noise was suppressed during the loop gain suppression period in order to prevent it perturbing the trigger pulse shape.

The combined r.f. pulses and noise were passed through an amplifier with a 3 dB bandwidth of 100 kc/s, which corresponded to the value of 1·2 times the reciprocal of the pulse length, and then demodulated by an envelope rectifier which gave a nearly linear envelope-to-output characteristic.

3.2.2. Setting of r.f. signal/noise ratio

The signal/noise ratio at the input to the envelope demodulator of the simulator was set to unity by measuring the d.c. component of the demodulated waveform. Noise and continuous signal frequency oscillations were applied, one at a time, and the noise level was adjusted, at the noise generator, to obtain equal indications in each case. The continuous signal was obtained by switching out the pulsed gate used in the normal simulator operation.

It can be shown that when the d.c. components have been equalized, and a linear rectifier is used for demodulation, we have

$$\frac{\text{r.m.s. amplitude of r.f. noise}}{\text{r.m.s. amplitude of r.f. oscillation}} = 1{\cdot}13 \quad \ldots(7)$$

Thus the r.f. signal/noise ratio for the continuous oscillation had been set to approximately −1 dB. However, the arrangement used to provide the switching out of the pulsed gate was designed to give a continuous signal whose amplitude was 1/1·13 times that for the pulses, and thus when normal pulsed working was resumed the signal/noise ratio was set to unity.

After obtaining this setting a large range of r.f. signal/noise ratios could be obtained by the use of the calibrated attenuators in the signal and noise channels.

4. The Assessment of Integrator Performance

The evaluation of the detectability of the pulse in the integrated video waveforms was made in the following way.

It was assumed that, so far as the observer was concerned, the pulse could lie at any position and the observer would choose a position corresponding to the maximum or peak level reached during a single deflection-modulated trace. The single traces were either recorded photographically or presented on a long persistence cathode-ray tube, and at least 300 single traces were obtained for each set of conditions.

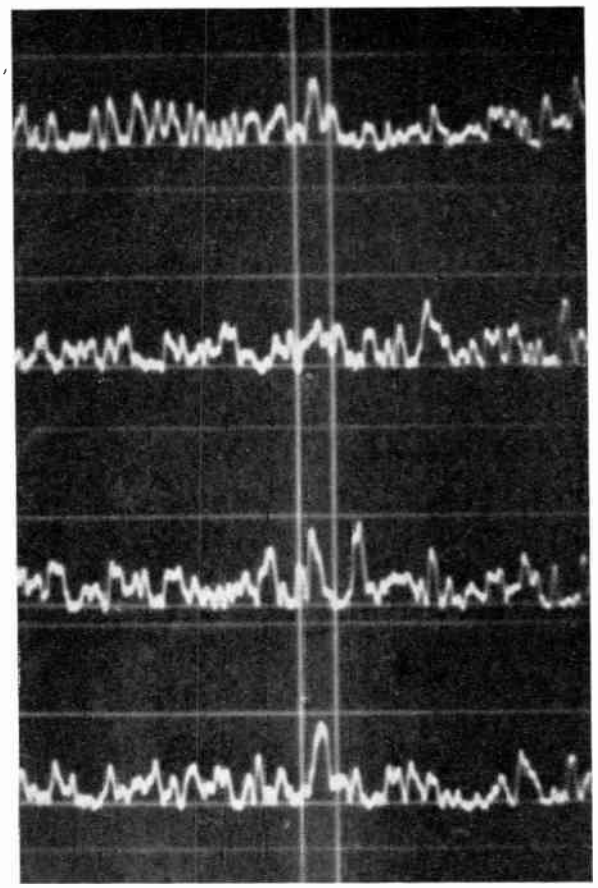


Fig. 6. Examples of video display waveforms.

Subjective aspects did not enter into this type of decision since the observer could study the records carefully, and if traces with two or more positions of equal peak amplitude were obtained they were not accepted as trials.

