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."

THE RADIO AND ELECTRONIC ENGINEER

The Journal of the Institution of Electronic and Radio Engineers

VOLUME 35 No. 5 MAY 1968

The Commercial Engineer

ASSUMING that an industrial research project warrants commercial development there still remains the problem of attracting customers to the end-product. Unless sufficient customers are forthcoming the cost of research and development cannot be justified and the project is not commercially viable. No industrial organization could exist for very long under such conditions and its operations would come to an end without sufficient capital for further development.

If the last statement over-simplifies the chain of commercial operation, it is nevertheless true of most enterprises which translate science into something 'for the benefit of man'. To produce engineering equipment for a few customers would, in most cases, be financially prohibitive.

This reasoning emphasizes the importance of the technical sales engineer. But there is a further consideration; the sales force is an integral part of a commercial feedback circuit. Under the name of 'market research', sales effort is increasingly directed towards ascertaining customer-need. Where a need is sufficiently established, and a potential customer may not at first be clearly aware of it, there is a demand for production which, in turn, requires development—often based upon applied research. Wider application of the demand determines, financially, whether the customer-need can justify production on a long or short-term basis, and this decides the ultimate cost to the consumer.

For too long, engineering institutions have tended to consider membership only in terms of engagement in research, development and, in more recent years, in production techniques. Too little attention has been given to the absolute need, economic and otherwise, to recognize the technical and financial contribution of the engineer who specializes in marketing—more familiarly known as a sales engineer.

In order to do his job satisfactorily, such a man needs to acquire academic knowledge and standards equivalent to that required for Graduateship of an engineering institution so that he may talk to his colleagues in research, development and production, not only to understand the products which he is marketing, but to feed back to them information on utilization problems, refinements, or on entirely new products to meet new needs. To the customer, the modern technical sales engineer is not just someone who wants to sell something irrespective of its fitness for the task in hand: he is often a *trusted ally* who can place specialist knowledge and advice at their disposal.

In a closely-knit economic unit, the marketing executive is as important as the research director or production manager. If such a man proves that he is a professional engineer, and is capable of demonstrating his professionalism in the terms of the E.U.S.E.C. definition of an engineer, then should not the Institutions encourage the attainment of professional status and the accolade 'Chartered Engineer' for those who are, by dint of their efforts and application of technical knowledge and know-how, able to produce the economic environment which makes possible, and pays for, research and development?

G. D. C.

INSTITUTION NOTICES

Election of Honorary Fellows

The Institution's 1968 Convention at Cambridge will be the occasion on which citations of Honorary Fellowship of the Institution will be presented to two Past-Presidents who will then sign the Roll of Honorary Fellows. The bye-laws of the Institution permit the election of one Honorary Fellow in each year.

For 1967 the Council conferred this honour on Mr. W. E. Miller, M.A., C.Eng., F.I.E.R.E.; Mr. Miller was President in 1953 and 1954, and his membership of the Institution dates from 1932. He recently retired from the Managing Directorship of Iliffe Electrical Publications Ltd.

Mr. Leslie H. Bedford, C.B.E., M.A., B.Sc.(Eng.), F.C.G.I., C.Eng., F.I.E.R.E. was elected Honorary Fellow earlier this year. He served as President in 1949 and 1950 and he is at present Director of Engineering in the Guided Weapons Division of the British Aircraft Corporation (Operating) Ltd.

A copy of the citation for each recipient of this highest recognition which the Institution can bestow is included in the illuminated Roll of Honorary Fellows kept in the Lecture Room at 9 Bedford Square.

Symposium on Electronic Weighing

The newly-formed Specialized Group on Instrumentation and Control is to open its first full session of meetings with a two-day Symposium on Electronic Weighing. This will be held on Wednesday and Thursday, 30th and 31st October, 1968, at the Middlesex Hospital Medical School, London, W.1.

Subjects to be covered will be the following:

> Weight sensors; static weighing; dynamic weighing; special applications; telemetry.

Already over a dozen papers are under consideration for inclusion in the final programme, and further offers are invited. Information on the detailed scope may be obtained by intending authors from the Secretary, Instrumentation and Control Group Committee, I.E.R.E., 9 Bedford Square, London, W.C.1. Requests for registration forms should be sent to the same address but it should be noted that these will *not* be available until September.

Conference on Electronic Switching and Logic Circuit Design

A one-day Conference will be held on Thursday, 24th October next, at the College of Technology, Letchworth, on 'Electronic Switching and Logic Circuit Design'. This is the third similar venture by the College which the Institution has supported, and offers of papers for consideration for inclusion in the programme are invited. These should, in the first instance, be sent to the Chairman of the Joint Organizing Committee, Mr. K. J. Dean, M.Sc., C.Eng., F.I.E.R.E., Department of Science and Electrical Engineering, College of Technology, Letchworth, Hertfordshire, and should be accompanied by a synopsis. As it is intended that all papers should be refereed prior to pre-printing, early submission of synopses is requested.

Conference on Computer Aided Design

The Institution of Electrical Engineers and the I.E.R.E., with the collaboration of the University of Southampton, are arranging a Conference on Computer Aided Design to be held at Southampton University from 15th to 18th April 1969. This will be held under the aegis of the United Kingdom Automation Council.

Further information and a call for papers notice will be available in due course from the I.E.R.E. or from the Conference Department, I.E.E., Savoy Place, London, W.C.2.

Australian National Radio and Electronics Engineering Convention

The biennial Convention of the Institution of Radio and Electronics Engineers Australia will be held in Sydney from Monday, 19th to Friday 23rd May, 1969.

The following technical areas will comprise the Convention programme:

> Basic sciences and techniques; industrial electronics; communications; electronic systems; computers and data processing; instrumentation; materials, components and production processes; biomedical electronics; and professional activities.

Contributions of technical papers falling within these areas are invited. Abstracts of each paper in about 1000 words, with diagrams, will be published prior to the Convention and complete papers will be published in post-Convention issues of the *Proceedings of the I.R.E.E. Australia.*

Further information is available from the General Secretary, Institution of Radio and Electronics Engineers Australia, Box 3120, G.P.O., Sydney, N.S.W. 2001. Summaries of proposed contributions should be submitted without delay. Completed manuscripts are to be submitted by 1st December 1968.

U.D.C.681.32:37/9

Computer Aided Layout of Microcircuits

By

W. J. CULLYER, Ph.D., B.Sc.,†

SYLVIA TUBBS†

AND

A. P. STOCKTON†

Summary: To eliminate some of the skilled human effort in arranging the layout of the components of a microcircuit on a substrate, it is proposed that the process should be aided by a digital computer. A fully automated system does not seem possible at present. The method which has been suggested, and partially developed, relies on the automatic production of a rough layout which is unconstrained in size. This is then rearranged by human intervention, to a suitable electrical and geometrical form, the designer communicating with the machine using a simple drawing language and viewing the progress of the work using an on-line graphical display. The process does not permit changes in topography and this still remains a major problem.

1. Introduction

Once the electrical design of a circuit has been settled, its reduction to microminiature form involves the layout and interconnection of components on a substrate. This step and the production of the corresponding art-work are lengthy and expensive processes. Thin film circuits can take several man-weeks of effort and the more elaborate silicon integrated circuits (s.i.c.) may require many man-months of work.

The risk of human error during these processes rises proportionally as circuits become more complex. So far, little has been done in this country to remove this bottle-neck and source of error from the production of microcircuits. Many of the stages involved can be described in precise mathematical terms and can be performed quickly and accurately by a digital computer. This paper presents a detailed proposal for programming a computer to do this and it describes the results so far obtained.

Steps such as the calculation of the mechanical dimensions of components can be programmed easily but it is the process of arranging these elements on the substrate which is difficult as it calls for both visual perception and imagination. These talents cannot be created easily in a computer program. For this reason, the discussion is confined to *computer-aided* layout rather than a fully automated system. Even if increasing interest is assumed, it will probably be some years before the layout of substrates without human intervention is achieved.

The philosophy advocated here and outlined in Fig. 1, is to form automatically a rough layout which is unconstrained in size and then to modify this with the aid of a computer. Any attempt to constrain the size of the trial layout results in conflicts which at present need human intervention to resolve them. When modifying this trial layout the designer communicates his experience and imagination to the

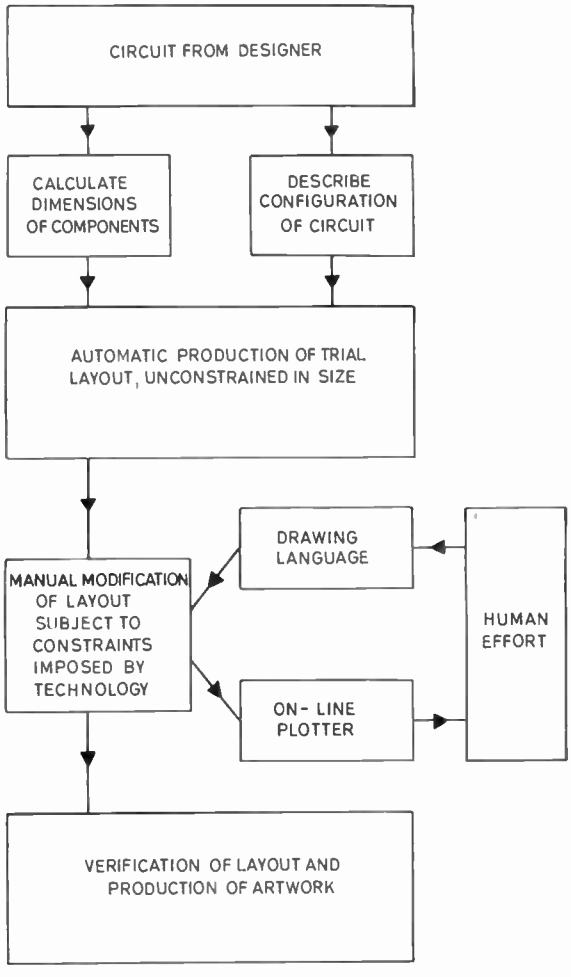


Fig. 1. Proposed process for computer-aided layout of microcircuits.

† Radio Division, Signals Research and Development Establishment, Ministry of Technology, Christchurch, Hants.

computer using a simple drawing language. A control typewriter or teleprinter forms a suitable input device and the progress of the work is viewed using an on-line plotter. During this phase of the work the designer does not have to worry about the underlying co-ordinate geometry of the situation or the constraints imposed by the particular technology. These facts form part of the program and prevent him from making impermissible moves.

When the final layout is decided upon the computer can verify that it is electrically correct before proceeding to produce detailed drawings. To eliminate any possible errors it is preferable that the drawings should be produced on-line. It will probably be two years before suitable peripheral equipment is available commercially. In the intervening period, off-line operation will be adopted.

The paper is concerned mainly with thin film circuits.

2. Outline of Steps in Proposed Process

2.1. *Data Required*

Initially, the object is to ask the designer for the smallest amount of information which is consistent with an unambiguous description of the circuit.

2.1.1. Constraints imposed by the technology

This information is assembled by the group responsible for fabrication and kept up to date as modifications are made to the technology. Examples of the data required are the sheet resistivity of resistive films, the minimum line-width that can be produced and the configuration of transistors. This body of information is not the direct concern of the designer of an individual circuit as it can be retained and used when necessary in the programs.

2.1.2. Values of components

The resistance and power dissipation of each resistor and the value of each capacitor must be given. With the increasing use of computer-aided design it is possible that these figures will have in turn been produced by design and analysis programs.[1]

2.1.3. Electrical structure

If the structure of a microcircuit is planar, the components can be arranged on a substrate without introducing any crossovers between conducting materials. It is unusual to find this and most practical microcircuits contain capacitors which are specified in the designer's circuit where they can be employed as crossovers.

Despite an outstanding recent paper by Sinden[2] on the topology of active R-C networks, there does not appear to be enough information to write a program which permits such changes in topography. A solution to this problem has recently been described in a paper by Nicholson.[3] The designer is forced to specify the relative positions of the components on the substrate and to indicate the approximate routes for interconnections.

A suitable way of describing the electrical structure is to draw the circuit on a grid (see Fig. 2), so that it may be defined in simple numerical terms. In the example of Fig. 2, the capacitor C3 is used as a crossover and this decision is not altered at any later stage in the process. The corresponding drawing as shown in Fig. 3 will be referred to as an 'alpha-numeric array' and it is clear by inspection that it can form the basis both for the initial placement of parts and also for deducing the precise interconnections within the circuit.

2.2. *Formation of Trial Layout*

Initially, the positions of the parts on the substrate are based on the grid defined when describing the electrical structure shown in Fig. 2. Sufficient spacing has to be allowed between successive rows and columns to accommodate the components.

These components can then be interconnected in the manner indicated in the alpha-numeric array. Experience has shown that the overall size of the complete layout at this stage will be two- or three-times larger in linear dimensions than the final arrangement.

2.3. *Modification of Trial Layout*

In modifying this trial layout the computer acts as a store for the parts of the puzzle and provides the means for moving them to positions decided by the eye. There is no attempt to produce an optimum solution and at present the extent to which the computer can automatically pack the components into a required area is an open question.

The process discussed here is applicable to computers equipped with a digital incremental plotter and an on-line control typewriter or teleprinter. The input must be in a form which can be understood easily by the user. A drawing language consisting of familiar words such as MOVE, TURN and GROW fulfils this need.

Using the digital plotter, communication with the designer can be achieved very accurately. This method has proved to be adequate for the development of the techniques described in this paper. A cathode-ray tube and light-pen interface could be used to provide the required communication and this would be appreciably faster.

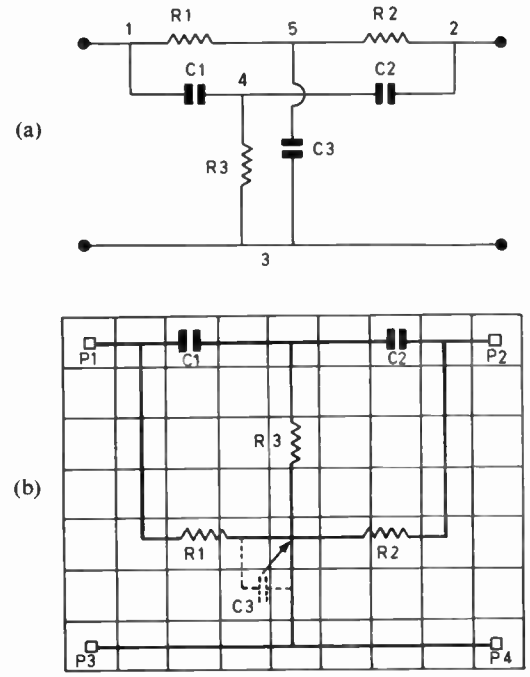


Fig. 2. Description of electrical structure.

2.4. *Practical Realization of Microcircuit*

Once a final layout has been decided upon, all of the information for the production of the microcircuit is held within the computer. It is then necessary to produce a series of masks to define the regions in which materials will be either diffused or evaporated to form the components. This problem will be discussed briefly in Section 6.

3. Numerical Description of Designer's Circuit

The circuit, shown in Fig. 2(b), has to be converted into a suitable form of input for the computer. Two possible methods are either to define the branches of the network by numbering the nodes or to define the drawing directly using a matrix. For the reasons given below, the latter approach is favoured.

3.1. *Numbering of Nodes*

By numbering the nodes of the network of Fig. 2(a) as shown, the circuit is defined electrically by the following data:

Component	Node A	Node B
R1	1	5
R2	5	2
R3	4	3
C1	1	4
C2	4	2
C3	5	3

P1	–	C1	–	–	–	C2	–	P2
	I			I			I	
	I			R3			I	
	I			I			I	
	–	R1	–	J3	–	R2	–	
				I				
P3	–	–	–	–	–	–	–	P4

Fig. 3. Matrix of alpha-numeric characters representing Fig. 2(b).

Transistors require three entries to fix them in the network.

This method is used widely for entering data into programs for circuit analysis but does not provide an adequate description for the purposes of layout. The major problem is that the precise positions of the crossovers are not defined. This in turn means that it is difficult to determine an acceptable set of paths for the interconnections.

3.2. *Use of Drawings Formed by Alpha-numeric Characters*

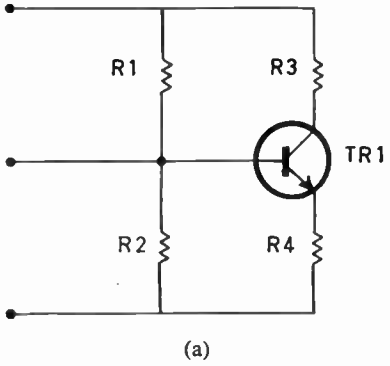
It is far better if the network is defined by an array which shows the exact placing of the crossovers and gives a close indication of the pattern of the conductors. Figure 3 is an example of such an array and comparison with Fig. 2(b) shows how this array of alpha-numeric characters has the appearance of the original circuit. Experience has shown that this method is easy to check. Within the program which creates the trial layout these alpha-numeric data is decoded into a matrix in the manner discussed in Section 4.2.

The alpha-numeric array is a complete mathematical description of the structure of the circuit, providing that it has been set up in accordance with the following rules:

(a) Each component, including solder pads, must occupy at least one cell of the array.

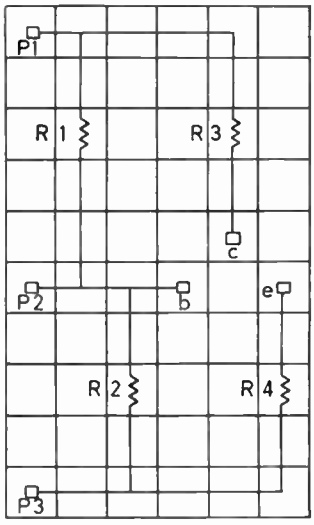
(b) Components which are drawn parallel to the *Y*-axis of the circuit diagram must have pieces of conductor in this orientation associated with them in adjacent cells in the same *column*, e.g. R3 and the cells above and below it in Fig. 3.

(c) Components which are drawn parallel to the *X*-axis of the circuit diagram must have pieces of conductor in this orientation associated with them in adjacent cells in the same *row*, e.g. C1 and the cells to the left and right of it in Fig. 3.



(a)

Fig. 4. Representation of circuit including a transistor (thin-film technology).



(b)

P1	—	—	—	—	
	I			I	
	R1			R3	
	I			I	
	I			TR1C	
P2	—	—	TR1B		TR1E
		I			I
		R2			R4
		I			I
P3	—	—	—	—	—

(c)

(d) Components and conductors which are in adjacent cells must be parallel, unless they are in electrical contact.

The specific notation which has been employed in the present work uses the following characters:

— (minus)	Horizontal conductor
I	Vertical conductor
P*n*	Solder pad number *n*, where *n* is an integer
R*n*	Resistor number *n*
C*n*	Capacitor number *n*
J*n*	Crossover in which capacitor number *n* is realized
T*n*E	Emitter of transistor *n*
T*n*B	Base of transistor *n*
T*n*C	Collector of transistor *n*
D*n*A	Anode of *n*th diode
D*n*K	Cathode of *n*th diode

The pads for transistors are grouped as shown in Fig. 4.

3.3. *Values of Components*

Tables of the values of the resistors and capacitors form part of the input data. The trial layout is formed by correlating this information with the electrical structure of the network, as indicated in Fig. 1.

4. Formation of Trial Layout

The problem is to produce a trial layout automatically from an array of the type shown in Figs. 3 and 4(c), together with the tables of the values of components. The only constraint on this layout is that it must be bounded on two adjacent sides as shown in Fig. 5.

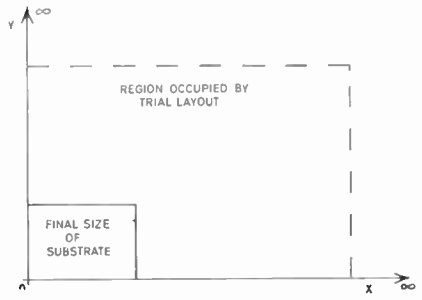


Fig. 5. Co-ordinate geometry of layout.

4.1. *Storage of Shapes*

Arithmetic using integers is carried out much faster in a computer than operations with floating-point numbers. Therefore, it is better to store the positions of the shapes as integers rather than retain dimensions in either inches or millimetres. The shapes can be defined either by listing the integral co-ordinates of their corners or by entries in a matrix. In both cases, the basic size of the mesh is equal to the minimum line-width which can be produced by the technology under consideration.

If a matrix is used, its size is governed by the physical size of the layout, and may be very large. When using the co-ordinates of corners, the amount of store is set by the number of pieces in the puzzle, not by the size of the picture they form. The programs

which have been written use arrays which hold the co-ordinates of corners as the primary method of storage. About 3000 locations are needed to store 100 shapes each having up to 30 corners, when these co-ordinates are packed two to a word.

The description of the shapes can be translated into blocks of entries in a matrix as a temporary measure to enable interconnections to be made.[2] The completed conductor is then translated back into the store containing co-ordinates.

Translation between the two methods of storage is straightforward.

4.2. *Decoding the Alpha-numeric Array*

Each cell of the alpha-numeric array is inspected and a corresponding code number is written into the descriptive numerical array, L. To distinguish between components which must be parallel to the X- and Y-axes, respectively, the 'cluster' of elements surrounding each resistor and capacitor is examined. A cluster is defined as the elements described in Fig. 6. It follows from the rules listed in Section 3.2 that the orientation of the component is specified unambiguously.