The probability of correct detection was obtained simply by determining the proportion of the number of acceptable trials in which the position of the peak waveform amplitude corresponded to the known pulse position or, more precisely, fell in the region occupied by the pulse above its half amplitude points. In order to illustrate the above comments a number of recorded traces are shown in Fig. 6. The actual position of the pulse is determined by the pair of lines crossing the traces. In the experiments the observed

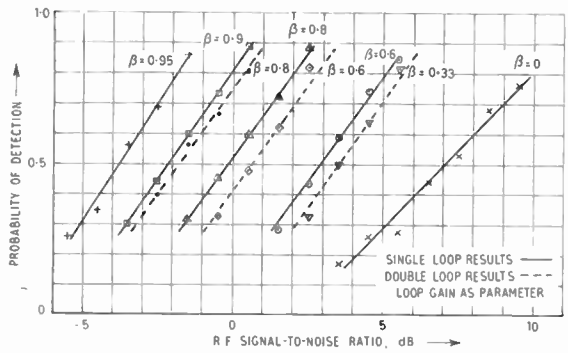


Fig. 7. Experimental results for loop integrator performance.

trace length corresponded to 520 μs or 43 times the pulse length.

It was obvious that the use of the above method for determining the probability of correct detection did not depend on the position of the pulse and the majority of the tests were conducted with the pulse at the centre of the displayed trace.

The results obtained for both single and double loop integration are shown in Fig. 7. It will be noted that straight lines have been drawn through the sets of points to approximate to the central portions of the expected curves.

5. Analysis of the Experimental Results

5.1. *The Effective Numbers of Integrations for Loop Integrators*

In the experimental work the integrated waveforms were the result of the weighted addition of an infinite number of video input waveforms. Actually we would like to determine how the threshold of detectability is affected by doubling a finite number of integrated traces, and we are therefore faced with the problem of defining an effective number of integrations for each of the experimental situations.

Fortunately the video signal/noise ratio, as defined previously, may be used to provide a means for comparing the performance of different integration processes.

In section 2 it has been shown that the video signal/noise improvement factors for single and double loop integrators operating on a continuously repeated input pulse are

$$F_{1\infty} = \sqrt{\frac{1+\beta}{1-\beta}}$$

and

$$F_{2\infty} = \sqrt{\frac{1+\beta}{1-\beta}} \cdot \frac{1+\beta}{\sqrt{1+\beta^2}}$$

If the threshold r.f. signal/noise ratio is now defined as being that required to give a probability of successful detection of 0·5, the experimental results of Fig. 7 may be used to produce two sets of points on a graph which has threshold r.f. signal/noise ratio as the abscissa and video signal/noise improvement factor as the ordinate. Each set of points corresponds to either the single loop or the double loop results.

The graph so produced is shown in Fig. 8 and it will be seen that both sets of points lie close to a single line. This line is virtually straight when decibel scales are used for each axis.

Since a single line is sufficient to define the performance of both single and double loop integrators we may suggest that it is the improvement factor which determines the results in our experiments, and that the weighting which is employed to obtain the given improvement factor is not particularly important.

If this suggestion is correct it is obvious that the equivalent number of uniformly weighted integrations, for any particular value of β in the single or double loop systems, may be defined as that number which results in the same video signal/noise ratio improvement factor. Therefore, since N uniformly weighted integrations increase the video signal/noise ratio by the factor $\sqrt{N}$, the effective numbers for the single and double loop systems are,

$$N_{e1} = \frac{1+\beta}{1-\beta} \qquad \text{.....(8)}$$

and,

$$N_{e2} = \frac{(1+\beta)^3}{(1-\beta)(1+\beta^2)} \qquad \text{.....(9)}$$

The use of the improvement factor as a measure of the effective number of integrations means that the effective number is doubled when the improvement factor is increased by 3 dB. Thus the slope of the line in Fig. 8 provides a measure of the reduction of the threshold r.f. signal/noise ratio resulting from the doubling of the effective number of pulses integrated. The value obtained for this reduction is 2·17 dB per doubling, and this result indicates that integration followed by detection, based on the criterion used, gives a similar improvement in threshold level to that given by visual correlation on the chemical recorder display.

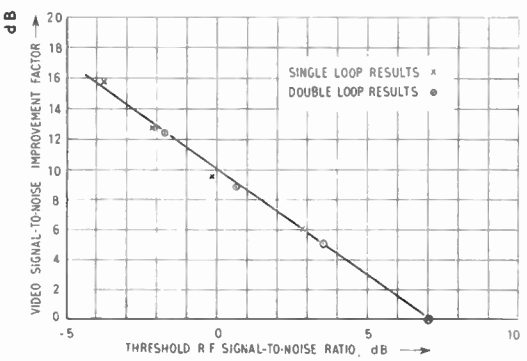


Fig. 8. Threshold r.f. signal/noise ratio versus video signal/noise improvement factor.