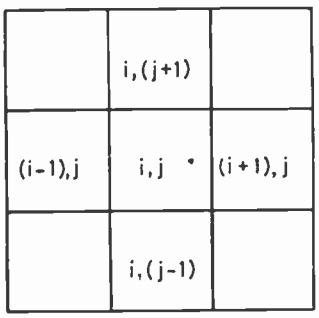


Fig. 6. Elements of a matrix which form a cluster.

Without going into the details of coding a typical array has the form of eqn. (1).

$$L = \begin{matrix} 601 & 999 & 999 & 999 & 999 & 0 \\ 0 & -999 & 0 & 0 & -999 & 0 \\ 0 & -101 & 0 & 0 & -103 & 0 \\ 0 & -999 & 0 & 0 & -999 & 604 \\ 0 & -999 & 0 & 0 & 411 & 0 \\ 602 & 999 & 999 & 412 & 0 & 413 \\ 0 & 0 & -999 & 0 & 0 & -999 \\ 0 & 0 & -102 & 0 & 0 & -104 \\ 0 & 0 & -999 & 0 & 0 & -999 \\ 603 & 0 & 999 & 999 & 999 & 999 \end{matrix} \quad \ldots\ldots(1)$$

The resistors, capacitors and pads for soldered connections can be placed initially by reference to eqn. (1). Also, the nodal description of the circuit, as discussed in Section 3.1, can be extracted by 'walking' along the conductors. The conductors are represented in eqn. (1) by the entries 999 and −999. Once the nodal description of the circuit has been established it is used not only to generate the interconnections initially but also as a reference when conductors have to be reformed at a later stage.

Drawings of the type shown in Figs. 3 and 4(c) can be used to provide data for programs employing nodal analysis, as well as giving a necessary and sufficient description of the circuit for the purposes described in this paper. The analysis of the circuit and formation of a trial layout can then proceed without recourse to numbering the nodes.

4.3. *Calculation of Dimensions of Components*

At this point, the constraints imposed by a particular technology must be introduced. Some of the relevant factors for thin film circuits are listed in Table 1 together with the numerical values used at S.R.D.E.

Table 1

Constraints imposed by technology

Factor	Numerical value used at S.R.D.E.
Sheet resistivity of resistive film	200 ohms/sq.
Minimum line width λ	0·01 in
Capacitance of a square capacitor of side λ	6·25 pF
Minimum spacing between components	0·02 in
Overlap between film making electrical contact	0·01 in
Permissible dissipation of resistive film/sq. of side λ	0·2 mW/10^{-4} in^2

When dealing with silicon integrated circuits the corresponding details would be much more extensive.

The precise rules for designing resistors and capacitors, which are to be fabricated by the defined technology, have to be embodied as procedures written by individual groups. There are so many empirical factors in the rules laid down by particular teams of microcircuit engineers that it is impossible to propose a general system. To follow the process described in this paper it is sufficient to note that the area required for an individual component can be calculated easily although a choice still has to be made about the aspect ratio of the region. For example, by using more meanders in a resistor it can be constrained to a shorter, broader region that it would otherwise occupy if it consisted of a straight strip of conducting film.

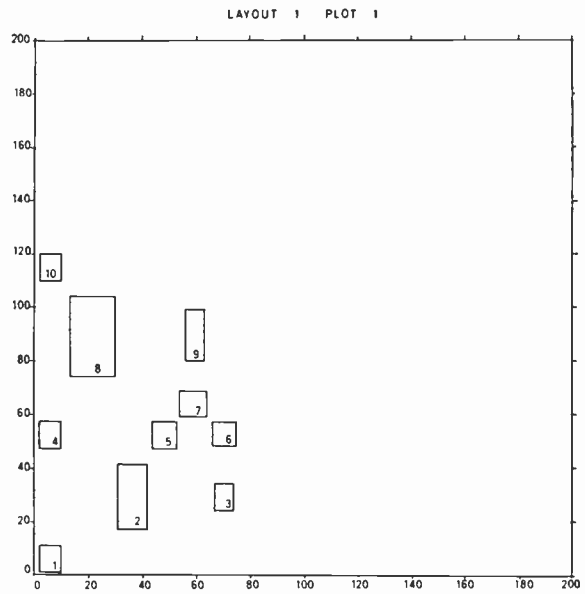


Fig. 7. Positions of rectangles bounding components of single-stage amplifier.

4.4. *Placing the Rectangles Bounding the Components*

The mechanical dimensions of the transistors, resistors, capacitors and pads, and the electrical structure of the circuit are combined to generate the trial layout. There is no point in handling detailed descriptions of these components at this stage. Instead, rectangles which are just large enough to contain these elements are manœuvred in the plane containing the substrate. In this context a pad used for a soldered connection is regarded as an element.

Referring to the numerical array in eqn. (1), it is possible to lay down a grid which represents the centre lines of the components in the corresponding rows and columns of the array. The centres of gravity of the bounding rectangles all lie on intersections of this grid. This process is illustrated by considering the amplifier shown in Fig. 4. Using the thin film techniques developed at S.R.D.E., defined by the parameters given in Table 1, the bounding rectangles take up the positions shown in Fig. 7. The detailed drawings of the components can be added inside these rectangles, once a degree of packing has been achieved.

4.5. *Interconnection of Components*

The data on which the interconnections are based are presented as a list of the nodes between which components are connected. This is derived from the alpha-numeric array defining the structure of the circuit, i.e. eqn. (1). As noted in Section 4.1, conductors are generated in a series of arrays in which the positions of the components are represented by blocks of entries. The basic technique was developed by Lee.[4] In essence, the process seeks to find the shortest path between a 'home' and a 'target' cell in a maze in which the obstacles are all of the other components.

The chief difficulty is the size of the array which is needed. Even for simple circuits the array is prohibitively large if the whole layout is represented at once. Attempts to interconnect a particular node in a single array also suffer from this defect when dealing with complex earth and supply lines.

The solution is to generate a large number of arrays sequentially. Each is just large enough either to interconnect two components or to join a component to an existing conductor. The new piece of conductor formed is placed in the store containing the co-ordinates of the corners of parts. A new array is established to determine the path for the next piece of conductor, even though this may belong to the same node. The separate pieces of a conductor can be combined as a later step.

By this means the components can be interconnected and a complete trial layout results. This is then refined by the process of man/machine communication described in Section 5.

5. Manual Modification of Trial Layout

5.1. *Nature of Program*

The modification of the trial layout is performed by the eye, the designer using a drawing language to communicate with the computer and viewing the progress of the work on a digital plotter. To implement this system a program has been developed in ALGOL on the Elliott *4100* series computer operated by S.R.D.E. This machine has 24,000 words of store and is equipped with a digital plotter which will produce drawings 10 inches wide.

The sequence of operations used when refining a trial layout is as follows:

(a) After studying a drawing of the trial layout, the designer decides on the initial moves required to pack the components into a smaller area.

(b) A description of the trial layout is read-in from paper-tape. This tape also includes several 'basic shapes' which do not form part of the actual layout but are used when generating new parts.

(c) The instructions needed for each move decided upon in (a) are punched on the paper-tape in terms of the drawing language defined in Section 5.5. These orders are read and on completion of each block of instructions a message is printed on the control typewriter. The result of this series of moves is displayed on the plotter.

(d) Control can now be transferred to the on-line typewriter. Using this as the input device and looking at the drawing on the plotter, the designer is able to make further improvements. As in (c), these instructions consist of words in the drawing language.

(e) It may become necessary to give further thought to the problem away from the computer. For this reason, provision has been made for filing the current state of the layout on paper-tape. The last drawing produced, together with this tape, enables the work to be continued from step (b) at a later date.

5.2. *Requirements for the Drawing Language*

Since a wholly numerical form of input would be difficult to learn and apply, a drawing language, used to command the program, is essential. The stress of implementing a system of man/machine communication on a high-speed computer can be very considerable.[5] Every means must be used to aid the designer's concentration on the problem.

It has been found that the following types of instruction are needed:

(a) Orders which are based on standard co-ordinate geometry. These are applied to single parts of the layout and fundamentally include the operations of rotation, translation, magnification and reflection.

(b) Executive instructions, to join simple parts to form complex shapes and also to complete the block structure of the language.

(c) Instructions which enable parts to be deleted from the picture and also allow for the correction of the errors which arise when using manual input.

(d) Control instructions for the plotter and output punch.

The words in the language should be simple and distinct, as well as self-explanatory. A program must be written to recognize these words in the context of a defined block structure.

Table 2

Drawing language

Instruction	Meaning
IMAGE, *I*	Reflection about *X*-axis when $I = 1$ and about *Y*-axis when $I = 2$
GROW, *I*, *J*	Magnification *I* times in *X*-direction and *J* times in *Y*-direction
TURN, *I*	Rotation anticlockwise by 90° when $I = 1$ by 180° when $I = 2$ by 270° when $I = 3$
MOVE, *I*, *J*	Translation by *I* in *X*-direction and by *J* in *Y*-direction
DO,	Create new shape from list of instructions and give it the next available number
MERGE, *M*	Create new shape from list of instructions and merge it with shape *M*
VIEW, *NP*, *NA*, A, B. . . .	Plot the shapes A, B. . . . The plot is on new axes if *NP* is *true* or on old axes if *NP* is *false* with a false origin in *NA* is *false* or a true origin if *NA* is *true*
DRAW, *NP*, *NA*, *I*	Plot the shapes after and including shape *I*
SCRAP,	Remove all statements since previous MERGE or DO statement
OMIT, *I*	Remove shape *I* from store
FILE,	Output a copy of all the shapes in store

5.3. *Structure of the Language*

Using a block of orders it is possible to carry out a complex series of operations on a single part. Each part is identified by a number. The block structure of the language takes the form:

*n — The number of the part to be operated on

INSTRUCTION A
INSTRUCTION B
. . . . . } List of orders
. . . . . .
EXECUTIVE WORD

Whilst instructions are being typed or read within such a block none of the operations are carried out. It is only when the block is terminated with the executive word that the list of orders is obeyed. This system gives the designer the chance to modify his thoughts.

The complete language which has been devised and used is specified in Table 2. The first four instructions may appear within a block between *n and either MERGE or DO but not elsewhere. Inspection of Table 2 shows how the requirements of Section 5.2 have been satisfied.

Once the instructions within a block have been obeyed, the part may either be left isolated using the executive word DO or connected to an existing piece using MERGE. To 'merge', in the sense used in this paper, is to create a single rectilinear shape from two other such shapes. These must touch along at least part of one edge and must be wholly exterior to each other. Examples of legal and illegal merging are shown in Fig. 8. The procedure used for merging parts must cope with re-entrant shapes, as shown in the fourth example of this illustration.

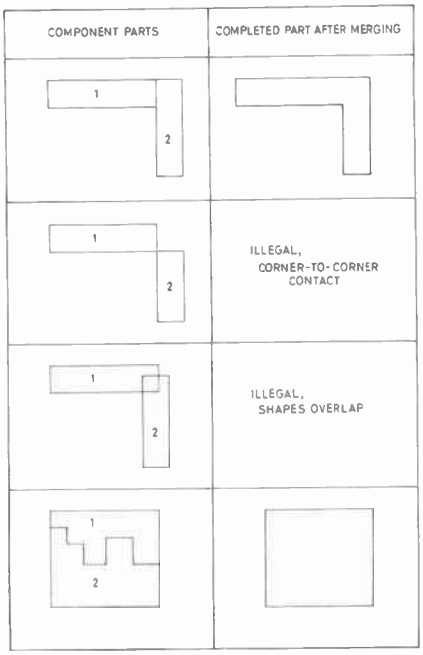


Fig. 8. Legal and illegal merges.

5.4. *Detailed Description of Drawing Language*

5.4.1. Geometric procedures

IMAGE, GROW, TURN and MOVE call upon procedures which deal with the basic co-ordinate geometry of the layout. They all operate with parameters which define the extent of the operation, as listed in Table 2.

Since reflection and rotation leave parts in the wrong quadrant and magnification involves unwanted translation, IMAGE, TURN and GROW are more complicated procedures than those used in conventional co-ordinate geometry. In the program which has been developed, the part is returned to the origin after any one of these procedures has been called. This is illustrated in Fig. 9 for the case of IMAGE. After one of these procedures has been used the shape is returned to its correct place in the layout using MOVE.

5.4.2. Executive instructions DO and MERGE

DO is a dummy statement, in the sense that it contributes nothing to the rearrangement of the layout. It is simply a means of completing a block and indicating that the list of instructions relating to the last part number should be obeyed.

The other executive instruction, MERGE, has the same function as DO but also commands the union of two shapes, as discussed in Section 5.3. It forms the heart of the language as it is the means by which the complex shapes encountered in microcircuits are created.

5.4.3. OMIT and SCRAP

The instruction 'OMIT, *I*' causes the shape numbered *I* to be deleted from all future drawings. This shape cannot be recalled and any attempt to operate on it causes the message 'NO SHAPE = *I*' to be printed. This is an important provision because at present there is no means of breaking a part and reducing its complexity. To achieve this, a new part is created and the instruction OMIT is used to remove the old version.

Within a block the designer may decide may decide that the steps he has taken are inadvisable. SCRAP is used to remove unwanted instructions back to and including '**n*' at the head of the block.

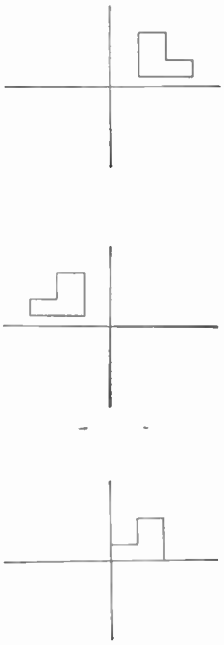


Fig. 9. Use of the procedure image.

5.4.4. Drawing procedures DRAW and VIEW

Procedures are required which enable layouts to be drawn on the plotter. As only straight lines are drawn, the outlines of the shapes can be produced with little difficulty. The far more challenging problems are the numbering of the parts for future reference and the organization of the scale factors.

The form of output is shown in Figs. 7 and 11. This is a reproduction of a sheet from the plotter and shows the practical layout of the circuit of Fig. 4 when realized using the S.R.D.E. thin-film technology specified in Section 4.2.

The designer uses the display in two slightly different modes:

(a) The command DRAW is used to see the complete layout, that is, all of the shapes in the store with the exception of the basic building blocks.

(b) To examine a small region of the drawing in more detail, or to add a shape to an existing drawing, the word VIEW is employed.

In either mode, provision is made for adopting a false origin and calling for a new drawing when required.

5.4.5. FILE

To enable a permanent record to be kept of the various stages of the layout, the co-ordinates of the corners of the shapes can be punched on paper tape by the command FILE. This tape is compatible with the read instructions at the beginning of the program and therefore work can be stopped and then restarted at a later date, as described in Section 5.2.

Another advantage of using FILE is that shapes which have been omitted are not copied on to the tape and therefore unwanted parts are not included when the program is run again.

5.5. *Example of the Use of the Drawing Language*

The program which forms the trial layout has been developed independently of the work on manual modification. At present the results obtained do not provide a complete implementation of the system discussed in Section 2. To illustrate the use of the drawing language details are given in Section 5.5.2 of the construction, from first principles, of part of the layout of the circuit shown in Fig. 4.

5.5.1 Application of basic shapes

To increase the efficiency of operation, basic shapes are provided as part of the data to the program. These shapes are needed in any layout and it is wasteful to generate them repeatedly.

The seven basic shapes at present used at S.R.D.E. are shown in Fig. 10 and the reasons for their existence are described in Table 3.

Table 3

	Basic Shape	Description
1	Unit square	This is the only basic shape that is a requirement of the language. It is the shape from which all rectilinear pieces can be constructed.
2, 3		These two rectangles are used extensively in the formation of meandered resistors.
4	Base pad	This is a pad to which the base contacts of a transistor can be bonded. The notch serves as identification.
5	Emitter pad	Similar to shape 4.
6	Collector pad	Similar to shape 4.
7	Contact pad	This is a square to which external connections can be bonded.

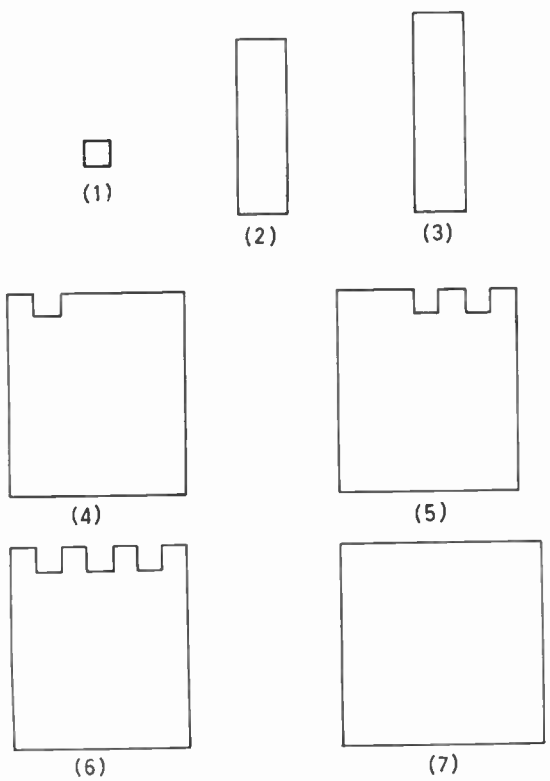


Fig. 10. Basic shapes.

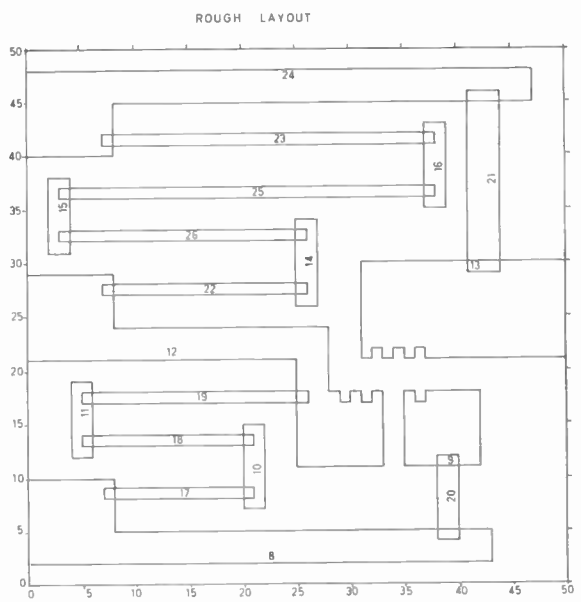


Fig. 11. Layout of circuit of Fig. 4(c).

These basic shapes are stored with their bottom left hand corners at the origin.

5.5.2. Detailed instructions to form part of the layout of the circuit of Fig. 4

If the circuit of Fig. 4 is realized using the S.R.D.E. thin film technology then a possible layout is shown in Fig. 11.

The construction from basic shapes of the conductors which link the transistor to the rest of the circuit could be achieved by the following moves. This process can be followed more easily by reference to Fig. 12.

*7	MOVE,	0	19	Basic shape 7 is moved to form
	DO,			the input pad of the circuit.
		(Part 16)		For reference it might be given
				the number 16.
*1	GROW,	17	3	A rectangular piece of conduc-
	MOVE,	8	19	tor is formed and merged with
	MERGE,	16		the input pad. The complete
		(Part 16)		shape retains the number 16.
*1	GROW,	3	5	Another piece of conductor is
	MOVE,	25	17	added to the configuration just
	MERGE,	16		produced.
		(Part 16)		
*5	MOVE,	25	9	The transistor base pad is now
	MERGE,	16		added to complete this conduc-
		(Part 16)		tor area.
*6	TURN,	2		The output conductor is formed
	MOVE,	30	20	by turning the collector pad,
	DO,			basic shape 6, through two
		(Part 17)		right angles and moving it into
*7	MOVE,	37	20	position. The output pad is
	MERGE,	17		moved and merged with the
	DO,			collector pad.
		(Part 17)		
*4	MOVE,	35	9	The emitter pad is moved into
	DO,			position and given the num-
		(Part 18)		ber 18.

6. Production of Master Drawings

Once the final layout has been produced, the computer holds all the information necessary to derive drawings of the individual masks. Each mask represents a separate layer of the microcircuit. It would be preferable if on-line equipment could be used and conventional art-work eliminated. This type of equipment is not likely to become available for some time on a commercial basis.[6,7]

As an interim measure, it is proposed that off-line operation is adopted. The basic information conveyed by means of paper-tape will be the co-ordinates of the

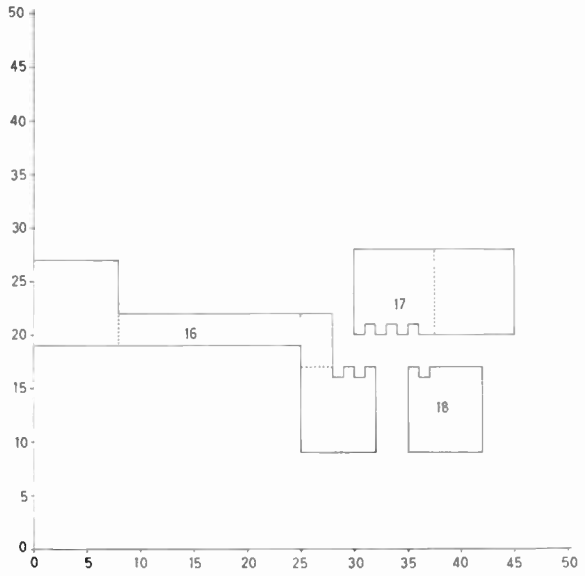


Fig. 12. Details of some of the conductors of Fig. 11.

corners of each part. In preparing this tape it is necessary to carry out a series of checks:

(a) Verification of the layout should be achieved by examining the interconnections to each part and checking these against the designer's original circuit.

(b) If metal-foil masks have to be produced for an out-of-contact technology, the mechanical design of each must be satisfactory. A mask which is weak in some areas will buckle in use and misalignment will occur between the various films. It is a challenging problem to devise a necessary and sufficient set of rules to check this point.

Failure to achieve a satisfactory answer in (b) means further manual modifications of the layout.

7. Conclusions

Enough progress has been made in the development of computer programs to aid the layout of microcircuits to suggest that these techniques can be extended towards the point where a largely automatic process is possible. At present, a high degree of manual intervention is involved, both to define the original electrical structure of the circuit and to produce the final layout.

8. Acknowledgments

Throughout, the investigation, the overall guidance of Mr. E. Fitch of the Applied Mathematics Group, S.R.D.E., has been of great help to the authors. Detailed assistance with the programming has been

given by Mr. P. Roberts, Mr. J. J. Morton and Mr. R. W. Spragg. Many of the rules for the design of film components and the layout of circuits produced by the S.R.D.E. thin-film facility were devised by Mr. D. G. Hodgson.

The structure of the drawing language was suggested to the authors by Mr. J. Wood of the Royal Radar Establishment.

Published by permission of the Controller, Her Majesty's Stationery Office.

9. References

1. A. Spitalny, 'Computer-aided design, Part 2: The computer excels as an analyst', *Electronics*, **39**, No. 24, p. 68, 28th November 1966.
2. F. W. Sinden, 'Topology of thin film RC circuits', *Bell System Tech. J.*, **45**, No. 9, p. 1639, November 1966.
3. T. A. J. Nicholson, 'Permutation procedure for minimizing the number of crossings in a network', *Proc. Instn Elect. Engrs*, **115**, No. 1, p. 21, January 1968.
4. M. L. Dertouzos, '*CIRCAL*: on-line circuit design', *Proc. Inst. Elect. Electronics Engrs*, **55**, No. 5, p. 637, May 1967.
5. C. Y. Lee, 'An algorithm for path connections and its applications', *Trans. Inst. Radio Engrs on Electronic Computers*, **EC-10**, No. 3, p. 346, September 1961.
6. J. S. Koford, P. R. Strickland, G. A. S. Sporzynski and E. M. Hubacher, 'Using a graphic artwork for manufacturing hybrid integrated circuits', Proc. A.F.I.P.S., Fall Joint Computer Conference, 1966.
7. N. E. Wiseman, 'Some applications of computers in electronics design', *The Radio and Electronic Engineer*, **34**, No. 4, p. 217, October 1967.

Manuscript first received by the Institution on 6th September 1967 and in final form on 20th February 1968 (Paper No. 1185/C103).

STANDARD FREQUENCY TRANSMISSIONS

(Communication from the National Physical Laboratory)

Deviations, in parts in 10^{10}, from nominal frequency for **April 1968**

April 1968	24-hour mean centred on 0300 U.T.			April 1968	24-hour mean centred on 0300 U.T.		
	GBR 16 kHz	MSF 60 kHz	Droitwich 200 kHz		GBR 16 kHz	MSF 60 kHz	Droitwich 200 kHz
1	—	—	+ 0·1	16	− 300·0	+ 0·1	+ 0·1
2	—	—	0	17	—	—	0
3	− 300·1	+ 0·1	+ 0·1	18	—	—	+ 0·1
4	− 299·9	0	0	19	− 300·0	0	+ 0·1
5	− 299·9	+ 0·1	+ 0·1	20	− 299·9	+ 0·2	0
6	− 299·9	− 0·1	0	21	− 299·9	+ 0·1	+ 0·1
7	− 299·9	+ 0·1	+ 0·1	22	− 299·8	+ 0·3	+ 0·1
8	− 300 0	0	+ 0·2	23	− 299·8	0	+ 0·1
9	− 299 9	0	0	24	—	+ 0·1	+ 0·1
10	− 299·9	+ 0·1	+ 0·2	25	− 300·1	+ 0·1	+ 0·1
11	− 300·1	+ 0·1	+ 0·2	26	− 299·8	+ 0·1	+ 0·1
12	− 300·0	0	+ 0·1	27	− 300·0	+ 0·2	+ 0·1
13	− 300·0	0	+ 0·1	28	− 299·9	0	0
14	− 299·9	0	0	29	− 300·0	+ 0·1	0
15	− 300·0	0	+ 0·1	30	− 299·9	+ 0·1	0

Nominal frequency corresponds to a value of 9 192 631 770·0 Hz for the caesium F,m (4.0)–F.m (3.0) transition at zero field.

Notes: (1) All measurements were made in terms of H.P. Caesium Standard No. 134 which agrees with the NPL Caesium Standard to 1 part in 10^{11}.
(2) The offset value for 1968 will be −300 parts in 10^{10} from nominal frequency.
(3) Between March 29, 1968 and April 2, 1968 MSF and GBR ceased emissions due to urgent aerial repairs.

The Institution's 1968 Convention

'Electronics in the 1970s'

SYNOPSES OF PAPERS

The following are synopses of a few of the papers which will be presented at the Convention which is to be held in Cambridge from 2nd to 5th July. An outline programme was published in the April issue of *The Radio and Electronic Engineer*, together with information about registration fees, accommodation, etc.; registration forms are now ready and requests for these should be made to the Institution either using the application form at the end of this issue or by telephone (Convention Registrar, 01-580 8443, extn. 3). The final programme and synopses of the remainder of the papers to be read will be published in the June issue of *The Radio and Electronic Engineer*.

Session 1. (Tuesday, 2nd July, morning.)

Symposium on COMPUTERS AND AUTOMATION

Progress in On-line Control by Computer

D. BEST, O.B.E., B.Eng., C.Eng., M.I.E.E. (*Automation Systems Division, Ferranti Ltd., Wythenshawe, Manchester*, 22.)

An account is given of the development of industrial process control by computer. Some of the control functions carried out by computer are described with examples from a number of industrial processes.

The special features of digital equipment for on-line industrial use are examined. In particular, reliability and fail-safe aspects are emphasized.

System design and programming activities form a major part of the task. Some of the methods used and problems encountered are illustrated.

The Application of Computers in Remote Indication and Control

H. D. MITCHELL, B.Sc.(Eng.), D.C.Ae., and W. RENWICK, M.A., B.Sc., C.Eng., F.I.E.E., F.I.E.R.E. (*Plessey Automation Group, Poole, Dorset.*)

For many years supervisory systems have been used to monitor and control the operation of complex and geographically scattered equipment. An early example is the electricity supply grid but similar requirements exist in such areas as oil or gas pipe-lines, railway electrification, water boards and motorway signalling.

This paper gives a short historical review of the techniques which have been and are now in use and discusses how the computer is coming to play a more and more important role in this field. The advantages which a computer offers when compared with previous methods, and some of the problems encountered in realizing an integrated system are considered.

The trend towards lower cost of electronic equipment makes it possible to re-appraise the system design concepts and some of the significant factors will be reviewed.

Simulation Techniques for Traffic Studies

M. G. HARTLEY, Ph.D., M.Sc.Tech., C.Eng., M.I.E.E., and E. T. POWNER, B.Sc., M.Sc.Tech., C.Eng., M.I.E.E. (*Department of Electrical Engineering and Electronics, University of Manchester Institute of Science and Technology.*)

Over the years interest in the behaviour of traffic in cities has increased substantially and several schemes involving the area control of traffic from a central computer are now being installed. Simulation work involving the evaluation of various control policies which might be applied in such schemes has a complementary role.

In this paper hardware and software simulation techniques are contrasted. Particular attention is paid to the hardware simulation equipment and associated software programs which have been developed by the authors and their colleagues.

The Representation of an Electronic Circuit Diagram in Digital Form to Permit Drafting by a Digital Computer

W. F. HILTON, D.Sc., Ph.D. (*British Aircraft Corporation (Operating) Ltd., Guided Weapons Division, Stevenage, Hertfordshire.*)

Layout of printed circuit boards in a Drawing Office is both tedious and frustrating. This arises from the cross-over of wires, which are only avoided by alterations to the layout, thus introducing fresh cross-overs.

One method of solving this layout problem by digital computer has been devised by the author, and is currently being programmed. Other groups are working on this same problem. It is desirable that a standard method of digitizing a circuit diagram be agreed upon, which would facilitate comparison of a given circuit as drafted by several (rival) programs.

The two main methods are those such as ECAP, which describe the wires (nodes), and the present method, which lists all connections from component A (including one say, to component B) and then later, when listing connections from B, again lists the same connection. This facilitates detection and elimination of errors.

It is hoped that as a result of discussion of this paper, a standard British coding can be evolved.

Session 3. (Wednesday, 3rd July, morning.)

SURVEY PAPERS

High-power Ultrasonics in Industry

ALAN E. CRAWFORD, C.Eng., F.I.E.R.E. (*Ultrasonics Products Group, Radyne-Delapena Ltd., Wokingham, Berkshire.*)

The uses of ultrasonic energy for industrial processing have shown sporadic growth since the initial exploitation in the early 1950s. As with other concentrated energy sources many effects exhibited as laboratory demonstrations have required extensive development before they can be used in industry.

Only a few applications have been established as practical processes although the present trend of development shows greater promise of success than the previous history of industrial ultrasonics.

The presented paper reviews applications and equipment over the past twenty years and proposes reasons for the previous slow growth. Applications are discussed in terms of primary and secondary effects and are considered in well-defined fields of gaseous, liquid and solid state phenomena. These include agglomeration dispersion, reaction stimulation and disintegration in gases and liquids, surface effects, machining and cutting aids, molecular fusion and solid–liquid interactions in solids.

The direct influence of electronic developments on equipment design is considered to be a major reason for the present increase in exploitation and this is discussed in detail.

Future trends in applications and the type of equipment required are assessed in terms of present-day knowledge. Possible advantages and comparison with other methods are given.

Session 4. (Wednesday, 3rd July, afternoon.)

Symposium on AUTOMATIC TEST EQUIPMENT

Co-ordinator and opening speaker: Colonel R. KNOWLES, C.Eng., F.I.E.R.E., R.E.M.E.

The introduction of increasingly complicated electronic systems has inevitably led to requirements for more advanced forms of automatic test equipment. For a number of years automatic testers have been widely used for the pre-flight check-out of aircraft and space vehicles, and confidence testing of weapon systems. More recently the use of automatic testing has been extended into the field of diagnostic maintenance to isolate equipment faults.

The development of automatic test equipment has been greatly assisted by the application of techniques originating in the computer and control equipment industry. There have also been significant advances made in the 'interface' area between prime system and automatic test equipment.

Users of automatic test equipment, namely the three armed services, will review their experience with this type of equipment and discuss their future requirements and plans. This will be followed by a review of the industrial aspects of Automatic Test Equipment.

Automatic testing may be regarded as a natural extension of automation from manufacturing processes into those of testing and inspecting (Quality Control). Automatic testing here refers to off-line applications as opposed to on-line data-logging, monitoring and alarm systems. Significant developments have occurred in these techniques during the last five years.

The present philosophy will be discussed and examples of current designs of automatic test equipment described. The application of digital computers to automatic testing will be introduced, and possible trends in automatic test methods, taking account of advances in computer technology and programming techniques, examined.

The implications of introducing automatic test equipment to industry will be discussed. These include work organization, physical arrangements, capital and running cost considerations, human relationships, operator and programmer training. Contracting procedure for automatic test equipment will be discussed, including definitions of 'delivery' within contracts containing hardware and software commitments. Possibilities of automatic methods for increasing the capability of Standards Laboratories will also be described.

Eight contributions will be included in this Symposium.

Session 5. (Thursday, 4th July, all day.)

Symposium on COMMUNICATIONS

Global Communications: Current Techniques and Future Trends

R. W. CANNON, C.Eng., F.I.E.E., F.I.E.R.E. (*Cable and Wireless Ltd., Mercury House, London, W.C.*1.)

The current situation in the demand for communications facilities is briefly surveyed. Each of the major facets of global communications is considered. A brief review of the outstanding technical characteristics of each method of communication is given and the relationship between each method and the world system is examined; in some cases an actual system is described. Finally, probable or actual future trends in each field are indicated.

The fields covered are: *Systems*—h.f. radiotelephone; h.f. radiotelegraph; submarine coaxial cable systems; communication satellite systems; line-of-sight microwave systems; tropospheric scatter systems; waveguides; lasers. *Exploitation of Systems*—Lincompex; automatic error correction; voice multiplexing; telegraph multiplexing; telephone switching; telegraph message switching; telex and leased circuits; data transmission; television.

The Control System of the National Grid and its Communication Links

P. F. GUNNING, C.G.I.A. (*Central Electricity Generating Board, Courtenay House, London, E.C.*4.)

The C.E.G.B. generates power to meet the instantaneous demands of the eleven Electricity Boards in England and Wales (maximum 35,800 MW in January 1968). With 230 generating stations and grid transmission at 400, 275 and 132 kV and with 700 grid switching and/or high voltage transforming stations, the C.E.G.B.'s integrated power system is the largest in the world under unified control.

To control and protect this countrywide power system, the C.E.G.B. requires a correspondingly large network of more than 2000 communication links. The paper briefly describes the C.E.G.B.'s three-tier control system which operates at district, area and national levels, and was arrived at over 30 years of control of the British 'Grid'. He explains some of the philosophy which has gone into the planning of the networks and into the design of the associated speech, remote control, general indication, telemetering and high-speed protection installations.

A Communication Network for Real-time Computer Systems

D. W. DAVIES, B.Sc., A.R.C.S. (*Division of Computer Science, National Physical Laboratory, Teddington, Middlesex.*)

Digital communication between computers and their users has until now been provided by exploiting the telephone network to carry digital data. This is an obvious and sensible development because of the wide coverage of the telephone network and the high capital cost of establishing any new, generally-available telecommunication service. In the long run it is expected that a specialized digital communication network will be required. This paper contains a specific proposal for a digital communication network designed for computers offering real-time services to remote users.

It has been argued that the immense capital cost of the existing telephone network implies that any new telecommunication facilities must be provided by gradual development from it. This is by no means obvious, and the argument put forward in this paper is that a specialized network operating as a separate system is necessary to provide the facilities required by real-time computers. To further the argument it is necessary to give a full account of the system design of the proposed new network and make rough estimates of its economic viability. At the present time, the missing element in the argument is a firm estimate of the amount of traffic that will be generated for the new network, and work is in hand to produce this estimate.

U.D.C. 621.382 :601.2.085

A Thin Film, Cold Cathode, Alpha-numeric Display Panel

By

R. W. LOMAX,
B.Sc., C.Eng., M.I.E.E., A.Inst.P.†

AND

J. G. SIMMONS,
B.Sc., Ph.D., F.Inst.P.‡

Summary: A vacuum-deposited thin film metal–insulator–metal (Al-SiO-Au) device, after an electro-forming process, emits hot electrons, and constitutes a cold cathode. An array of 25 of these Al-SiO-Au cold cathodes, each of approximately 0·1 cm² area, is used with a phosphor screen in an alpha-numeric display panel, which is clearly visible under normal laboratory lighting conditions. The normal cathode bias is about 12 V. The normal screen potential is about 3 kV. Screen brightness varies directly with the screen potential and exponentially with cathode bias. Pulse operation shows that the rise and decay times of the light output are limited only by the phosphor used. The thickness and composition of the cathode insulator film determine the operating characteristics, notably life. Cathodes of high emission current density, for example, 250 μA cm^{-2}, may operate for about 1 hour. Some with initial high ratio of emission current to current circulating through the cathode, for example, up to one per cent, may operate for 1000 hours. Cathodes of emission current density that is approximately uniform over the cathode area, and therefore suitable for a display, operate for several hundred hours. Failure mechanisms are due to continued forming during operation, and dielectric breakdown.

1. Introduction

1.1. *Objectives*

In a recent series of papers,[1-5] thin film metal–insulator–metal devices under certain operating conditions have been shown to manifest extremely interesting phenomena, which are of potential technological significance. These phenomena include temperature independent conductivity, voltage-controlled negative resistance, reversible voltage,[1,2] thermal-voltage storage effects,[1] and hot electron emission.[3-5] The technological significance of these phenomena has been demonstrated in the form of binary and analogue memories[2] and cold cathode emitters.[5] This paper follows up the previous cold cathode work by way of a thin film, cold cathode, alpha-numeric display. The work embodies all our previous discoveries on cold cathode emission,[3-5] and no new physical phenomenon is reported here; rather we wish to demonstrate how we have exploited the effect to produce a rather novel and technologically important system.

Before describing the actual display, it is appropriate, at this stage, to describe briefly the relevant characteristics of the thin film cold cathodes[3-5] of which the display is constructed.

1.2. *The General Cold Cathode Characteristics*[3-5]

The thin film cold cathode consists basically of a thin film insulator of 200 to 3000 Å (1 Å = 10^{-10} m) thickness sandwiched between metal electrodes, usually of between 200 to 1000 Å thickness. We have been concerned mainly with studying the Al-SiO-Au system. This is by no means unique, not only from the point of view of the metal electrodes, but also from that of the insulator. We have observed that a variety of insulators other than silicon monoxide may be used to obtain the effects to be described in this paper. For instance, magnesium fluoride, sodium chloride, tantalum oxide and aluminium oxide have been used. However, in this paper we shall be concerned solely with the Al-SiO-Au system.

Before the cathode will emit electrons, there is the necessity to subject its insulator film to an electro-forming process, which consists of applying a voltage in excess of about 5 volts across the insulator film in a vacuum better than about 10 Nm^{-2}, or approximately 10^{-1} torr, with the gold electrode biased positive. This causes the insulator resistance to fall considerably, for example, from typically 10^6 ohm to 10^3 ohm; in addition, the cathode is found then to exhibit a pronounced voltage-controlled negative resistance region in the range 5 to 8 V, and a high resistance region above this. The current-voltage characteristic is symmetrical about its origin. For

† Standard Telecommunication Laboratories, Harlow, Essex.
‡ Physics Department, University of Lancaster.

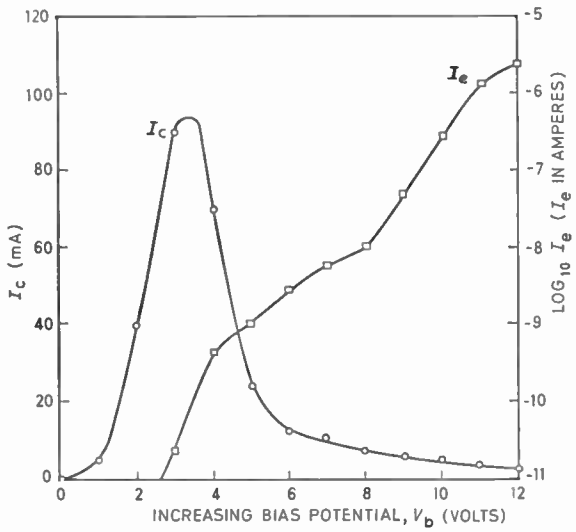


Fig. 1. Graph of circulating current, I_c, against $\log_{10} I_e$, for a typical cathode. (I_e = emission current.)

voltages above about 2·8 V, the cathode is found also to emit electrons, which may be collected by a suitably biased anode. The emission is from localized spots on the surface of the top electrode. The circulating current, I_c, and the emission current, I_e, are shown as a function of the cathode bias potential, V_b, in Fig. 1. It should be noted that the I_e scale is logarithmic, showing that I_e varies quite markedly with small changes of V_b.

The emission process can be shown on two counts to be due to field-accelerated hot electrons rather than electrons thermally excited from the top metal electrode. Firstly, I_e in Fig. 1 is observed to increase as V_b increases in the range of from 3 to 7 V, although the power dissipated in the cathode actually decreases. Secondly, on reversing the polarity of V_b, no emission is observed, even though the $I_c - V_b$ characteristic is symmetrical about its origin, that is, the same power is generated in the cathode independently of the polarity of V_b. It is assumed that, if the cathode is heated, the top and bottom electrodes are still at approximately the same temperature.

So far only a general description of the cathode characteristics has been given; for further details reference may be made to the literature.[3-5] The rest of the paper is devoted to the description and performance of the alpha-numeric display using the cold cathode as an element.