5.2. *Comparison of Experimental and Theoretical Results*

In order to provide a further check on the validity of the use of the video signal/noise improvement ratio as a means for comparing the results of integration, detection probabilities will be calculated for various r.f. signal/noise ratios with N waveforms integrated. Thus if N is chosen to correspond to the effective number of integrations for a value of loop gain used in the experimental work, the theoretical results may be expected to be close to the experimental results for the given value of loop gain.

The calculations are simplified a little if we consider the integration of N video waveforms which are produced after an envelope demodulation process with a square-law characteristic. Thus the instantaneous amplitude Y_v of the video waveform is related to the instantaneous amplitude R of the r.f. waveform by the expression

$$y_v = \frac{R^2}{2\psi_0} \qquad \text{.....(10)}$$

where ψ_0 = mean square amplitude of the r.f. noise. This relation has been normalized for convenience in the work which follows.

It must be admitted that a linear demodulator characteristic was used in the experimental work but Marcum[7] has shown that the results obtained with linear and with square law demodulator characteristics are almost identical and, therefore, the assumption of a square law characteristic is not expected to introduce appreciable errors.

We now proceed using the probability density functions, derived by Marcum, for the integrated noise and the integrated signal-plus-noise. Thus, at the signal position the amplitude Y of the integrated waveform has the signal-plus-noise distribution,

$$P_N(Y) = \left(\frac{Y}{Nx}\right)^{\frac{N-1}{2}} e^{-Y-Nx}\, I_{N-1}(2\sqrt{NxY}) \quad (11)$$

where x = r.f. signal/noise power ratio

and I_{N-1} = modified Bessel function of order $N-1$,

and at all other positions, where noise only is present, the integrated noise distribution is

$$P_N(y) = \frac{y^{N-1}\, e^{-y}}{(N-1)!} \qquad \text{.....(12)}$$

Since the video noise is the result of the envelope demodulation of bandwidth limited r.f. noise, two instantaneous values of the video noise waveform will be correlated if they are separated by a very small time interval. There will be an appreciable amount of correlation (as expressed by the autocorrelation function of the video waveform) for intervals which are much less than the reciprocal of the r.f. bandwidth in cycles per second, but the correlation will be small when the interval takes this value and remains small for larger intervals.

Therefore, it is often assumed that independent samples of the noise are obtained at intervals of one pulse length when the receiver r.f. bandwidth has been made approximately equal to the reciprocal of the pulse length. Thus during a 520 μs trace, corresponding to that used in the experimental work there are 520/12 or approximately 43 instants at which the noise amplitudes may be regarded as being independent, since a pulse length of 12 μs was used.

Hence if an amplitude Y is obtained at the pulse position, the probability of all the independent noise amplitudes at the remaining 42 instants being less than Y is

$$\left[\int_0^Y P_N(y)\,dy\right]^{42}$$

where $P_N(y)$ is given by eqn. (12).

For a large number of experiments all possible values of Y will be obtained with frequency determined by the probability distribution of Y at the pulse position. Hence the overall probability of the amplitude at the pulse position being greater than all other amplitudes, is given by

$$P(x, N) = \int_0^\infty P_N(Y)\left[\int_0^Y P_N(y)\,dy\right]^{42} dY \quad \text{...(13)}$$

where $P_N(Y)$ and $P_N(y)$ are given by eqns. (11) and (12).

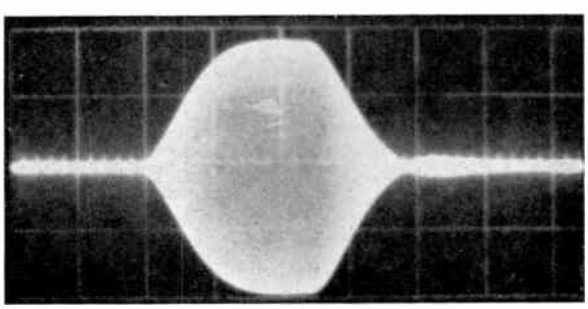


Fig. 9. Pulse shape used in experiments.

At this point it should be stated that the r.f. signal/noise ratio used to obtain the experimental results of Fig. 7 was the peak value, which was only reached during a small part of the pulse duration. This is illustrated in Fig. 9 which is a photograph of the signal pulse at the output of the simulator r.f. amplifier for a constant amplitude input pulse.

In the experimental work it was necessary to regard the pulse position as being defined by a decision interval during which the instantaneous voltage amplitude of the pulse, in the absence of noise, was

about one pulse length in duration and the experiments were concerned with a period in which the instantaneous power rose from one quarter of its maximum value, to the maximum value, and then fell to its original level.

From the theoretical point of view a peak video amplitude, in the presence of noise, will have a certain probability of being at a particular place in the decision region, and obviously the probability will be greatest at the position where the pulse reaches its maximum amplitude. Thus, ideally, the expression of eqn. (13) provides the probability that the amplitude at a particular point in the decision interval is greater than the amplitude at all of the independent sampling points. To obtain the probability when all peaks in the decision interval are accepted, the above single point result should be weighted by the probability of a peak occurring at the chosen instant and the result summed for all instants in the decision interval, taking due account of the variation of the r.f. signal/noise ratio during this period.

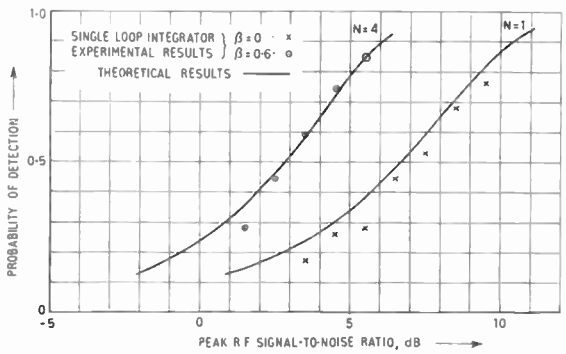


Fig. 10. Comparison of experimental and theoretical results.

This would be a most difficult task and matters are simplified by assuming that the pulse has a constant power level in the decision interval, this being the mean level evaluated for the actual pulse shape used. This mean level, not the peak value, was used in eqn. (13), and for the pulse shown in Fig. 9 it was found that the mean power level during the decision interval was 1 dB below the peak value.

Values of the probability of correct detection, obtained in the above manner, have been computed by numerical methods. The inner integral was expressed in terms of Pearson's incomplete gamma function, for which tables existed, and this reduced the labour involved quite considerably. However the computation was still quite lengthy, and the values of N considered were limited to two.

The values of N taken were 1 and 4 and the results of the computations are shown in Fig. 10. Also shown in this figure are the experimental results obtained with the single loop integrator operating at loop gains of 0 and 0·6, since these loop gains correspond to effective integration numbers of 1 and 4. The experimental points lie very close to the theoretical curves, thereby justifying the definition of the effective number of integrations.

6. Conclusions

In the experimental work two forms of the delay loop integrator were used to produce a single trace video display which was assessed by an observer in a logical manner using a peak level criterion of detection.

The weighting of the integration performed by the loop integrators was non-uniform and it was found that an effective number of uniformly weighted integrations could be defined in terms of the improvement produced in the video signal/noise ratio, thus permitting the comparison of the performance of integrators with uniform and non-uniform weighting.

As far as threshold detection is concerned it may be concluded that the electronic integration of video waveforms, followed by a single trace display, is equally as efficient as the use of the chemical recorder display, and both systems appear to be more efficient than type A or p.p.i. displays in which phosphor integration occurs. However, it must be remembered that the detection criterion used here is only applicable in cases where a pulse is known to exist and its position must be determined.

7. Acknowledgments

The author wishes to record his appreciation of the assistance given by the Mullard Research Laboratories in providing the mercury delay line and the associated amplifiers, and the many valuable discussions with his colleagues, particularly Dr. J. W. R. Griffiths.