2. Constructional Details of the Display Panel

2.1. *Fabrication of the Cold Cathode Array*

Cathodes are made by the deposition of the constituent thin films on to glass substrates in an oil diffusion pumped vacuum system, which is equipped

Table 1

Typical depositions for a 5 × 5 array of cathodes

Deposition number	Name	Material	Approximate thickness (Å)
1	Underlayer	Silicon oxide	2000
2	Contacts	Chromium	200
3	Contacts	Gold	500
4	Bottom electrode leads	Aluminium	3000
5	Lead insulation	Silicon oxide	8000
6	Bottom electrode	Aluminium	1000
7	Insulator film	Silicon oxide	400
8	Thickened insulator	Silicon oxide	1200
9	Top electrode	Gold	200
10	Top electrode leads	Gold	500

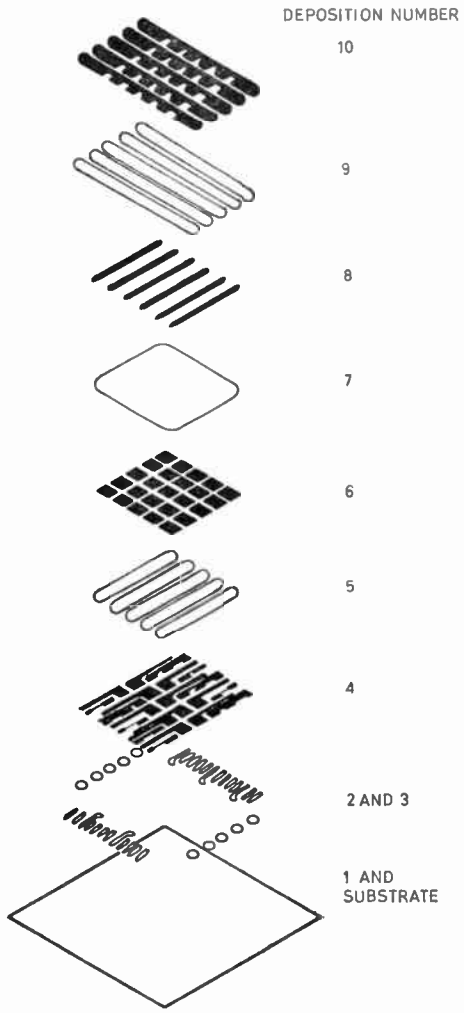


Fig. 2. Exploded view of a 5 × 5 array of cathodes.

with a liquid nitrogen trap. The metals are evaporated from electrical resistance-heated tungsten ribbon boats at a chamber pressure of about 10^{-6} torr. The silicon monoxide is evaporated from an alumina crucible heated by an overhead tungsten filament. The insulator is deposited at a rate of approximately 10 Å s^{-1} at a pressure of approximately 10^{-3} N m^{-2} or 10^{-5} torr maintained by a constant leak of oxygen gas in to the vacuum chamber. The ratio of deposition rate to oxygen pressure for a particular substrate temperature decides the film composition, which has a strong influence on the cathode operating characteristics and performance.

The display has 25 cold cathodes of area approximately 0·1 cm^2 in the form of a 5 × 5 array, an exploded view of which is shown in Fig. 2. The array is deposited on a 49 mm by 49 mm glass substrate. The various films making the array are deposited usually in the order given in Table 1. The film thicknesses deposited and their rates of deposition are monitored during deposition by an oscillating quartz crystal film thickness monitor.

Because the $I_c - V_b$ characteristics of the cathodes are symmetrical about their origins, it is not possible to use co-ordinate switching techniques on an array at present; it is thus necessary to make individual connections to each elemental cathode in the present array. This difficulty is expected to be overcome by use of thin film diodes in series with each elemental cathode in an array.

Deposition 1 is an underlayer deposited directly upon the substrate to enhance its smoothness; this, in turn, increases the voltage withstood by cathodes before the occurrence of dielectric breakdowns in the cathode insulator film. Metal depositions 2 and 3 provide contacts; the chromium, which adheres strongly to the underlayer, gives a good bonding surface for the gold. Deposition 4 provides the bottom electrode leads, which pass the contacts to the bottom electrodes of the cathodes. This film is made relatively thick to keep the series resistance of the lead to a low value. The necessity of this is discussed in Section 4.1.2.

After deposition 4, the vacuum chamber is opened and some masks are changed, since all the masks cannot be accommodated simultaneously in the coating unit to fabricate a complete array. Deposition 5 insulates the bottom leads for all but the very edge of each, which connects to the appropriate subsequently deposited bottom electrode.

Deposition 6 provides the bottom electrodes for the elemental cathodes.

Deposition 7 is the cathode insulator film. Its thickness depends on whether a high emission current or a high emission efficiency (I_e/I_c) is required. An

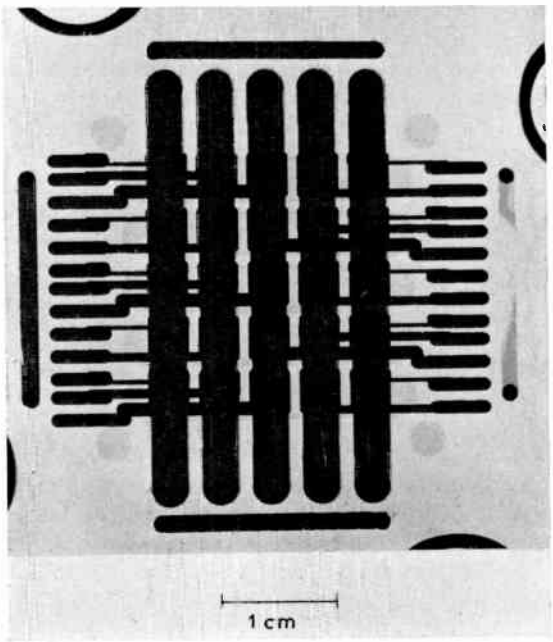


Fig. 3. A 5 × 5 array of cathodes.

insulator film thickness of about 200 Å or less gives rise to a cathode of high emission, which is capable of providing an emission current density of greater than 250 μA cm^{-2}. The emission is distributed very uniformly over the cathode area. An insulator of this thickness, however, also gives rise to a high circulating current density of the order of 1 A cm^{-2}. This causes cathode heating, which may result in cathode failure.

A cathode with insulator film of about 1000 Å thickness or more can withstand a high bias voltage, before suffering electrical breakdown. As can be seen by consideration of both $I_c - V_b$ and $I_e - V_b$ characteristics in Fig. 1, the greatest emission efficiency is obtained at high values of V_b. For thick insulator film cathodes operating at a bias voltage of about 19 V, emission efficiencies of up to about 1% have been recorded. Unfortunately, at present, cathodes with thick insulator films emit only at a small number of points on the cathode area and the total emission current is no greater than that from cathodes of thinner insulator film. Moreover, cathodes having thick insulator films, because of their lack of uniformity of emission, have a poor appearance in a display.

For a display, the usual thickness for the cathode insulator film is about 400 Å, which is a compromise between the two extremes.

The insulator film composition also affects the operating characteristics. This is discussed in Section 4. It has been found that the insulator film is

thinner where it passes over the edge of the bottom electrode, and is the origin of premature electrical breakdown in the cathode. In order to obviate this, deposition 8 thickens the insulator film over the edges of the bottom electrodes where they are crossed by the top electrodes. In addition, it also covers the bottom electrodes, where the latter are in contact with the bottom electrode leads, and thus defines the width of the operating area of the cathodes.

Deposition 9 provides the top electrodes of the cathodes, which are only about 200 Å thick. Films of this thickness tend to have pin-holes or depressions, which increase the probability of emission into the vacuum of a hot electron injected into the top electrode from the insulator film. This is because emission is from the walls of pin-holes or depressions in the top electrode.[3-5] Deposition 10 thickens the top electrode leads between the contacts and the cathode area, and so reduces the series resistance of the lead. The necessity of this is discussed in Section 4.1.2.

Figure 3 is a photograph of a completed 5×5 array of cathodes.

2.2. *Assembly of a Complete Display Panel*

A phosphor screen is used to display the emitted electrons. The screen is held parallel to the 5×5 array of cathodes, at a distance of approximately 2·5 cm, by four sheets of mica, the whole structure forming a box arrangement. The phosphor screen consists of a 49 mm by 49 mm glass substrate, which has been coated with an electrically-conducting layer of tin oxide, upon which phosphor has been deposited. The phosphor is deposited in a powder form by allowing it to settle out of a mixture of weak aqueous solution of barium nitrate and potassium silicate. The phosphor used depends on the particular display application. Generally, a copper-activated willemite powder is used.

The display so formed, is operated normally in the laboratory in a demountable ion-pumped vacuum system at a pressure of about 10^{-3} N m^{-2} or 10^{-5} torr.

3. Operating Characteristics

3.1. *Steady-bias Operation*

Figure 4 shows some photographs of alpha-numeric characters, which are obtained with the display panel. With this simple 5×5 array, all numerals and letters of the alphabet can be displayed legibly. Figure 5 shows the display panel operating in the laboratory under normal lighting conditions, illustrating that the display is clearly visible. The circular and fuzzy nature of the individual patches making up the figures is due to coherent scattering[3-5] of the electrons, as they pass through the top electrode. The

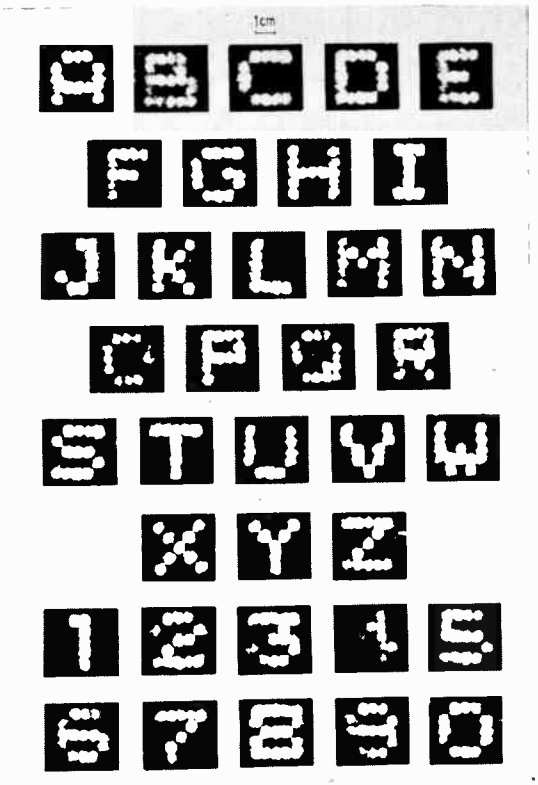


Fig. 4. Some alpha-numeric characters obtained with the display panel.

effect is minimized by either increasing the accelerating potential or by placing the phosphor screen closer to the cathodes. The latter procedure, however, although increasing emission definition, results in some loss of character resolution; the closer the screen is to the cathodes, the greater the density of array elements required. Figure 5 illustrates the brightness of the display in room ambient lighting condition.

Figure 6 shows the circuit used to obtain the displays shown in Figs. 4 and 5. V_a is the accelerating potential applied to the phosphor screen. Only five elemental cathodes are shown in cross-section. The limiting resistor prevents overload of the emission current circuit if a cathode suffers a dielectric breakdown. This produces gaseous products in the vacuum space between the cathode and the phosphor screen. Without the limiting resistor an arc may form between them; with the limiting resistor, if the current tends to rise to high values in the emission current circuit, the phosphor screen potential falls to practically zero and prevents arcing.

When the cathodes are used in a display panel, the brightness of the panel depends firstly on the intensity of the electron emission, which varies exponentially with V_b, and secondly on the energy of the electrons

striking the screen, which is proportional to V_a. Thus, as far as its use in a display is concerned, the emission current from a cathode, and hence its emission efficiency, need not be high if V_a is high instead, but the emission over the cathode area needs to be uniform. Cathodes emitting large currents have operated at accelerating potentials as low as 500 V and have given displays visible under normal lighting conditions. For cathodes of lower emission current, V_a should be about 3 kV.

3.2. *Alternating-bias Operation*

V_b is normally a steady positive bias of about 10 to 15 V; however, pulsed voltages and sinusoidally alternating voltages of similar peak values have been used as well. The total light output is then lower, because the maximum emission density is obtained only when the bias is at its peak positive value. In one experiment, an optical display was obtained with an alternating voltage of 7 V r.m.s. across the cathode and 1000 V r.m.s. accelerating potential; the voltages were in phase, and were obtained from a common small mains voltage transformer.

3.3. *Pulsed-bias Operating*

Figure 7 shows photographs of traces on a cathode-ray oscilloscope. In Fig. 7(a) the upper trace displays a square voltage pulse of approximately 10 V peak and 70 μs duration applied to the cathode. The lower trace shows the resulting circulating current pulse of approximately 7·5 mA. The circuit used is shown in Fig. 8. The spikes at the leading and trailing edges of the current pulse are caused by the charging and discharging of the cathode capacitance which has a value of about 10^{-9} F.

Figure 7(b) illustrates the cathode emission current response to voltage pulsing, the emission current being collected by a metal anode placed 2·5 cm away from the cathode, and at a positive potential of 90 V with respect to it. The circuit used for this measurement is shown schematically in Fig. 9. The voltage pulse is once again 10 V peak and of 70 μs duration. The resulting emission current during the pulse corresponds to about 1 μA. The response at the anode of the emission current to the voltage pulse is seen to be virtually instantaneous, which further verifies that emitted electrons are not thermally excited from the top electrode but rather are 'hot' electrons.

To determine the phosphor screen light output response the circuit shown schematically in Fig. 8 was used again. Figure 7(c) illustrates the phosphor screen light output response to voltage pulsing. The upper trace once again depicts a 10 V, 70 μs applied voltage pulse. The lower trace shows the output current pulse from a 56 AVP photomultiplier, which was placed with its photo-cathode about 10 cm in front of the phosphor screen. The phosphor on the screen was type P24 and the accelerating potential was 2·8 kV. It can be seen that the light had a decay-time of approximately 6 μs, this being due to the type of phosphor used.

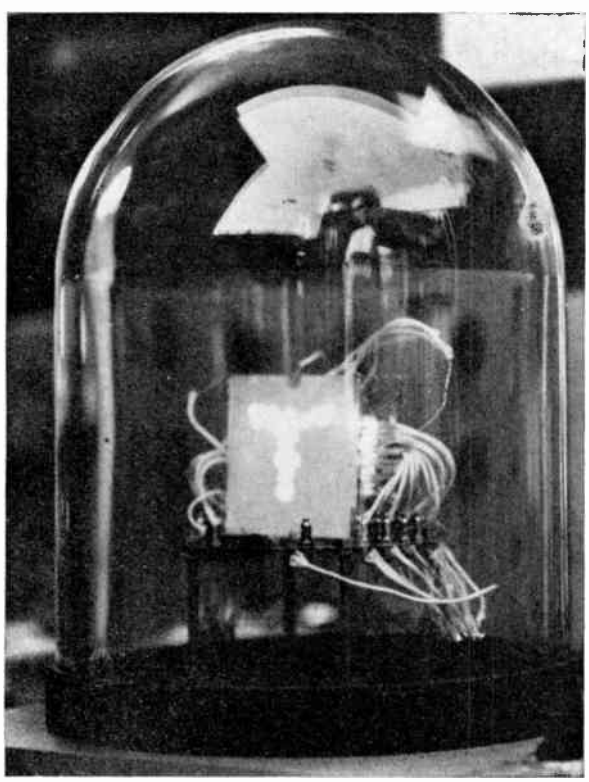


Fig. 5. Operating display panel under room ambient lighting conditions.

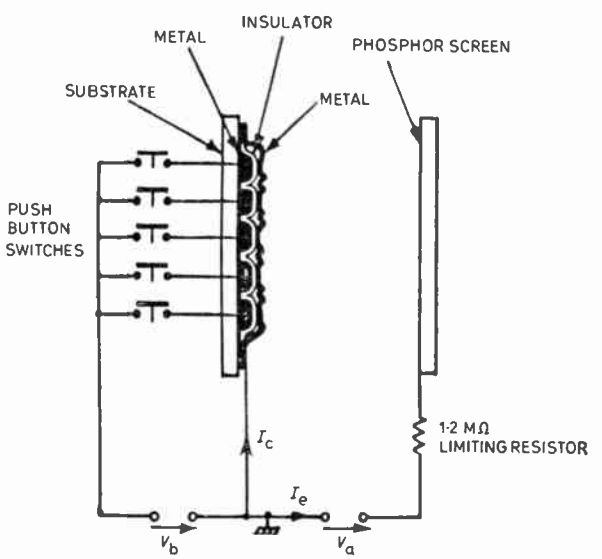
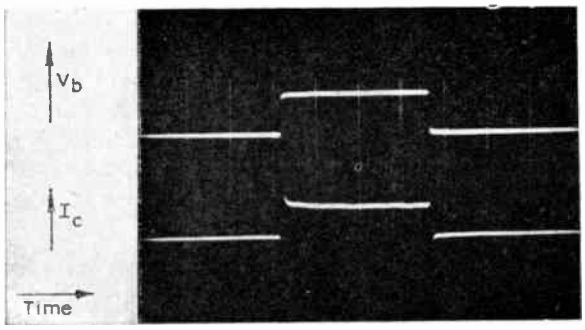
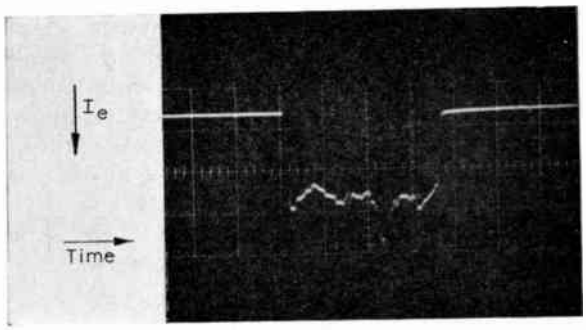


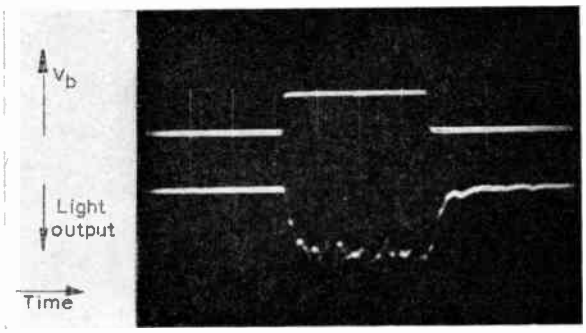
Fig. 6. Diagram of circuit used for the display panel.



(a) Upper trace, V_b pulse; lower trace, I_c pulse.



(b) Electron emission pulse.



(c) Upper trace, V_b pulse; lower trace, light output.

Fig. 7. Traces on a cathode-ray oscilloscope.

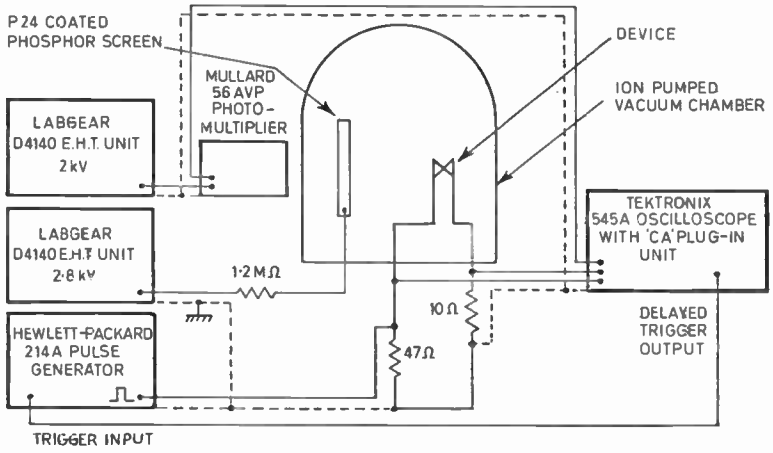


Fig. 8. Diagram of circuit used to record V_b, I_c and light output pulses.

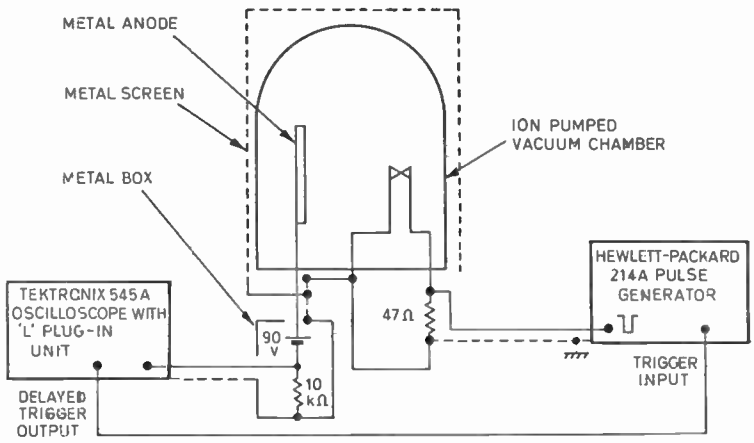


Fig. 9. Diagram of circuit used to record I_e pulse.

4. Cathode Performance

4.1. *Failure Mechanisms*

4.1.1. Dielectric breakdown

Self-healing dielectric breakdowns are observed in the cathode, if the bias potential is increased to too high a value. These occur where the cathode insulator film is locally thin, and could be due to foreign matter on the substrate, or pinholes in underlying depositions. These self-healing dielectric breakdowns are similar to those observed in other capacitor-like structures.[6-10] After their occurrence, the cathode is found not to be shorted between the electrodes, and is still operable. When observed under the microscope, these self-healing dielectric breakdowns are seen to be circular holes, which often pass completely through both top and bottom electrodes as well as the insulator film. The bias potential at which self-healing dielectric breakdowns start to occur frequently, varies with insulator thickness.