8. References

1. J. L. Lawson and G. E. Uhlenbeck, "Threshold Signals", M.I.T. Radiation Laboratory Series, Vol. 24 (McGraw Hill, New York 1950).
2. R. Payne-Scott, "The visibility of small echoes on radar p.p.i. displays", *Proc. Inst. Radio Engrs*, **36**, p. 180, February 1948.
3. J. W. R. Griffiths, "Detection of pulse signals in noise: the effect on visual detection of the area of the signal paint", *J. Brit.I.R.E.*, **17**, p. 330, June 1957.
4. P. McGregor, "A note on trace-to-trace correlation in visual displays", *J. Brit.I.R.E.*, **15**, p. 329, June 1955.
5. D. G. Tucker, "Detection of pulse signals in noise: trace-to-trace correlation in visual displays", *J. Brit.I.R.E.*, **17**, p. 319, June 1957.
6. D. C. Cooper and J. W. R. Griffiths, "Video integration in radar and sonar systems", *J. Brit.I.R.E.*, **21**, p. 421, May 1961.
7. J. I. Marcum and P. Swerling, "Studies of target detection by pulsed radar", *Trans. Inst. Radio Engrs (Information Theory)*, **IT-6**, No. 2, April 1960. (Special Monograph Issue.)

Manuscript received by the Institution on 15th June 1962. (Paper No. 801/SS27.)

POINTS FROM THE DISCUSSION

Dr. E. J. Risness: There appears to be an essential difference between the role of the observer in these measurements, as compared with other detectability experiments such as those of Griffiths and Nagaraja presented later in this Symposium, in that the criterion to be used by the observer is in this case exactly known. Hence the outcome of the experiments can in principle be calculated theoretically (the author has in fact done so) and one is then tempted to ask why it was necessary to carry out the experimental measurements at all. Three possible reasons can be distinguished: (*a*) the theoretical calculations are unreliable, or too difficult, (*b*) it is not known whether an experimental equipment can be made to come up to theoretical prediction, (*c*) it is desirable to demonstrate a working equipment. I think all three reasons have contributed (though strictly speaking (*a*) and (*b*) are mutually exclusive). Perhaps Dr. Cooper could comment on this.

The Author (*in reply*)*:* I believe Dr. Risness is correct when he states that the outcome of the detectability experiments can in principle be calculated theoretically. However, the theoretical approach which has been given in the paper is based on a number of simplifying assumptions and it is the validity of these assumptions which has been verified by the experimental work.

Therefore, in reply to the specific points made by Dr. Risness the following comments are appropriate.

(*a*) An exact theoretical treatment will be exceedingly laborious.

(*b*) A practical attempt to meet a theoretical prediction sometimes shows up errors in the theoretical postulates.

(*c*) It is necessary to have a working equipment in order to follow up item (*b*).

Radio Engineering Overseas . . .

The following abstracts are taken from Commonwealth, European and Asian journals received by the Institution's Library. Abstracts of papers published in American journals are not included because they are available in many other publications. Members who wish to consult any of the papers quoted should apply to the Librarian, giving full bibliographical details, i.e. title, author, journal and date, of the paper required. All papers are in the language of the country of origin of the journal unless otherwise stated. Translations cannot be supplied. Information on translating services will be found in the Institution publication "Library Services and Technical Information".

TELEVISION NETWORK PLANNING

A recent article by two German engineers describes a method for planning a television network which is an extension of the technique previously developed by the authors for v.h.f. networks. The present method enables account to be taken in a systematic manner not only of the interference between channels in the same band, but also the interference due to emissions in another band, for example, that due to radiation from receiver oscillators. The account is concluded with an example of application to one of the basic networks of the 1961 Stockholm Conference (E.B.U. network). This application shows particularly that the use of channel widths of 7 Mc/s in Bands I and III and of 8 Mc/s in Band IV/V is more advantageous than a unique channel width of 8 Mc/s in all the bands.

"Establishment of a channel-distribution plan for a network of television transmitters in different bands", H. Eden and H. W. Fastert. *European Broadcasting Union Review*, 76-A Technical, pp. 272–7, December 1962.

PARAMETRIC AMPLIFICATION

The low thermal noise level in parametric amplifiers makes them particularly suitable for use in communication circuits using tropospheric diffusion. For the purpose of such an application, a study has been made by a French engineer of a reflection-type parametric amplifier, which uses a circulator working in the 830 to 960 Mc/s band. Calculations have been made of the noise temperature and of the gain within the 3 dB limit of such an amplifier in optimum conditions. Practical measurements made on the characteristics of this amplifier agree well with the calculated values (noise temperature of the order of 80° K band gain 3 dB over 100 Mc/s).

"Wide-band parametric amplifier for tropospheric communication", R. Baud. *Onde Electrique*, **42**, pp. 987–91, December 1962.