Sometimes chains of breakdowns are seen over the area of a failed cathode. When observed under the microscope the chains always appear to originate at a self-healing breakdown. Breakdowns seem to be temperature dependent.[10] A self-healing dielectric breakdown can increase the temperature of the cathode, in its locality, which in turn can initiate another breakdown nearby. If such a breakdown increased the temperature of the cathode in its locality as well, a chain of breakdowns could occur.

Some chain breakdowns appear to have meandered over the area of a cathode, until little area is left. This, of course, causes failure of the cathode. For a particular cathode area, the probability of occurrence of a chain breakdown is partly dependent on the deposition conditions of the cathode insulator film.

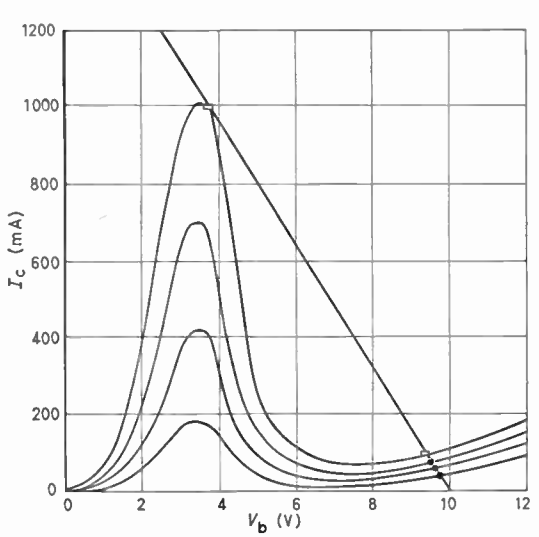


Fig. 10. Different states of forming of a cathode and the series resistance load-line.

These conditions are determined by the substrate temperature, and the ratio of the rate of deposition to the partial pressure of oxygen in the deposition chamber. These are just the parameters that decide the composition of the film.[11] The composition of vacuum-deposited 'silicon monoxide' is found always to lie somewhere between silicon and silicon dioxide. The cathodes that showed least probability of occurrence of a chain breakdown are apparently those with a composition nearer to silicon dioxide than silicon.

The probability of occurrence of a chain breakdown may be said to be linked with the 'hardness' or 'softness' of silicon oxide capacitors investigated by Klein and Gafni.[10] The greater the probability, the quicker the area decreases and the sooner is the breakdown complete.

4.1.2. Continued 'forming'

The cathode circulating current appears to increase during operation, presumably because the 'electroforming' process continues. This eventually causes failure of the cathode for one of two reasons. Firstly, after the circulating current has reached a high value, a whole cathode may blow up suddenly in what could be a series of practically simultaneous dielectric breakdowns. It would be reasonable to explain this by assuming that the cathode is heated by the circulating current passing through it, and the resulting increase in temperature eventually initiates dielectric breakdown over the whole cathode area.

Secondly, the increased circulating current can cause the bias potential across the cathode to fall because of the increased voltage drop in the thin film leads to the cathode. This reduces greatly the emission current, as can be observed from the $I_c - V_b$ characteristic shown in Fig. 1. It can also be the reason for total failure, as will now be explained by the use of Fig. 10, which illustrates a series of $I_c - V_b$ characteristics of a cathode for different states of 'forming'; the longer the cathode operates, the more it is 'formed' and the higher is the circulating current. Figure 10 also shows the load-line for the series resistance due to the thin film leads to the cathode. The single operating points for the first three states of forming are shown by black dots on Fig. 10, assuming a constant voltage supply of 10 V. If a state of forming, as illustrated by the upper curve in Fig. 10, is finally attained in the cathode, a high current point, near to the peak of the $I_c - V_b$ characteristic is possible. The cathode may switch its operation to this point and produce an increase in the power dissipated in the thin film leads. This increase can be of the order of one hundred times. The power dissipated in the cathode increases by only several times, cathode failure then results from rupture

of the thin film leads. Such failures can occur when reswitching on cathodes, which have already been operating for some time.

4.2. *Life-times of Cathodes*

If the conditions of substrate temperature, deposition rate and partial pressure of oxygen during deposition of the cathode insulator film are incorrect, or are controlled poorly, a cathode may exhibit a low breakdown voltage. On attempting to operate such a cathode at usual bias potential, many chain breakdowns occur and cause almost immediate failure. This could be said to be a 'soft' cathode. For a long life-time, 'hard' cathodes with an insulator film composition nearer to silicon dioxide are required.

As stated in Section 2.1, the thickness of the cathode insulator film decides the cathode characteristics. If the insulator film is thin, a high-emission cathode is obtained. As the circulating current is also high, the cathodes become overheated. They fail eventually owing to mechanisms given in Section 4.1, which are enhanced by higher temperatures. High-emission current devices may operate for up to an hour.

Cathodes that are of the thickness required to give them a reasonably uniform emission over all of their area, have a longer life. They fail mainly on account of apparently increased forming during operation, as explained in Section 4.1.2. Usually, they last for several hundred hours. 'Hard' cathodes have often failed owing to the thin film leads burning up or the substrate under the cathode cracking without there being a sign of dielectric breakdown on the cathode.

Cathodes of thick insulator film, of inherent initial high emission efficiency, (I_e/I_c), may last for several hundred hours, even above 1000 hours. They have little tendency to continue 'forming' during operation. After a few minutes, they often reach stable operating conditions which continue throughout their life. Most of them fail if an over-voltage is applied, as can happen when switching on. They emit at a small number of distinct points, and it is possible that these are points of thin cathode insulator film. This seems to be a likely case, for these points would be the first to experience self-healing dielectric breakdowns. Often the emission disappears suddenly, yet the cathode looks little damaged. This fact suggests that the emitting points have indeed suffered self-healing dielectric breakdowns.

5. Conclusions

It has been shown that, with suitable control during manufacture, thin-film cold-cathode alpha-numeric displays are quite feasible. The basic structure of the elements making up the display is remarkably simple and quite easy and cheap to fabricate. Before commercial exploitation can be contemplated several problems remain to be solved. The main emphasis must be on improving the life-time of the cathodes, which was only a few minutes in earlier studies. Clearly, the life-time is a function of the preparation of the insulator, and probably also of the electrode materials. Furthermore, silicon monoxide is probably not the optimum insulator material, but has been used because it is the material we are most familiar with at this stage. It is also hoped that use of a thin film diode placed in series with each cathode would allow manufacture of co-ordinate switched arrays of high element density. It should be pointed out that the display produced, having individual connections to each elemental cathode of the array, has proved useful in its own right, for alpha-numeric characters are generated continuously. The display does not need to be scanned and the circuitry required is very simple.

6. Acknowledgments

The authors would like to thank Mr. R. Jewitt for the deposition of the cathodes.

This work has been carried out under a C.V.D. contract and is presented by permission of the Ministry of Defence (Navy Department).

7. References

1. J. G. Simmons and R. R. Verderber, 'New conduction and reversible memory phenomena in thin insulating films', *Proc. Roy. Soc.*, **A301**, pp. 77–102, October 1967.
2. J. G. Simmons and R. R. Verderber, 'New thin-film resistive memory', *The Radio and Electronic Engineer*, **34**, No. 2, pp. 81–9, August 1967.
3. J. G. Simmons, R. R. Verderber, J. Lytollis and R. W. Lomax, 'Coherent scattering of hot electrons in gold films', *Phys. Rev. Letters*, **17**, No. 9, p. 675, 6th September, 1966.
4. J. G. Simmons and R. R. Verderber, 'Observations on coherent electron scattering in thin film cold cathodes', *Appl. Phys. Letters*, **10**, No. 7, p. 197, 1st April, 1967.
5. R. R. Verderber and J. G. Simmons, 'A hot electron, cold cathode emitter', *The Radio and Electronic Engineer*, **33**, No. 6, pp. 347–51, June 1967.
6. H. Strab and H. Maylandt, 'Present Stage of Technique of Metallised Paper Capacitors for Power Systems', Conf. Int. des Grands Reseaux Electriques, Paris, 1958. Report No. 109.
7. D. R. Kennedy, 'Effects of Moisture on Electrical Properties of Anodic Aluminium Oxide', Electrical Research Association, Leatherhead, Report, Ref. L/T 374, 1958.
8. G. Siddall, 'Vacuum deposition of dielectric films for capacitors', *Vacuum*, **9**, No. 5–6, p. 274, 1959.
9. L. Young, 'Anodic Oxide Films', p. 125 (Academic Press, New York, 1961).
10. N. Klein and H. Gafni, 'The maximum dielectric strength of thin silicon oxide films', *Trans. Inst. Elect. Electronics Engrs on Electron Devices*, **ED-13**, p. 281, February 1966.
11. E. Ritter, 'Uber die chemische Reaktion bei der Bildung dünner Schichten durch Kondensation', *Monatshefte für Chemie*, **95**, p. 795, 1964.

Manuscript received by the Institution on 12th January 1968. (Paper No. 1186/CC5.)

U.D.C. 621.382.2

A Theoretical Analysis of the Ideal Step-recovery Diode in the Series-mode of Operation

By

W. M. VAN LOOCK, M.Sc.†

AND

A. CARDON, B.Sc.†

Summary: An analysis is given of frequency multiplication with the ideal step-recovery diode in the series-mode of operation. Only one conduction angle per period of the fundamental frequency is allowed. The series resistance of the diode is assumed to be different from zero. General and explicit formulae are given for the input impedance, the output power and the power efficiency. The validity of the approximations is discussed throughout. It is shown that the efficiency, η, of the power conversion is proportional to $1/n^2$, and that the maximum value for η is about 15% for a times 10 multiplier in this series-mode of operation.

List of Symbols

τ	life-time of minority carriers
Q	stored charge in the diode
n	multiplication factor
$E' \sin(\omega t+\alpha)$	source voltage, where α is the phase angle with respect to the voltage $E_1 \sin \omega t$ at the input of the diode
ω	fundamental angular frequency
$\bar{Z}_g = R_g + jX_g$	complex generator impedance
$\bar{F}_1$	ideal filter which is an open-circuit for angular frequency ω and a short-circuit for all other frequencies and d.c.
$\bar{F}_N$	same as $\bar{F}_1$ but for the nth harmonic of ω, $\bar{F}_N$ is always short-circuited except for $n\omega$
$\bar{Z}_L$	complex load impedance
V_0	diffusion potential of the s.r.d., the effect of which is represented by a voltage source
V_b	bias voltage which is derived from the generator or from an external source
$E_n \sin(n\omega t+\phi_n)$	voltage over the ideal filter and load, generated by the nth harmonic current
ϕ_n	phase shift of the nth harmonic
R_s	series resistance of the diode
θ	angle at which the forward conduction starts
ϕ	angle at which the transition occurs
θ', ϕ'	same as θ and ϕ, when the amplitude of the nth harmonic is zero (see Fig. 2b)
$\Delta\theta$	difference between θ and θ'
ψ	conduction angle
E_p	total bias equal to V_0+V_b
S, s	amplitude and normalized amplitude respectively of the step or transition,

$$\mathscr{I}_1 = \frac{E_1}{R_s} \qquad \mathscr{I}_n = \frac{E_n}{R_s} \qquad \mathscr{I}_p = \frac{E_p}{R_s}$$

$$T = \frac{2\pi R_s R_L}{R_L^2+X_L^2} \qquad U = \frac{X_L}{R_L} T$$

$$G_g + jB_g = \frac{1}{\bar{Z}_g}$$

† Laboratory for Electromagnetism and Acoustics, University of Ghent, Ghent, Belgium.

1. Introduction

Non-linear impedances have been extensively analysed in frequency multipliers.[1-6] In this paper, the ideal non-linear capacitor or the step recovery diode (s.r.d.)[7-12] is considered. During the forward conduction of a s.r.d., charge is being stored which is removed when the current reverses. When all stored charge is removed, depending on the current and on the life-time of the minority carriers, the s.r.d. abruptly switches to an open-circuit. This discontinuity can be used in frequency multiplier circuits.

The analysis of these multipliers is always carried out with a negligibly small nth harmonic[7,8] or is limited to a computer solution for discrete values of the multiplication factor n.[9] However in this paper the general and useful formulae which considerably reduce the computing work will be derived. The theory is valid for multipliers using the step recovery diode in a series connection and where only one conduction angle per cycle is allowed.

2. Principles of Operation

The basic series circuit to which this study is limited is shown in Fig. 1. Only $E_1 \sin \omega t$ and $E_n \sin(n\omega t+\phi_n)$ are allowed across the ideal filters

$\bar{F}_1$ and $\bar{F}_N$. When the diode conducts, the voltage sources V_0 and V_b are short-circuited and a direct current in the reverse direction can flow until all the stored minority carriers are removed and the diode switches to an open-circuit condition. For the generality of the analysis, it is not assumed that the diode series resistance, R_s, equals zero because this leads to Dirac pulses.

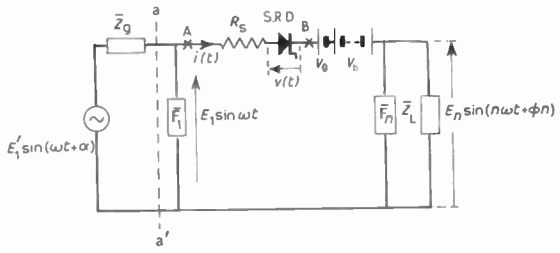


Fig. 1. Basic circuit configuration for the series-mode of operation of the s.r.d. multiplier.

The voltage between A and B is deduced from Fig. 2(a), and the voltage $v(t)$ over the diode is shown in Fig. 2(b). If E_n cannot be neglected with respect to E_1, the conduction angle is determined by $\phi-\theta$, which is different from $\phi'-\theta'$.

In Fig. 3, where E_n is large, multiple steps occur in one cycle. To obtain only one step independently of ϕ_n, one must have

$$n^2E_n^2 \leqslant E_1^2-E_p^2 \qquad \text{.....(1)}$$

For $\omega t = 0$ the voltage $v(t)$ across the diode becomes zero and the conduction begins. During the conduction period, the current $i(t)$ through the diode is given by the expression

$$i(t) = \frac{1}{R_s}[E_1 \sin \omega t - E_n \sin(n\omega t+\phi_n)-E_p] \quad ...(2)$$

3. Determination of Conduction Period

In Fig. 2(a) it can be seen that the angle θ is determined by:

$$E_1 \sin\theta - E_n \sin(n\theta+\phi_n) = E_p \qquad \text{.....(3)}$$

The conduction period ends when $\omega t = \phi$, i.e. when all stored charge is withdrawn. The value for ϕ can be determined from the solution of

$$i(t) = \frac{dQ}{dt} + \frac{Q}{\tau}$$

$$= \frac{1}{R_s}[E_1 \sin \omega t - E_n \sin(n\omega t+\phi_n)-E_p] \quad ...(4)$$

The solution of eqn. (4) is

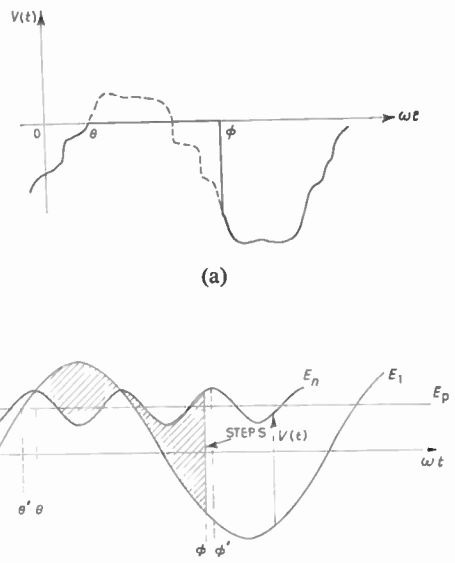


Fig. 2. Waveforms at the diode.

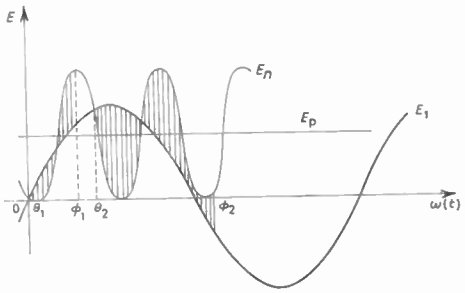


Fig. 3. Series-mode with multiple conduction angles.

$$Q = K e^{-t/\tau} + \frac{\tau}{R_s}[\varepsilon_1 \sin(\omega t-\beta_1)- -\varepsilon_n \sin(n\omega t+\phi_n-\beta_n)-E_p] \quad ...(5)$$

with

$$\varepsilon_1 = E_1(1+\omega^2\tau^2)^{-\frac{1}{2}}$$

$$\varepsilon_n = E_n(1+n^2\omega^2\tau^2)^{-\frac{1}{2}}$$

$$\beta_1 = \arctan \omega\tau$$

$$\beta_n = \arctan n\omega\tau$$

K is a constant of integration. The condition that the stored charge is zero at $\omega t = \theta$ and $\omega t = \phi$ leads to

$$\exp\left(\frac{\phi-\theta}{\omega\tau}\right) = \frac{\varepsilon_1 \sin(\theta-\beta_1)-\varepsilon_n \sin(n\theta+\beta_n)-E_p}{\varepsilon_1 \sin(\phi-\beta_1)-\varepsilon_n \sin(n\phi+\phi_n-\beta_n)-E_p} \qquad \text{.....(6)}$$

For $\omega t > \phi$ the current through the diode remains zero until ωt becomes equal to $2\pi+\theta$ and conduction starts again for the next cycle.

4. General Solution

The condition for power match of the input circuit yields

$$E_1 = E_1' \frac{\sqrt{R_g^2 + X_g^2}}{2R_g} \qquad \text{.....(7)}$$

$$\alpha = \arctan \frac{X_g}{R_g} \qquad \text{.....(8)}$$

Calculation of the nth harmonic power therefore requires the knowledge of six unknowns, namely,

$$E_n,\ \theta,\ \phi,\ \phi_n,\ R_g,\ X_g$$

In consequence, the general solution requires four new equations besides eqns. (3) and (6).

During the conduction angle, $\psi = \phi - \theta$, the current $i(t)$ is given by eqn. (4). For all other angles, the current $i(t)$ in the diode is zero. Therefore,

$$0 < \omega t < \theta \quad \text{and} \quad \phi < \omega t \leqslant 2\pi \qquad \text{.....(9)}$$

This periodic current can be expressed as a sum of harmonics:

$$i(t) = \sum_{m=-\infty}^{+\infty} I_m e^{jm\omega t}$$

$$I_m = \frac{\omega}{2\pi} \int_0^T i(t) e^{-jm\omega t} dt \qquad \text{.....(10)}$$

The coefficients I_0, I_1 and I_n in eqn. (10) can be written as

$$I_0 = \mathscr{I}_1 A_0 + \mathscr{I}_n B_0 + \mathscr{I}_p C_0 \quad (= 0 \text{ if } \omega\tau = \infty) \qquad \text{.....(11)}$$

$$I_1 = \mathscr{I}_1 A_1 + \mathscr{I}_n B_1 + \mathscr{I}_p C_1 \qquad \text{.....(12)}$$

$$I_n = \mathscr{I}_n A_n + \mathscr{I}_n B_n + \mathscr{I}_p C_n \qquad \text{.....(13)}$$

Explicit expressions for A_0, B_0, C_0, A_1, B_1, C_1 and A_n, B_n, C_n are given in Appendix 1.

It is to be noticed (Fig. 1) that

$$R_g = \mathrm{Re}\left\{\frac{E_1}{2jI_1}\right\} \qquad \text{.....(14)}$$

$$-X_g = \mathrm{Im}\left\{\frac{E_1}{2jI_1}\right\} \qquad \text{.....(15)}$$

The fundamental current is given by

$$\frac{E_1'}{2R_g} \sin(\omega t + \alpha) = \frac{1}{2j}\left[e^{j(\omega t+\alpha)} - e^{-j(\omega t+\alpha)}\right] \frac{E_1'}{2R_g}$$

and is to be identified with eqn. (12).

The current and voltage of the nth harmonic are related to the load impedance $\bar{Z}_L = R_L + jX_L$ by the following equations:

$$R_L = \mathrm{Re}\left\{\frac{E_n}{2jI_n} e^{j\phi n}\right\} \qquad \text{.....(16)}$$

$$X_L = \mathrm{Im}\left\{\frac{E_n}{2jI_n} e^{j\phi n}\right\} \qquad \text{.....(17)}$$

The explicit expressions for eqns. (16) and (17) are given in Appendix 2. Equations (16) and (17) together with eqns. (3), (6), (14) and (15) allow determination of E_n, θ, ϕ, ϕ_n, R_g, X_g, given E_1', E_p, Z_L, τ, R_s and n. The solutions must satisfy eqn. (1).