TUNNEL DIODE OPERATION

The presence of negative resistance in the characteristic of a tunnel diode is not only useful in many practical applications but is also responsible for the appearance of unstable operating areas, which must be controlled if the full advantages of the diode are to be realized. By reference to a simple circuit diagram, a French engineer examines the different possible unstable conditions and classifies them in relation to the parameters of the diode and of the circuit in which it is used.

"Non-linear operating conditions in tunnel diodes", J. Revuz. *Onde Electrique*, **43**, pp. 47–55, January 1963.

TRANSISTOR RING COUNTER

Complementary *pnp-npn* transistor pairs, used in circuits where they are allowed to saturate, provide reliable switching in a low-cost ring counter developed by Atomic Energy of Canada. Switching speed is dependent on associated components and physical configuration, and circuits for 200 kc/s (using three stages) and 1 Mc/s (using 10 stages) are shown.

"Ring-counter circuit uses complementary pair transistors as a *pnpn* structure", J. Leng and J. C. Irvine. *Canadian Electronics Engineering*, **7**, pp. 45–9, January 1963.

TRANSISTOR PERFORMANCE

It is shown in a recent German paper how thermal relaxation effects can influence the small-signal behaviour of transistors at low frequencies. These effects are important in the frequency range 0 or 1 Mc/s especially for mesa and planar transistors having low thermal time-constants and high cut-off frequencies ($F_{\beta 1} \simeq 500$ to 1000 Mc/s). Because of such thermal relaxation the frequency and bias dependence of most fourpole-parameters (especially the *y*-parameters) may be very different from what is expected from the well-known equivalent circuits, which are all derived under the assumption of a time-constant junction temperature. Formulae are stated for the *h*- and *y*-parameters, where this effect is considered. For example a strong frequency dependence of the transconductance $y_{21e} \simeq -y_{21b}$ was found by analysis and experiment at low frequencies. The output admittances h_{22e} and y_{22e} which are normally capacitive, may further become inductive because of this effect. Finally, a new method is proposed for determining the thermal equivalent circuit of transistors.

"The influence of thermal relaxation on the small-signal behaviour of transistors ('mitlauf' effect)", O. Müller. *Archiv der Elektrischen Ubertragung*, **17**, pp. 13–28, January 1963.

TUNNEL-DIODE MEASUREMENTS

A German engineer has described a method devised for measuring the barrier-layer capacitance of tunnel diodes, in particular within the dropping section of their current/voltage characteristics in which the capacitance is used for detuning a resonant circuit. It turns out that, between peak current and valley current, the capacitance passes through a minimum at the point of maximum negative conductance. A possibility is pointed out how to measure the characteristic in a stable circuit. A checking measurement finally confirms the result of the test.

"A method of measuring the characteristics and barrier-layer capacitance of tunnel diodes", K. Christ. *Archiv der Elektrischen Ubertragung*, **17**, pp. 42–8, January 1963.

RADIO ASTRONOMY

The February 1963 issue of *The Proceedings of the Institution of Radio Engineers Australia* contains 22 papers on radio astronomy. The guest editor for this issue was the late Dr. J. L. Pawsey, F.R.S., assistant chief, Division of Radiophysics, C.S.I.R.O., Australia. The Division has itself contributed 14 of the papers and these give an impressive view of Australian work in this field.

"The Australian 210-foot radio telescope"—E. G. Bowen and H. C. Minnett

"The Australian 210-foot reflector and its research program"—J. G. Bolton

"Receivers in radio astronomy"—B. F. C. Cooper

"Development of parametric amplifiers for radio astronomy"—B. J. Robinson

"A 1400 Mc/s continuum radiometer"—F. F. Gardner and D. K. Milne

"Cross-type radio telescopes"—B. Y. Mills

"The 19 Mc/s Mills cross and absorption in interstellar gas"—M. M. Komesaroff

"A compound interferometer with a 1·5 minute of arc fan beam"—N. R. Labrum, E. Harting, T. Krishnan and W. J. Payten

"The Sydney University cross-type radio telescope"—B. Y. Mills, R. E. Aitchison, A. G. Little and W. B. McAdam

"Solar observations at a wavelength of 20 cm with a crossed-grating interferometer"—W. N. Christiansen and R. F. Mullaly

"Techniques for the investigation of solar radio bursts at metre wavelengths"—K. V. Sheridan

"Apparatus for investigating the angular structure of radio sources"—P. A. G. Scheuer, O. B. Slee and C. F. Fryar

"A multi-channel hydrogen line (λ21 cm) receiver"—R. X. McGee and J. D. Murray

"An equipment for combined geophysical and astronomical measurements of meteors"—A. A. Weiss and W. G. Elford

"The v.l.f. radio emissions from the Earth's outer atmosphere"—G. R. A. Ellis

"Radio astronomy in France"—E. J. Blum, A. Boischot and J. Lequeux

"Current radio astronomical research in the Netherlands"—G. Westerhout

"The Benelux cross antenna project"—W. N. Christiansen, W. C. Erickson and J. A. Högbom.