With matched input condition the input power is given by

$$P_{in} = \frac{1}{2}\left(\frac{E_1'}{2R_g}\right)^2 R_g \qquad \text{.....(18)}$$

This input power is partly converted into the nth harmonic power P_n

$$P_n = \frac{1}{2} \frac{E_n^2}{(R_L^2 + X_L^2)} R_L \qquad \text{.....(19)}$$

and is dissipated in the load Z_L. The difference between P_{in} and P_n is dissipated in R_s. Therefore, R_s should be as low as possible. However, for a given s.r.d., R_s cannot be changed and we have to optimize the parameters of the circuit to obtain the maximum possible efficiency or power output.

The efficiency η of the power conversion is given by the ratio of the nth harmonic power and the available power. From eqns. (18) and (19), therefore,

$$\eta = \frac{P_n}{P_i} = 4\left(\frac{E_n}{E_1'}\right)^2 \frac{R_g}{R_L^2 + X_L^2} R_L \qquad \text{.....(20)}$$

Calculation of the optimum efficiency for a given diode (R_s is given) is difficult because of the form of eqn. (20). If the optimum solution is wanted for η, when E_1', $\omega\tau$, R_s and n are given, nine parameters must be determined which requires three supplementary equations:

$$\frac{\partial\eta}{\partial E_p} = \frac{\partial\eta}{\partial R_L} = \frac{\partial\eta}{\partial X_L} = 0 \qquad \text{.....(21)}$$

5. Approximate Solution

5.1. *Limitations*

Because of the complicated form of the equations which lead to the exact solution of the optimum efficiency of the multiplier (Fig. 1), we must proceed to an approximate solution.

Instead of $E_n \ll E_1$ we introduce the less restrictive condition

$$\frac{E_n}{n} \ll E_1 \qquad \text{.....(22)}$$

To simplify determination of the angles θ and ϕ, it can be seen that, if E_n could be neglected with respect to E_1 in eqn. (3),

$$\sin\theta = \frac{E_p}{E_1} \qquad \text{.....(23)}$$

Because this too coarse an approximation is used in the calculations to follow, we have to justify eqn. (23).

Equation (23) gives in fact θ' and the exact value of θ is given by eqn. (3). If $\Delta\theta = \theta - \theta'$, it can be shown using eqns. (1), (3) and (23) that the maximum error $\Delta\theta$ is less or equal to $1/n$ radian. For $n = 2$ this approximation is not tolerable. In this case E_n must be smaller than given by eqn. (1). Therefore a good idea of the maximum allowable E_n for a given maximum $\Delta\theta$ (e.g. 0·1 radian) can be obtained from the relationship

$$\Delta\theta = \frac{E_n}{\sqrt{E_1^2 - E_p^2}} \qquad \text{.....(24)}$$

This equation limits the amplitude of E_n or in other words eqn. (24) defines the maximum allowable degree of coupling of the nth harmonic tuned circuit. Because E_n is assumed to be small, the nth harmonic load must have a low impedance which in practice can be accomplished by suitable coupling to the resonator or by use of a transmission line section.

5.2. *Conduction Angle*

According to eqn. (22), if we neglect ε_n in eqn. (6) which gives with eqn. (23)

$$\cot\theta = \frac{-\omega\tau\sin\psi - \cos\psi + (\omega\tau)^2(1 - e^{-\psi/\omega\tau}) + 1}{\sin\psi - \omega\tau\cos\psi + \omega\tau\, e^{-\psi/\omega\tau}} \qquad \text{...(25)}$$

θ versus ϕ and $\omega\tau$ as a parameter can also be derived from Fig. 6 of Reference 7. From this reference it appears that the curves for $\omega\tau = 500$ or $\omega\tau = \infty$ practically coincide. This conclusion is independent of the other approximations. Because it is easy to have $\omega\tau = 500$ with practical diodes, we will put $\omega\tau = \infty$. Thus, eqn. (25) becomes

$$\tan\theta = \frac{1 - \cos\psi}{\psi - \sin\psi} \qquad \text{.....(26)}$$

5.3. *Matching Conditions and Efficiency*

The quantities described in Appendix 2 can be simplified as follows when condition (22) holds:

$$H = K = 0$$

$$G = L = \phi - \theta = \psi$$

$$D = \frac{2}{n}(\sin\phi\sin n\phi - \sin\theta\sin n\theta)$$

$$F = \frac{2}{n}(\sin\theta\cos n\theta - \sin\phi\cos n\phi)$$

$$I = \frac{2}{n}(\sin n\theta - \sin n\phi)$$

$$M = \frac{2}{n}(\cos n\phi - \cos n\theta)$$

Equations (54) and (55) from Appendix 2 become

$$E_n(T+\psi)\cos\phi_n + E_n U\sin\phi_n = E_1\frac{2}{n}(\sin\theta - \sin\phi)\cos n\phi \qquad \text{.....(27)}$$

$$-E_n U\cos\phi_{n+} + E_n(T+\psi)\sin\phi_n = E_1\frac{2}{n}(\sin\phi - \sin\theta)\sin n\phi \qquad \text{.....(28)}$$

Squaring and adding eqns. (27) and (28) leads to

$$\left(\frac{E_n}{E_1}\right)^2 = \frac{4}{n^2}\frac{s^2}{\psi^2 + 2\left(\pi\frac{R_s}{R_L}T + U\psi\right)} \qquad \text{.....(29)}$$

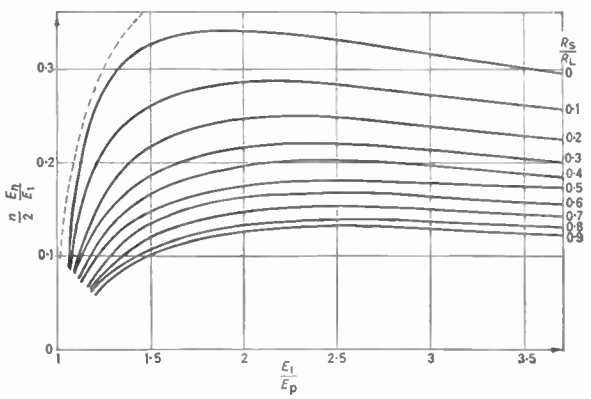


Fig. 4. Output voltage at the nth harmonic as expressed by $\frac{nE_n}{2E_1}$ versus the polarization as $\frac{E_1}{E_p}$ and the diode resistance $\frac{R_s}{R_L}$.

Equation (29) is a simple expression for the output voltage. Computation of $n/2(E_n/E_1)$ as a function of the normalized bias, E_p/E_1, for different values of R_S/R_L and for $X_L = 0$ has been carried out on a IBM 1620 computer. The results are shown in Fig. 4. The dashed line corresponds with eqn. (1). It is evident that the values of E_n/E_1 satisfy the general requirement of eqn. (1). Curves have been obtained in Fig. 7 of Ref. 8 for $n = 2$ with the general solution. They are of similar shape to those shown in Fig. 4, which are for any value of n.

With $X_L = 0$, it is easily verified that eqns. (27) and (28) lead to

$$\tan\phi_n = -\tan n\phi$$

which relates the phase angle ϕ_n of the output voltage to the angle ϕ.

Also, by neglecting E_n/n, a simple expression for the current I_1 at the fundamental frequency can be obtained but will not be given here. Because

$$G_g + jB_g = \frac{2jI_1}{E_1}$$

the following expressions for the components of the input admittance can be written.

$$G_g = \frac{1}{2\pi R_s}[\psi - \tfrac{1}{2}\sin 2\phi + \tfrac{1}{2}\sin 2\phi + 2\sin\theta(\cos\phi - \cos\theta)] \quad ...(30)$$

$$B_g = \frac{1}{2\pi R_s}\left[-\frac{\cos 2\phi + \cos 2\theta}{2} + 2\sin\theta\right] \quad ...(31)$$

For $E_p = 0$, $\theta = 0$ and $\phi = 2\pi$, G_g becomes $1/R_s$, as expected for $E_p = E_1$, $\theta = \phi = \pi/2$, the input conductance vanishes. Thus, absence of polarization gives an input conductance equal to R_s, and $E_p = E_1$ makes the diode into an open-circuit (infinite input resistance). The input susceptance B is $-1/2\pi R_s$ (inductive) for $E_p = 0$, and $1/2\pi R_s$ for $E_p = E_1$. These values are limits and have no physical meaning as eqn. (29), together with eqn. (1), impose further limitations.

The output power P_n at the nth harmonic is given by

$$P_n = \frac{2E_1^2}{n^2}\left[\frac{s^2 R_L}{\psi^2(R_L^2 + X_L^2) + 2\pi^2 R_s^2 + 2\pi R_L R_s \psi}\right] \quad .....(32)$$

This power is proportional to $1/n^2$. It can be seen from this expression that the output power P_n exists if $R_s = 0$. With the aid of eqns. (32) and (26) this output power can easily be calculated. A plot of P_n for $X_L = 0$ could give

$$\frac{n^2}{2} P_n \frac{R_s}{E_1^2} \text{ versus } \frac{R_s}{R_L}$$

with $\sin\theta$ as a parameter.

Equation (32) gives the output power as a function of E_p for the nth harmonic and a given (constant) source voltage E_1. The results have been verified by sweeping the bias voltage and observing the output power. Equation (32) can be written in terms of ψ only, because one can verify that the normalized step can be written as

$$s^2 = \frac{2(1-\cos\psi) - \psi\sin^2\psi}{(\psi - \sin\psi)^2 + (1-\cos\psi)^2} \quad .....(33)$$

Equation (32) shows one maximum for the output power when the bias voltage is swept. Additional maxima in the experimental curve[11] are due to other mechanisms such as the non-linear character of the junction capacitance.

As an example of solution of eqn. (32), Fig. 5 displays

$$\frac{n^2}{2} P_n \frac{R_s}{E_1^2} \text{ versus } \sin\theta$$

(i.e. the normalized bias voltage) for

$$\frac{R_s}{R_L} = \frac{1}{\pi}$$

Equation (32) now reduces to

$$\frac{n^2}{2} P_n \frac{R_s}{E_1} = \frac{s^2}{\pi(\psi + 2)^2}$$

In Fig. 1, for a constant voltage generator at aa', the maximum obtainable output power P_n can be determined from eqn. (32). X_L, R_L and the bias voltage are adjusted for maximum power conversion. The step s in eqn. (32) must be as large as possible. With the aid of eqn. (33), the step can be expressed as a function of the conduction angle only. The optimum working conditions are

$$\frac{\partial P_n}{\partial \psi} = \frac{\partial P_n}{\partial R_L} = \frac{\partial P_n}{\partial X_L} = 0 \quad .....(34, 35, 36)$$

Equation (36) results in $X_L = 0$: we therefore conclude that all reactances at the nth harmonic must be tuned out in the output circuit for maximum output power at the nth harmonic.

The condition $R_s = 0$ has no realistic meaning; R_s will never be zero because of the non-zero value of the series diode resistance. If therefore a non-zero resistance R_s is considered, the optimum resistance from eqn. (35) turns out to be

$$R_L = \frac{2\pi}{\psi} R_s \quad .....(37)$$

This result is of the greatest importance in a practical multiplier. Substitution of eqns. (37) and (33) in the expression for the output power yields

$$P_n = \frac{E_1^2}{4\pi n^2} \cdot \frac{1}{R_s} \cdot \frac{\sin\psi + 2(1-\cos\psi)^2}{\psi[(\psi - \sin\psi)^2 + (1-\cos\psi)^2]} \quad .....(38)$$

The maximum value of P_n is obtained for $\psi = 3{\cdot}9$ rad. This maximum value does not coincide with the conduction angle for maximum step amplitude, which is $\psi = 4{\cdot}1$ rad. From eqn. (26) the bias for maximum output power is given by

$$\sin\theta = \frac{E_p}{E_1} = 0{\cdot}38$$

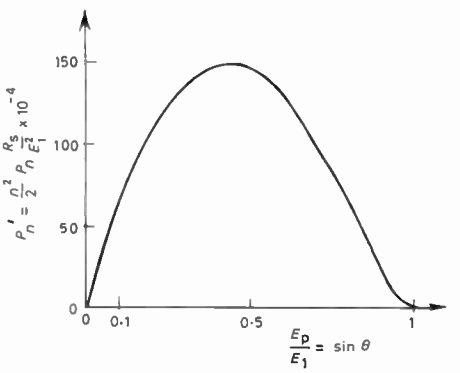


Fig. 5. Output power versus normalized bias.

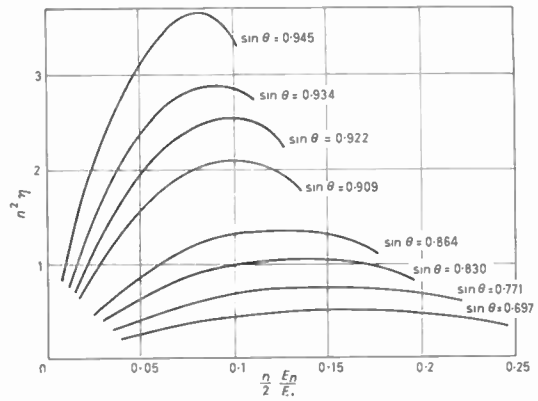


Fig. 6. Power efficiency of the multiplier.

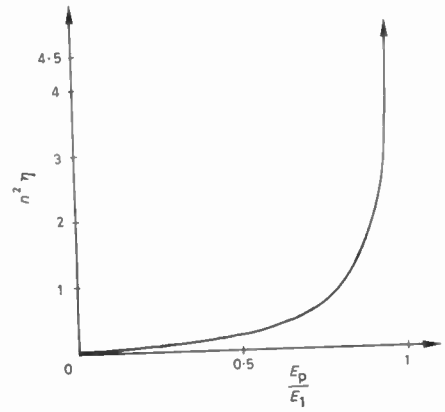


Fig. 7. Power efficiency versus normalized bias.

This result is independent of the harmonic number n. The normalized maximum output power can now be calculated.

$$\frac{n^2}{2} P_n \frac{R_s}{E_1^2} = \frac{1}{2\pi} \cdot 0{\cdot}406 \qquad \text{.....(39)}$$

The maximum output power corresponds with a constant-voltage source at the input aa′ of the multiplier. This can be realized with a generator of negligible internal impedance. If the internal impedance Z_g of the generator is resistive, a transformer and a matching circuit at the input of the multiplier, are used to effect matching.

The optimum input resistance R_g will now be calculated. Substituting ψ, ϕ and θ for $\psi = 3{\cdot}9$ rad in the expression for the input power, namely,

$$P_{in} = \tfrac{1}{2}E_1^2 \frac{1}{2\pi R_s}\left[\psi - \frac{\sin 2\phi}{2} - \frac{\sin 2\theta}{2} + 2\sin\theta\cos\phi\right] \qquad \text{.....(40)}$$

yields P_{in} from which the optimum input resistance can be determined.

$$P_{in} = \frac{1}{4\pi R_s} E_1^2 \times 2{\cdot}41 \qquad \text{.....(41)}$$

The efficiency is given by the ratio of P_n and P_{in}. Thus

$$\eta = \frac{1}{n^2} 0{\cdot}17$$

is the value which corresponds with the maximum output power for the nth harmonic, the input being matched and with $R_L = (2\pi/\psi)R_s$ and $E_p/E_1 = 0{\cdot}38$. This efficiency for the present mode of operation is rather poor. Therefore, the expression for the efficiency must be optimized with respect to the bias voltage.

The efficiency, as defined by eqn. (20), can be written as follows. From eqns. (36) and (37):

$$n^2\eta = \frac{8\pi(\sin\theta - \sin\phi)^2}{\left(\psi^2 \frac{R_L}{R_s} + 4\pi^2 \frac{R_s}{R_L} + 4\pi\psi\right) \times} \times (\psi - \tfrac{1}{2}\sin 2\phi \,.\, \tfrac{1}{2}\sin 2\theta + \sin\theta\cos\phi) \qquad \text{.....(42)}$$

Using eqn. (26) and taking ψ as the independent variable, eqn. (42) can be easily calculated. Figure 6 displays $n^2\eta$ as a function of the normalized output voltage E_n/E_1. The load is optimized with eqn. (37) and the normalized bias $\sin\theta$ is used as a parameter. These curves are limited to the right by condition (1).

The input power can be expressed as a function of the conduction angle only. Using eqn. (37) yields for $n^2\eta$

$$n^2\eta = \frac{s^2}{\psi} \frac{1}{N} \qquad \text{.....(43)}$$

where

$$N = \psi^3 - \sin\psi(2+\cos\psi)\psi^2 + \\ + 2(2\sin^2\psi + \cos\psi - 1) - 2\sin\psi(1-\cos\psi)$$

and s is given by eqn. (33). Values of $n^2\eta$ are now easily calculated for different values of ψ. The corresponding $\sin\theta$ or normalized bias is then calculated from eqn. (26) (see Fig. 7).

It is seen that $n^2\eta$ tends to infinity near $\sin\theta = 1$. Near this point the approximation loses its validity. However, when putting eqns. (37), (29) and (23) in eqn. (1), it turns out that the series-mode of operation exists only if

$$\frac{s}{\psi} \leqslant \cos\theta \quad \text{or} \quad \sin\theta \leqslant 0{\cdot}89$$

The maximum possible power efficiency for a s.r.d. multiplier in the series-mode of operation is therefore given by:

$$\eta \leqslant \frac{15}{n^2}\% \qquad \text{.....(44)}$$

The accuracy of the approximation is

$$\Delta\theta = \frac{1}{n} \times 0{\cdot}45 \text{ rad}$$

For this expression to be valid, n must be high.

The theoretical maximum efficiency with the diode losses, R_s, included is given by eqn. (44) for the ideal step recovery diode in this series-mode of operation with one conduction angle per period and for $n \geqslant 4$.

For a times 10 multiplier, the maximum obtainable efficiency is clearly approximately 15%.

Condition (44) is the ultimate limitation for the efficiency η in harmonic multiplication ($n \geqslant 4$). For $n < 4$, one must go back to eqn. (24) and verify if the approximation on θ is tolerable.

7. Conclusion

The power efficiency in the circuit configuration of Fig. 1 has been shown to be proportional to $1/n^2$ for the s.r.d. multiplier. General explicit formulae have been given. A precise limit for the series-mode of operation with one conduction angle per period has been put forward together with an analysis of some interesting approximations. The graphs resulting from the formulae allow the design of a multiplier with any desired multiplication factor n. It has also been shown that the maximum output does not coincide with the point of maximum efficiency. The most important conclusion is that the efficiency is maximum at the limit of the series-mode of operation, i.e. for the maximum allowable voltage, as can be seen from Fig. 7.

Efficiencies better than those given by eqn. (44) (e.g. see other articles[10, 12]) are due to the capacitance or the non-linear capacitance of the diode which can give an energy transfer during the off-condition. Better efficiencies are probably also obtained by other modes of operation, e.g. the shunt mode.[12]

The efficiency will be lowered by including the effect of a constant capacitance in the reverse direction; this calculation is carried out by Miyakawa[13] for the particular cases when $n = 10$ and $n = 20$.

With the given formulae, it is easy to include the effect of the quality factor of the input and output circuits.

Further work is in progress to include the effect of a capacitance and the effect of multiple conduction angles.

8. Acknowledgment

The authors wish to thank Professor J. Van Bladel, Director of the Laboratory, for his kind permission to publish this paper.

9. References

1. H. Torrey and C. Whitmer, 'Crystal Rectifiers', pp. 111–235, (McGraw-Hill, New York, 1948).
2. C. H. Page, 'Harmonic generation with ideal rectifiers', *Proc. Inst. Radio Engrs*, **46**, p. 1738, October 1958.
3. D. B. Leeson and S. Weinreb, 'Frequency multiplication with nonlinear capacitors', *Proc. I.R.E.*, **47**, pp. 2056–84, December 1959.
4. A. Van der Ziel, 'On the mixing properties of non-linear condensers', *J. Appl. Phys.* **19**, No. 11, pp. 999–1006, November 1948.
5. J. M. Manley and H. E. Rowe, 'Some general properties of nonlinear elements, Part 1: General energy relations', *Proc. I.R.E.*, **44**, No. 7, pp. 904–13, July 1956.
6. J. M. Manley and H. E. Rowe, 'General energy relations in nonlinear reactances', *Proc. I.R.E.*, **47**, No. 12, pp. 2115–6, December 1959 (Letters).
7. Stewart M. Krakauer, 'Harmonic generation, rectification, and life-time evaluation with the step recovery diode', *Proc. I.R.E.*, **50**, No. 7, pp. 1665–76, July 1962.
8. D. L. Hedderly, 'An analysis of a circuit for the generation of high-order harmonics using an ideal nonlinear capacitor', *Trans. I.R.E. on Electron Devices*, **ED-9**, pp. 484–91, November 1962.
9. J. Roulston, 'Frequency multiplication using the charge storage effect: an analysis for high efficiency, high power operation', *Internat. J. Electronics*, **18**, No. 1, pp. 73–86, January 1965.
10. R. D. Hall, 'Step-recovery diodes add snap to frequency multiplication', *Microwaves*, pp. 70–5, September 1965.
11. B. Levine, 'Simplified technique for evaluation diode r.f. performance', *Hewlett-Packard J.*, **18**, No. 4, pp. 12–3, December 1966.
12. S. Hamilton and R. Hall, 'Shunt-mode harmonic generation using step recovery diodes', *Microwave J.*, **10**, pp. 69–78, April 1967.
13. T. Miyakawa, 'An analysis of voltage controlled multipliers with punch-through diodes', *Fujitsu Scientific and Tech. J.*, **3**, No. 2, pp. 43–70, September 1967.
14. W. M. Van Loock, 'Working conditions of a s.r.d. multiplier allowing one conduction angle per cycle', *Proc. Inst. Elect. Electronics Engrs.* (To be published, April 1968), (Letters).

10. Appendix 1

From eqn. (10)

$$I_0 = \frac{\omega}{2\pi}\int_0^T i(t)\,dt = \frac{\omega}{2\pi}\int_{\theta/\omega}^{\phi/\omega} i(t)\,dt = I_1 A_0 + I_n B_0 + I_p C_0$$

$$A_0 = \frac{1}{2\pi}(\cos\theta - \cos\phi) \qquad \text{.....(45)}$$

$$B_0 = -\frac{1}{2\pi n}(\cos n\theta - \cos n\phi) \qquad \text{.....(46)}$$

$$C_0 = \frac{1}{2\pi}(\theta - \phi) \qquad \text{.....(47)}$$

$$I_1 = \frac{\omega}{2\pi}\int_{\theta/\omega}^{\phi/\omega} i(t).e^{-j\omega t}\,dt = I_1 A_1 + I_n B_1 + I_p C_1$$

$$A_1 = \frac{\omega}{2\pi}\left[\frac{1}{2\mathrm{j}}\left(t + \frac{\mathrm{e}^{-2\mathrm{j}\omega t}}{2\mathrm{j}\omega}\right)\right]_{t=\theta/\omega}^{t=\phi/\omega} \quad \text{.....(48)}$$

$$B_1 = -\frac{\omega}{2\pi}\left[\frac{1}{2\mathrm{j}}\left(\frac{\mathrm{e}^{\mathrm{j}[(n-1)\omega t+\phi n]}}{\mathrm{j}(n-1)\omega} - \frac{\mathrm{e}^{-\mathrm{j}[(n+1)\omega t+\phi n]}}{\mathrm{j}(n+1)\omega}\right)\right]_{t=\theta/\omega}^{t=\phi/\omega} \quad \text{...(49)}$$

$$C_1 = \frac{1}{2\pi\mathrm{j}}[\mathrm{e}^{-\mathrm{j}\phi} - \mathrm{e}^{-\mathrm{j}\theta}] \quad \text{.....(50)}$$

$$I_n = \frac{\omega}{2\pi}\int_{\theta/\omega}^{\phi/\omega} i(t)\,\mathrm{e}^{-\mathrm{j}m\omega t}\,\mathrm{d}t = I_1 A_n + I_n B_n + I_p C_n$$

$$A_n = \frac{\omega}{4\pi\mathrm{j}}\left[\frac{\mathrm{e}^{\mathrm{j}(1-n)\omega t}}{\mathrm{j}(1-n)\omega} + \frac{\mathrm{e}^{-\mathrm{j}(1-n)\omega t}}{\mathrm{j}(1+n)\omega}\right]_{t=\theta/\omega}^{t=\phi/\omega} \quad \text{...(51)}$$

$$B_n = -\frac{\omega}{4\pi\mathrm{j}}\left[t\,\mathrm{e}^{\mathrm{j}\phi n} + \frac{\mathrm{e}^{-\mathrm{j}2n\omega t-\phi n}}{\mathrm{j}2n\omega}\right]_{t=\theta/\omega}^{t=\phi/\omega} \quad \text{...(52)}$$

$$C_n = \frac{1}{2\pi n}(\mathrm{e}^{-\mathrm{j}n\phi} - \mathrm{e}^{-\mathrm{j}n\theta}) \quad \text{.....(53)}$$

11. Appendix 2

From eqns. (16) and (17)

$$\frac{2\pi R_s(R_L\cos\phi_n + X_L\sin\phi_n)}{R_L^2+X_L^2}E_n = E_1.F - E_n[\sin\phi_n.K + \cos\phi_n.L] + E_p.M \quad (54)$$

$$\frac{2\pi R_s(-X_L\cos\phi_n + R_L\sin\phi_n)}{R_L^2+X_L^2}E_n = E_1.D - E_n[\sin\phi_n.G + \cos\phi_n.H] + E_p.I \quad \text{...(55)}$$

$$F = \frac{\sin(1-n)\phi - \sin(1-n)\theta}{1-n} - \frac{\sin(1+n)\phi - \sin(1+n)\theta}{1+n}$$

$$K = \frac{1}{2n}(\cos 2n\theta - \cos 2n\phi)$$

$$L = \phi - \theta + \frac{1}{2n}(\sin 2n\theta - \sin 2n\phi)$$

$$M = \frac{2}{n}(\cos n\phi - \cos n\theta)$$

$$G = \phi - \theta + \frac{1}{2n}(\sin 2n\phi - \sin 2n\theta)$$

$$H = \frac{1}{2n}(\cos 2n\theta - \cos 2n\phi)$$

$$I = \frac{2}{n}(\sin n\theta - \sin n\phi)$$

$$D = \frac{1}{n-1}[\cos(1-n)\phi - \cos(1-n)\theta] - \frac{1}{n+1}[\cos(1+n)\phi - \cos(1+n)\theta]$$

12. Appendix 3

The proof of eqn. (1) can be found in work described in another paper.[14]

The equation for $\Delta\theta$ can be derived as follows. Because $\theta = \Delta\theta + \theta'$, eqn. (3) gives

$$E_1\sin\Delta\theta.\cos\theta' + E_1\cos\Delta\theta\sin\theta' = E_p + E_N\sin(N\theta+\phi_N) \quad (56)$$

Equation (23) must be written as

$$\sin\theta' = \frac{E_p}{E_1} \quad \text{.....(57)}$$

For a reasonably small value of $\Delta\theta$, eqns. (56) and (57) lead to:

$$\Delta\theta.E_1\cos\theta' \simeq E_N\sin(N\theta+\phi_N) \quad \text{.....(58)}$$

Using eqn. (57), eqn. (58) can be written as

$$\Delta\theta = \frac{E_N\sin(N\theta+\phi_N)}{\sqrt{E_1^2-E_p^2}} \leqslant \frac{E_N}{\sqrt{E_1^2-E_p^2}}$$

which gives an idea of the maximum absolute value of $\Delta\theta$ as eqn. (23).

Manuscript first received by the Institution on 22nd September 1967 and in final form on 8th February 1968.
(Paper No. 1187/CC6.)

U.D.C. 621.317.76:621.376.3

F.M. Deviation Measurements

By

P. BRODERICK,
B.Sc., A.Inst.P.†

Reprinted from the Proceedings of the Joint I.E.R.E.-I.E.E. Conference on 'R.F. Measurements and Standards' held at the National Physical Laboratory, Teddington, on 14th–16th November 1967.

Summary: The technique used for a frequency modulation measurement depends upon various factors such as the value of modulation index and frequency deviation. Even with the same set of variables more than one method of measurement may be available and the choice of method is decided as much by its ease of application as by its inherent accuracy. Some of the more commonly used methods of frequency modulation measurement are discussed and the conditions under which each is used are examined. The errors arising in the use of these methods and the precautions which must be taken to minimize these errors are also examined. Finally, a commercial type of f.m. meter is described. This instrument depends for standardization on one of the basic methods but covers a wide range of application with good accuracy.

1. Introduction

To understand the reasons for the existence of the various methods of frequency deviation measurement and to examine more easily the limitations of these methods, it is helpful to summarize some of the basic equations leading to the equation for a frequency modulated wave.

Thus, if:

$$e(t) = E\cos\theta(t) \qquad \ldots\ldots(1)$$

Then, $e(t)$ will be a simple sinusoid if $\theta(t)$ is a linear function of time and

$$\frac{d\theta}{dt} = \omega_i$$

where ω_i is the instantaneous angular frequency. However, if the angular frequency is not constant but defined by

$$\omega_i = \omega_c + \Delta\omega(t) \qquad \ldots\ldots(2)$$

where $\Delta\omega(t)$ is the deviation of the angular frequency from the centre frequency ω_c, then $\theta(t)$ is no longer linear and gives rise to frequency modulation.

Thus,

$$\frac{d\theta}{dt} = \omega_c + \Delta\omega(t)$$

$$\theta = \omega_c t + \int \Delta\omega(t)\,dt \qquad \ldots\ldots(3)$$

If now the deviation $\Delta\omega(t)$ is a sinusoidal function of time given by

$$\Delta\omega(t) = \Delta\omega_{max}\cos\omega_m t \qquad \ldots\ldots(4)$$

where ω_m is an audio angular frequency ($2\pi f_m$) and $\Delta\omega_{max}$ is the maximum value of the deviation ($2\pi\Delta f_{max}$) then eqn. (3) becomes

$$\theta = \omega_c t + \frac{\Delta f_{max}}{f_m}\sin\omega_m t$$

and eqn. (1) becomes:

$$e(t) = E\cos\left(\omega_c t + \frac{\Delta f_{max}}{f_m}\sin\omega_m t\right) \qquad \ldots\ldots(5)$$

The quantity $\frac{\Delta f_{max}}{f_m}$ is the modulation index usually called β and Δf_{max} is the peak deviation of the instantaneous frequency from the centre or carrier frequency f_c. The methods of deviation measurement split up generally into two types. One aims to measure directly the maximum frequency difference between the centre frequency and the peak frequency swing on either side of centre. The other uses derivable properties of the modulation index to determine the index for known modulation frequencies whence the peak deviation is deduced.

2. Ranges of Values of Deviation Met in Practice

A brief outline of the ranges of deviation and modulation frequencies used in a variety of applications is shown in Table 1.

Table 1

Frequency deviation range and modulation frequency for different applications

Application	Deviation range	Modulation frequency range
Broadcasting	up to ± 75 kHz	30 Hz–15 kHz
F.m. stereo	up to ± 75 kHz	50 Hz–53 kHz (S.C.A.-75 kHz)
Telemetry	up to 2 MHz	up to 2 MHz
V.h.f. mobile	up to ± 15 kHz	Normal 300 Hz–3 kHz Occasional –6 kHz
Services	up to 30 kHz	300 Hz–3·4 kHz
Multi-channel systems	Single channel test tone ± 280 kHz r.m.s.	300 Hz–3·4 kHz

† Marconi Instruments Ltd., St. Albans, Hertfordshire.

From this it is clear that there is a need for measurements of deviations from as low as a few hundred hertz up to a few hundred kilohertz. In each case there is a wide range of modulation frequencies so that modulation indexes from close to unity to several thousand are used.

The standard direct methods of deviation measurement generally rely for accuracy upon the range of modulation index so that the choice is based on this criterion. Where accuracy is not too important then it can be relevant to argue which is the best method but then the criterion is convenience.

The methods to be discussed here are examined mainly from the accuracy view point and include the use of a spectrum analyser, an oscilloscope, a counter and an indirect method used in an instrument designed to cover a wide range of deviation and modulating frequencies. Using the spectrum analyser the most accurate means of measuring deviation provided the proper precautions are taken is the Bessel zero method and so a more detailed examination of this method has been done.

Similarly in the oscilloscope and counter methods, the method using the counter to measure average deviation could be regarded as the best refinement of the technique of comparing a known frequency with frequencies within the deviation frequency range. Accordingly, this too has been given more detailed examination.

3. Spectrum Analyser Methods

The bandwidth of an f.m. spectrum is given approximately by $W = 2(\Delta f_{max} + f_m)$. This neglects all sidebands less than 1% in amplitude of the unmodulated carrier. The frequency spacing between any two sidebands is given by f_m so that the number of significant sidebands visible on a frequency-time display of width W would be

$$\frac{2(\Delta f_{max} + f_m)}{f_m} = 2(\beta + 1) \qquad \text{.....(6)}$$

A spectrum analyser gives a display of amplitude versus frequency. The frequency is plotted horizontally over a given range—display width—and the amplitude of each frequency component in the range is plotted vertically. If now the individual lines corresponding to sidebands in a wideband spectrum can be displayed with sufficient resolution so that they can be counted, then we have a means of deducing β. Thus if $\Delta f_{max} = 10$ kHz and $f_m = 1$ kHz, a display of 22 kHz will show 22 spectral lines ($\beta = 10$) and the analyser must be set to display clearly 22 separate spectral lines.

The resolution is the ability to separate out two frequency components so that they can be distinguished. If a wide band of frequencies is observed through a filter of bandwidth B then components separated in frequency by less than B will not appear separate at the filter output. Two such responses could be distinguished separately by two filters of bandwidth less than B whose centre frequencies were separated by B. The display on a spectrum analyser is effectively the result of sweeping a narrow band filter (i.f. filter of bandwidth B) across an input spectrum of width W. Here a similar argument applies about the resolution of adjacent frequency components with the added complication that the effect of sweeping reduces the ability to resolve. The resolution gets worse as the sweep speed increases. It may be shown that the resolution of a spectrum analyser is given by:

$$R = B\left[1 + 0{\cdot}195\left(\frac{W}{TB^2}\right)^2\right]^{1/2} \qquad \text{.....(7)}$$

where it is assumed that the i.f. filter response is Gaussian in shape with a 3 dB bandwidth of B. W is the sweep width and T is the time for one sweep. This is not exactly the case in practice but is a good approximation and is useful in indicating the relative values of W, T and B for good resolution in the practical case. It is found in practice that for good resolution

$$\frac{W}{TB^2} \simeq \frac{1}{2} \qquad \text{.....(8)}$$

To clearly resolve two lines separated by f_m the i.f. bandwidth B must be less than f_m. A typical h.f. spectrum analyser would have a choice of values for B, for example 6 Hz or 150 Hz. Hence W/T would be variable from 18 to about 10^4 in eqn. (8).

At low modulating frequencies, e.g. less than 100 Hz, the lower i.f. bandwidth would be selected whereas the larger one would be selected for f_m greater than about 500 Hz. The scan-time T and sweep width W are also limited in practice. Thus for a typical spectrum analyser the maximum value of T would be about 30 s and the maximum W about 30 kHz.

Accordingly with a 6 Hz i.f. the maximum W for good resolution is given by $W/30 = 18$, i.e. $W = 540$ Hz. With $f_m = 50$ Hz ($\beta = 4{\cdot}4$) the number of visible lines would be 10. With 150 Hz i.f. and $W_{max} = 30$ kHz good resolution would be obtainable with a time $T = 3$ s and the number of lines clearly visible with $f_m = 500$ Hz ($\beta = 29$) would be 60.

Clearly it is a tedious procedure having to count numerous lines and one would only use the method as a quick check on β with a limit of about 24. Having limited the range of β on practical grounds of sweep-time for adequate resolution and number of lines it is reasonable to count, the trouble arises in deciding which lines are less than 1% of the unmodulated carrier. As β increases the error due to this uncertainty

decreases but for the range of values of β to which we have just restricted ourselves, the error is unlikely to be less than $\pm f_m$, i.e. $\pm 5\%$. As β increases to extremely large values then for a Δf within the range of the instrument the spectrum takes on a definite contour as all the lines merge together. This can also be used to get an idea of Δf but is not really any more accurate than the former method.

4. Bessel Zero Method

If we take eqn. (5) as the general equation for a frequency-modulated wave and expand it into a series we get:

$$E = J_0(\beta)\cos\omega_c t + J_1(\beta)[\cos(\omega_c+\omega_m)t - \\ -\cos(\omega_c-\omega_m)t] + J_2(\beta)[\cos(\omega_c+2\omega_m)t + \\ +\cos(\omega_c-2\omega_m)t] + \ldots \quad \ldots\ldots(9)$$

where the coefficient $J_n(\beta)$ is the Bessel function of the first kind and nth order with argument β. A plot of the first three coefficients is shown in Fig. 1. From this the values of β at which each coefficient is zero can be read. In particular, the amplitude of the carrier $J_0(\beta)$ is zero at β equal to 2·405; 5·52; 8·65; 11·79; etc. If at any one of these points the modulating frequency f_m is known, the deviation is calculable.

Usually the method is used to set up a given deviation which can be used as a reference to standardize, for example, a deviation monitor. In this case the f.m. spectrum is displayed using a narrow sweep width (enough to see the carrier and first-order sidebands only) and fixed modulation frequency. The deviation is then varied by varying the amplitude of the modulating signal until the deflection of the centre frequency first goes to zero, whence $\beta = 2{\cdot}405$. As the deviation is increased the carrier frequency deflection reappears and again vanishes at the deviation appropriate to a modulation index of 5·52 and so on for other zero positions.

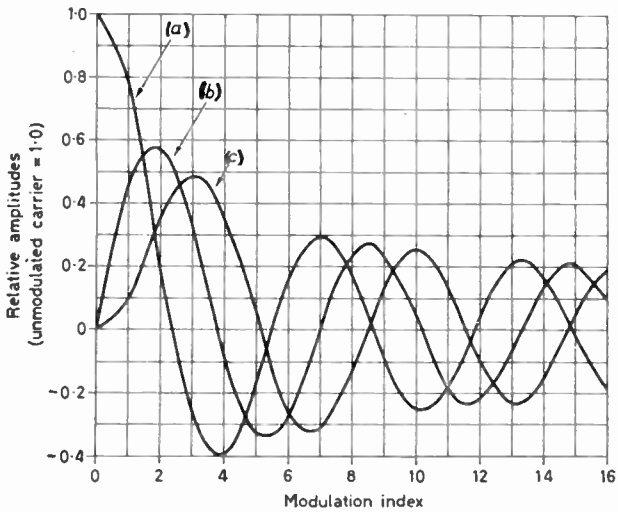


Fig. 1. Amplitude of frequency components of an f.m. wave.
(a) Carrier
(b) First-order sideband components
(c) Second-order sideband components

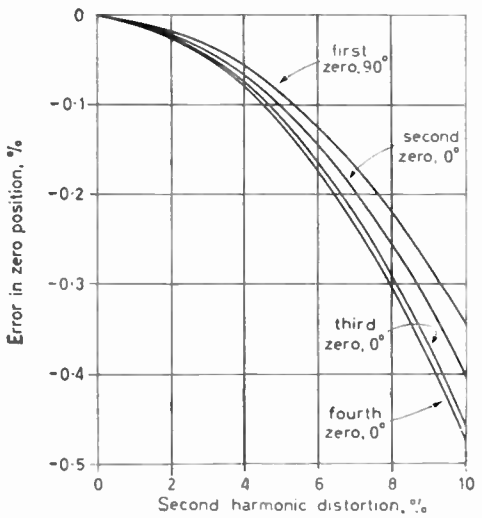


Fig. 2. Shift of carrier minimum for second-harmonic distortion at the worst-case phase angle.

4.1. *Distortion Effects*

If the modulating signal is not a pure sinusoid and/or there is non-linearity in the modulating process the f.m. spectrum will be distorted and the effect of this will be to alter the values of β at which the carrier amplitude is zero. The result of applying a multi-tone modulation frequency is to produce a much more complicated spectrum. The coefficients of each sideband term are no longer single Bessel functions but are series of products of Bessel functions. The problem of identifying where a carrier zero occurs, if indeed one exists is, therefore, more involved.

It has been shown[1] that if the distortion is second harmonic the carrier component splits up into in-phase and quadrature parts with the result that a true zero of carrier amplitude is not obtainable though a minimum exists. Figure 2 shows the error in the position of the minimum for second harmonic distortion at those phase angles which produce the largest shift in position (with respect to the zero position for no distortion) for the first four zeros. Even for 10% second harmonic distortion the percentage shift in all carrier zeros is less than $-0{\cdot}5\%$.

Odd harmonic distortion is also shown to produce a change in the value of β at which the carrier amplitude is zero. Figure 3 shows the results for percentages

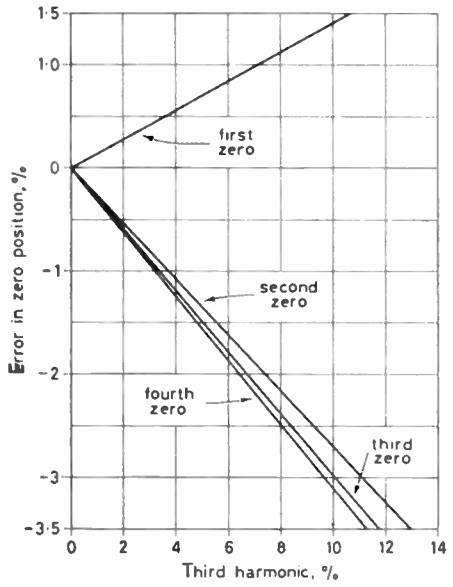


Fig. 3. Shift in carrier zero for third-harmonic distortion at $\Delta\phi_3 = \pi$.

of third harmonic up to 14%, again for the 'worst case' phase angle for carrier zeros up to the fourth. The error in the zero position for all zeros levels off to between 3% and 3½% for 10% distortion.

Spurious amplitude modulation in a generator is usually related to the modulation frequency by some low harmonic number being generated often by variation in the Q-factor of the tuned circuit as it is loaded by, for example, a variable capacitance diode. The degree of spurious amplitude modulation depends upon the frequency deviation.

Spurious a.m. also produces errors or shifts in the positions of the zeros or minima from the undistorted case. Figures 4 and 5 show the results of spurious a.m. at the modulating frequency and at its second harmonic, these being the common cases. The errors shown are those for the 'worst case' phase angles.