"The realization of giant radio telescopes by synthesis techniques"—A. Hewish

"Radio astronomy at Jodrell Bank"—R. D. Davies

"Some radio telescopes in the U.S.S.R."—P. D. Kalachov

"Radio astronomy in Japan"—T. Hatanaka

Radio Astronomy issue, *Proceedings of the Institution of Radio Engineers Australia*, **24**, pp. 94–251, February 1963.

TELEVISION SATELLITE TRANSMITTERS

To secure adequate coverage of Norway by the v.h.f./f.m. service, the Norsk Rikskringkasting has to undertake the development of low-power rebroadcasting stations. The first such transmitter built in the N.R.K. laboratories, which utilized the transposer or "frequency-converter" principle, employed ordinary thermionic valves. This was followed by a much more compact version suitable for installation in a weatherproof casing in the open air. Finally, the second version was redesigned using transistors exclusively in place of thermionic valves.

The circuit ultimately adopted and the function of the various stages, which utilized thirteen transistor type OC 171 and two type 2N 1493, the output power rating being 0·5 watt, is described by an engineer of the N.R.K. laboratories in a recent paper. The unit can be used to feed a valve amplifier having an output of ten watts. The paper concludes with some notes on the behaviour of the equipment.

"Transistorized rebroadcasting transmitters for f.m. broadcasting", T. Ovensen. *European Broadcasting Union Review*, No. 75-A, pp. 198–204, October 1962. (In English.)

RECORDING TELEVISION SIGNALS

In the apparatus described in a recent Dutch paper, a 2·5 cm wide magnetic tape is wound at a speed of 38 cm/s in a helical line around almost the entire periphery of a stationary drum 305 mm in diameter, while a ferroxcube recording head, carried on the edge of a disc rotating around the drum axis at 50 rev/s, travels in a slit in the drum. In this way the surface of the tape is filled with adjacent tracks about 1 m long set obliquely across the tape, each of which contains a complete television frame.

"An experimental apparatus for recording television signals on magnetic tape", F. T. Backers and J. H. Wessels. *Philips Technical Review*, **24**, pp. 81–3, 1962–3. (In English.)

TELEVISION TEST EQUIPMENT

A pulse-and-bar signal generator for the testing of 625-line standard television transmission facilities using transistorized circuits has been described by an Australian engineer. The device supplies either T or $2T$ sine-squared shaped standard H-composite signals, synchronized from either an internal crystal oscillator or any external television or driving signal having H-sync components. Apart from the test signal, c.r.o. double-trigger pulses of variable timing are provided to obtain a superimposed pulse-and-bar display on commonly employed oscilloscope types. The generator may be either rack-mounted or used as portable equipment and both mains or batteries are provided as primary power sources. Module construction of circuit assemblies facilitates the application of plug-in printed circuit techniques.

"Transistorized pulse and bar generator", A. J. Seyler. *Proceedings of the Institution of Radio Engineers Australia*, **23**, pp. 36–42, January 1963.

SITE NOISE PREDICTION

In the v.h.f. and u.h.f. ranges, one of the major parameters associated with the prediction of radio-telephone system performance is the effective noise factor. This

factor is a combination of the receiver's internal noise factor, a relatively small and constant component, and another noise factor, called the site noise factor, which assesses the intensity of the noise fields enveloping the receiver's antenna. Although this latter component has long been known to vary from site to site, its magnitude for planning purposes has hitherto been arbitrarily set at one of several discrete levels, the one chosen being determined solely by the size of the city in the vicinity of the site. A recent Australian paper shows that predictions of greater accuracy are obtained when the site noise is related to the nearby traffic density, and as a result of measurements taken at various sites in Victoria and New South Wales, suggests the law between them. The measurement technique and the individual results and their anomalies are also discussed in detail.

"Site noise and its correlation with vehicular traffic density", A. G. Ellis. *Proceedings of the Institution of Radio Engineers Australia*, **24**, pp. 45–52, January 1963.

NOISE IN SPACE-CHARGE DIODES

The reduction factor F of the shot noise current in plane-parallel space-charge diodes was measured at a frequency of 2·24 Gc/s by an engineer at Berlin Technical University who developed special measurement equipment with a high sensitivity for these noise measurements. In the paper the transformation of the noise current from the cathode-anode path to the input of the measuring receiver is stated, and the measured values of the reduction factor were compared with numerical results from the transit time theory. The parameters were cathode current density and cathode anode spacing which could be varied during operation in one type of diode. The results from theory were also available as values depending on distance.

"Noise measurement in the region of transit time effects in plane-parallel diodes", G. Wittig. *Nachrichtentechnische Zeitschrift*, **16**, pp. 8–13, January 1963.

TELEVISION VIEWING STANDARDS

With due consideration of the effects in television from ambient lighting on the reproduction characteristic, on comfort in viewing and on general impression of the picture, it is shown in a German paper, that optimum viewing conditions are obtained when the screen is so dark that it can be illuminated from the front like a photograph. For this purpose a screen tinting with a factor of 1·2 is sufficient. In view of the relations explained here it is proposed to specify standards for the viewing conditions during critical judgements of television pictures, the standards covering screen tinting, light from the front and peak light density of the picture. Recommended guide values are a screen tinting of 1·2, a front light of 1·5 asb and a peak light density of 150 asb (1 asb = 0·0929 foot-lambert).

"Guide values for the adjustment of optimum conditions for viewing television pictures", H. Grosskopf. *Nachrichtentechnische Zeitschrift*, **16**, pp. 35–9, January 1963.

COLOUR TELEVISION PERCEPTION

In a recent Czech paper a description is given of the method and results gained in an inquiry into differential colour discrimination of the human vision, under television conditions. The work aims at ascertaining a practical criterion of permissible colour distortion of the television picture. The judgement of ten picked spectators, a set of ten colour specimen areas framed in the real picture and an experimental colour television set-up were used for the experimental investigation. The resulting practical measure of perceivable colour differences is shown in tables and graphically in a colour diagram in terms of colour coordinates differences. By comparing the results of the present investigation and of earlier work, an average differential colour discriminating ability of the human eye $ds = 20$ according to the MacAdam criterion results for television conditions.

"On the perceivable colour difference in television pictures", J. Pazderak. *Slaboproudy Obzor (Prague)*, **24**, No. 2, pp. 69–76, February 1963.

DIRECTIONAL AERIALS

Because with linear aerials used at present the directional effect is determined by their geometrical dimensions, German engineers are investigating the possibilities of controlling the properties of a directional aerial by inserting non-linear elements. For this purpose the aerial is subdivided into at least two parts, the outputs of which are combined after a non-linear conversion. Using an example of frequency multiplication it is shown theoretically and practically that reception patterns can be produced which correspond to those of linear aerials with apparently multiplied dipole spacings. The second part of the paper deals with effects resulting from a simultaneous incidence on the aerial system of two waves with frequencies within the reception range of the aerial. These waves may be coherent or incoherent. The mixing terms produced at the non-linear elements are also investigated.

"Non-linear directional aerial systems for controlling the radiation pattern in the case of reception", H. Meinke, H. Eisenmann and E. Hechenleitner. *Nachrichtentechnische Zeitschrift*, **16**, pp. 1–7, January 1963.

RADIO CARBON TECHNIQUES

The sciences of geology, archeology, ethnology, soils and climatology require accurate datum points on the time scale. Of the means available for this purpose radiocarbon (^{14}C) is the most suitable for periods up to 40,000 years ago. In systems using proportional counters and carbon dioxide gas, it is necessary to use high-gain non-overloading linear amplifiers, electronic pulse-stretching, delaying, sorting, blocking, discriminating and multi-channel recording. The instrumental features of the unit at the University of New South Wales are described in a recent Australian paper.

"Electronic instrumentation for radiocarbon dating", J. Bell, J. W. G. Neuhaus and J. H. Green. *Proceedings of the Institution of Radio Engineers Australia*, **23**, pp. 718–21, December 1962.