Frequency or amplitude modulation distortion by higher harmonics generally produces smaller errors. The greatest error from all the effects mentioned is due to a.m. at the second harmonic of the modulating frequency and is less than 4% for 10% distortion.

In practice, distortions like 10% would not be tolerated if the generator was being used as a means of setting up an f.m. spectrum with a known deviation for standardizing purposes. However, the results do indicate that if significant distortions do exist these will hardly be noticed in the measurement of a zero whereas on a peak deviation monitor the distortion could be immediately apparent if the phasing of the distortion happened to directly affect the peak. Clearly, the use of a reasonably pure modulating tone and good modulator will minimize any ambiguity of this kind.

Allowing for the need of a good quality modulating signal and modulator, the method is a convenient one in that zeros are easily located with adequate sensitivity—as the carrier amplitude is decreased nearer zero the i.f. gain of the analyser can be increased for increased sensitivity. The requirement of a local oscillator with high stability as in other methods is also removed. The method is suitable for the measurement of very low modulation index, e.g. 2·405; however, for high modulation index the order of carrier zero required gets inconveniently large and zeros above about 8 (corresponding to $\beta = 24{\cdot}35$)

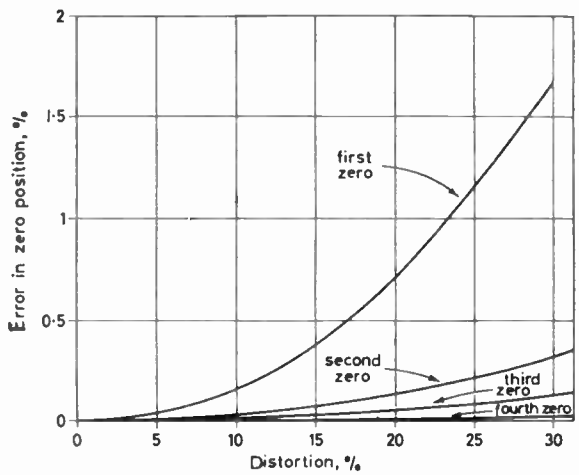


Fig. 4. Shift in carrier zero for spurious a.m. distortion at the modulating frequency.

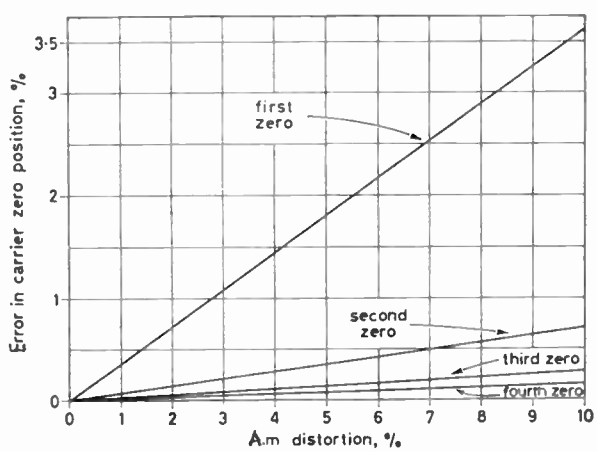


Fig. 5. Shift in carrier zero for spurious second-harmonic a.m. distortion.

are seldom used. At higher modulation indexes zeros of the first-order sidebands can be located since these will occur at higher values of β for the same order carrier zero. In any event a modulation index about 24 is usually considered the useful range for this method. Spectrum analysers are available covering all bands from l.f. and v.h.f. up to microwaves so that the method is useful in all ranges corresponding to the applications listed in Table 1.

5. Oscilloscope Methods

There are several variations in the systems using oscilloscopes to measure frequency deviation but basically they all consist in frequency changing the output of a frequency-modulated source using a local oscillator tunable over the full deviation range and locating on a display the position of zero i.f. Thus in Fig. 6 the signal generator output is mixed with the output of a stable oscillator and the resultant spectrum fed into an oscilloscope. The signal generator is frequency modulated using an external oscillator which is also used to provide the time-base on the oscilloscope. The resultant display for the case where f_{LO} is tuned to f_c is shown in the figure.

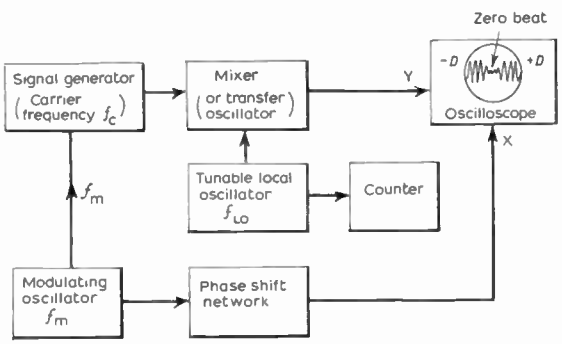


Fig. 6. Oscilloscope method.

The discrimination of a null at either end of the trace, in practice, is never better than about 2% and gets progressively worse as the deviation increases for a constant modulation frequency. A decrease in deviation or increase in modulation frequency results in there being fewer cycles occupying the scan of the time-base. The null therefore becomes shallower and occupies a greater horizontal distance. Clearly as the modulation index tends to unity the null will occupy the whole scan and the measurement of deviation will be grossly in error. In practice, high modulation frequencies are undesirable anyway because of the problem of phase correction, the absence of which causes the null to fluctuate in position as the peak of the deviation and modulation frequency drift in and out of phase.

For reasonably quantitative measurements a modulation frequency less than 1 kHz should be used with minimum deviations of greater than 10 kHz. For more accurate results using a stable oscillator and a counter, modulation frequencies of the order of 100 Hz are necessary.

The procedure is to tune f_{LO} over the full deviation range $\pm D$. When it is tuned to either end of the range the zero beat will appear on the screen at either end of the trace. The difference in the settings of the local oscillator at these positions gives the peak-to-peak deviation. As a refinement to the method, the local oscillator output can be monitored on a counter and the difference in counts at each end of the deviation range measured.

In one variation of the method the mixer and local oscillator are replaced by a single transfer oscillator whose output is the required frequency-changed spectrum. The stability of the local oscillator in this case is not as good as can be obtained by selecting an independent oscillator with good stability. Furthermore, the use of a mixer can provide a bandwidth extending to d.c. on its output so that, in principle at least, the measurement of low deviation is possible which is not the case where a transfer oscillator is used with a l.f. cut-off characteristic. Any l.f. cut-off above d.c. provides an uncertainty in the position of a null or zero beat.

A further practical point in this system is the need for a phase-adjusting network to ensure that the peaks of the deviation as seen on the oscilloscope occur at positions corresponding to the peaks of the modulating signal.

5.1. *Limitations*

The method is limited to the measurement of large deviations—the limit being imposed by both oscillator stability and the discrimination in positioning the null. If during the time the local oscillator is tuned from one end of the deviation range to the other and the null located, the oscillator frequency or centre frequency of the generator drifts, then there is an uncertainty in the measurement of peak deviation. For example, if the oscillator drifts by 1 part in 10^6 in this time, then if its actual frequency is 100 MHz the uncertainty in the measured deviation is 100 Hz, so that deviations below 10 kHz would be impracticable in accurate measurements.

One other limitation of the method is its insensitivity to any distortion in the modulating waveform or asymmetry in the resultant f.m. spectrum. Such asymmetry can result in the positive peak being different from the negative peak and a direct measurement of peak-to-peak deviation will not show this. If the unmodulated carrier frequency is available from

the generator, it can be used as the local oscillator to locate the position of the centre frequency on the screen. The difference between this point and either end of the trace will indicate any difference between positive and negative peak deviations.

6. Counter Averaging Method

In the method shown in Fig. 7 the average deviation is measured on a counter and converted to the peak deviation.[2] A variation in the method enables peak deviation also to be measured directly.

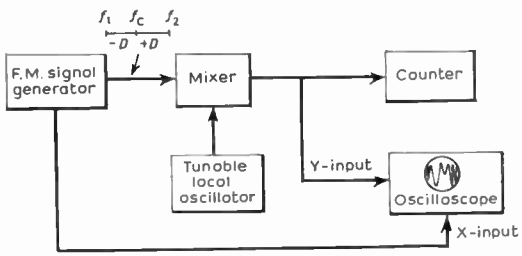


Fig. 7. Counter averaging method.

The signal from the generator can be represented as in Fig. 7 where f_1 and f_2 are the extremities of the frequency shift impressed on the carrier f_c. The difference or deviation D is proportional to the amplitude of the modulating signal f_m. The rate of variation from f_1 to f_2 is, of course, at the frequency of f_m.

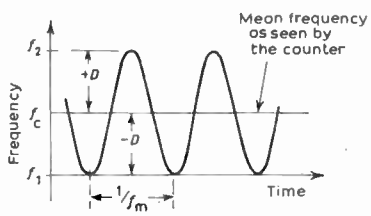
This spectrum is mixed with the local oscillator frequency which is tunable over the full deviation range. The resultant spectrum depends on the frequency difference between the local oscillator and the signal generator output spectrum. If the local oscillator is tuned to the carrier frequency f_c the mixer output will be an f.m. spectrum centred about 0 Hz so that the frequency increases from 0 to $\pm D$. If the local oscillator is tuned to f_1 or f_2 the mixer output will vary from 0 to $\pm 2D$. Figures 8(a) and 8(b) are frequency/time sketches of the mixer output for the two cases mentioned.

The counter will indicate the average frequency in the mixer output. The results of frequency changing the f.m. spectrum in the two cases mentioned and shown in Figs. 8(a) and 8(b) is to give average readings on the counter of $2D/\pi$ and D respectively. In the case of Fig. 8(a) the counter will display the full-wave average of the input since it will not distinguish between positive and negative frequencies. The peak deviation D can then be obtained by multiplying the displayed value by $\pi/2$ assuming the variation of input frequency is sinuosidal.

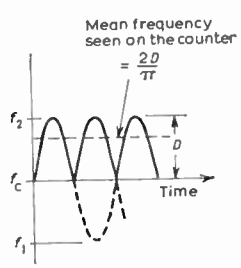
The oscilloscope is used to position the local oscillator frequency within the range f_1 to f_2. If the time-base is provided by the modulating frequency then the display will show a null point positioned in the trace dependent upon the setting of the local oscillator frequency in the range f_1 to f_2. A null at either end of the trace ($f_{LO} = f_1$ or f_2) is easy to discriminate when the deviation is large and a large number of cycles occupies one scan of the time-base. For low deviations it is easier to set the null point in the centre of the display. In the latter case the counter will read minimum.

Although the display is a convenient means of locating the correct frequency of the local oscillator the counter itself can be used to do this. Thus, if the local oscillator is tuned outside the deviation range the counter will read the difference frequency between the oscillator and f_c. If now the oscillator frequency is moved towards the carrier, the counter reading will decrease by the same amount linearly until the local oscillator is equal to the peak deviation, after which changes in the local oscillator frequency do not result in equal changes in the counter reading.

In this method the counter must be able to count at the frequency of the peak deviation and its sampling time must be long compared to the modulating frequency. At high frequencies it is clear that the stability of the local oscillator is all-important if an independent oscillator is used, i.e. one not derived from the unmodulated carrier. If the modulating frequency is distorted or the output spectrum of the modulator is asymmetrical then a different interpretation of the counter display must be given if error is



(a) Local oscillator tuned to f_c.



(b) Local oscillator tuned to f_1.

Fig. 8. Frequency/time for mixer output.

to be avoided. Finally, the question arises concerning the effect of a counter sampling time not very long compared to the lowest modulating frequency.

6.1. *Effects of Distortion or Asymmetry of Spectrum on Indicated Deviation*

6.1.1. Odd harmonics

Error in the measurement of deviation due to odd harmonics depends upon the setting of the null in the deviation range. If one tunes the oscillator to either end of the deviation range the counter will still record the mean deviation which in this case will also be the peak deviation. This is because each half-cycle is affected the same way and the mean value is unaltered.

If the local oscillator is tuned to the carrier frequency which is the usual case, then error will result if the multiplying factor mentioned above, namely $\pi/2$, is used to evaluate the peak deviation. The magnitude of the error will depend upon the amount of distortion and its phase relationship with respect to the fundamental. Generally, one does not know the amount of distortion present in the spectrum, nor is the phase relationship known. Accordingly, a useful procedure is to evaluate the 'worst case' error. This will relate the error to the percentage distortion for the particular phase angle which gives the greatest error. In practice an operator will read the mean value on the counter and deduce the peak value by multiplying by $\pi/2$. Clearly the phase angle for the distortion which gives the greatest peak to mean ratio will result in the largest error. Figure 9 shows the case of third harmonic distortion where the phase angle between the harmonic and the fundamental is π.

In this case the resultant mean for the first half cycle will be less than that for the undistorted waveform because of the extra half cycle of harmonic below the axis in this region. The peak, however, will be increased and the ratio of peak to mean will be a maximum.

Considering Fig. 9:

$$\text{Resultant mean} = \frac{2}{\pi}\left(a_1 - \frac{a_2}{3}\right)$$

$$\text{Resultant peak} = a_1 + a_2$$

$$\text{Assumed peak} = \frac{\pi}{2} \times \text{resultant mean}$$

Therefore, error in estimate of peak

$$= \frac{(a_1+a_2) - \left(a_1 - \frac{a_2}{3}\right)}{a_1+a_2}$$

$$= \frac{4a_2}{3(a_1+a_2)}$$

If $a_2 = ka_1$, then percentage error $= \dfrac{400k}{3(1+k)}$.

For example, for 5% distortion

$$\% \text{ error} = 6{\cdot}3\%$$

The above results will apply for percentage distortions up to that which causes extra crossovers on the time axis. Thus, in Fig. 9 if $a_2 > \frac{1}{3}a_1$ the slope of the harmonic at 0 and π would exceed that of the fundamental so that the resultant waveform would have crossovers between 0 and π. Since the counter treats positive and negative frequencies similarly, it effectively rectifies the resultant waveform and the effect on the mean due to the extra positive lobes so created would need to be evaluated after locating their position as a function of the amount of distortion. In practice, however, the percentage distortion will be very much less than $33\frac{1}{3}\%$ so this complication does not arise.

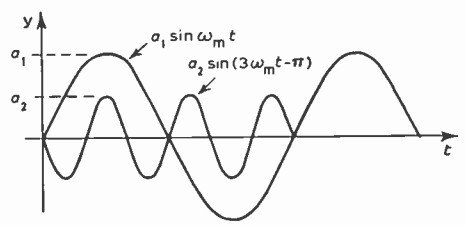


Fig. 9. Third harmonic distortion. Phase angle between harmonic and fundamental is π radians.

In the case of higher odd harmonics similar effects can be shown in that the effect on the mean is the result of an extra positive or negative half-cycle of the harmonic. However, extra half-cycles of higher-order harmonics occupy less area than lower harmonics and have less effect on the average value. Thus the effect on the average over 0 to π due to the area under one half-cycle of the third harmonic $= 2a_2/3\pi$, whereas for one half-cycle of fifth harmonic it is $2a_2/5\pi$, i.e. the effect on the mean varies inversely as the degree of the harmonic. However, the effect on the peak is the same for the same percentage distortion. Consequently the effect on the peak to mean ratio will get less compared with third harmonic. The criterion regarding crossovers still applies here in that there will be no extra crossovers in 0 to π provided $a_n > (1/n)a_1$. For the situation where there are a number of harmonics together as in

$$e = a_1 \sin\theta_1 + a_2 \sin 2\theta_2 + a_3 \sin 3\theta_3 + \ldots$$

No crossovers between 0 and π occur provided

$$a_1 \geqslant 2a_2 + 3a_3 + \ldots$$

Figure 12 shows the resultant errors in the estimate of peak deviation for different percentages of third harmonic distortion in the 'worst case'.

6.1.2. Even harmonics

If the local oscillator is tuned to f_1 and f_2 then, depending upon the phase angle of the distortion, the answers in the two cases will be different. For zero phase angle between the harmonic and fundamental, the mean values seen on the counter will equal the mean deviation and also the peak deviation. If the phase angle is not zero this is not the case as can be seen for example in Fig. 10 (for $\pi/2$ phase angle). Here the mean value is less than half the peak-to-peak value for the case where the oscillator is tuned to f_1. Tuning to f_2 would give a mean greater than half the peak-to-peak value. The peak deviation referred to f_c would be the average of two such results. All these considerations also apply to higher-order even harmonics.

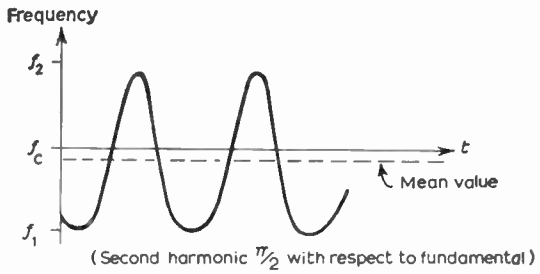


Fig. 10. Second harmonic $\pi/2$ with respect to fundamental.

If the local oscillator is tuned to f_c the problem of relating the mean to the peak arises. Again the error will be largest where there is the greatest change in the ratio of peak to mean value. When the fundamental and harmonic are in phase and provided that the harmonic is small enough not to result in any extra crossings of the time-axis other than 0, etc., then there is no change in the mean value as seen by the counter. This will apply for second harmonic distortion up to 50% and fourth harmonic up to 25%.

It may be shown that the greatest change in the ratio of peak to mean for, say, second harmonic, occurs at the phase angle where the peaks of the fundamental and harmonic coincide; Fig. 11(a) shows this case. Here the peak value of the first half-cycle is increased by the sum of the fundamental and distortion whereas in the second half-cycle it is decreased. Clearly, the peak-to-peak has not altered but the waveform as seen by the counter, namely Fig. 11(c), has altered such that the peak to mean value is different. The mean frequency seen by the counter is the mean value of Fig. 11(c) which is readily

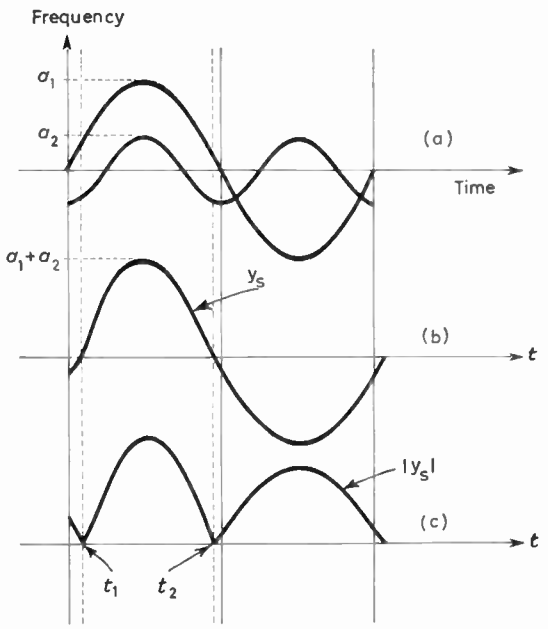


Fig. 11. Fundamental with second harmonic at the 'worst-case' phase angle.

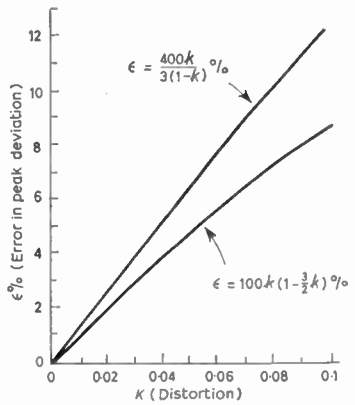


Fig. 12. 'Worst-case' errors due to 2nd and 3rd harmonic distortions.

evaluated when the intersections of the time axis are deduced for each percentage distortion. Appendix 1 shows that the percentage error for this type of distortion is given generally by

$$\varepsilon = k(1-\tfrac{3}{2}k)100\% \qquad \text{.....(10)}$$

where k = ratio of harmonic to fundamental amplitudes.

Figure 12 shows a plot of the 'worst case' error in the measurement of peak deviation for second harmonic distortion up to 10%. It also shows, for comparison, the error when using the same method

where third harmonic distortion is present. Clearly, third harmonic distortion has a larger effect on the measured value.

6.1.3. Higher even harmonics

The fourth harmonic causes no error in the measurement of peak deviation if, as in the case of second harmonic, the phase angle is zero and provided the magnitude is less than 25% which is the case in practice. If the harmonic is out of phase, error will arise similar to that in the case of second harmonic. The 'worst case' condition is the same as before and the maximum error is the same as for the same amount of second harmonic. This applies for all significant even harmonics.

6.2. *Counter Sampling-time Comparable with Lowest Modulation Frequency*

The effect of this is significant at very low modulation frequencies when using a counter with a short sampling-time. Even then, repeated samples can be averaged with increasing accuracy in the final results. It is shown in Appendix 2 that for less than 1% error in the measurement of mean deviation in the absence of any distortion, the modulating frequency and the counter sampling time are related by:

$$T \geqslant 5/f_m \qquad \text{.....(11)}$$

6.3. *Range of Modulation Index*

The signal at the input of the counter will be an f.m. spectrum with zero centre frequency in the case where the local oscillator is tuned to the carrier frequency. When the deviation is very much greater than the modulating frequency there will be many cycles of the f.m. signal occupying the peak deviation range. The counter records a count for each positive excursion of the trigger level. The average number of excursions in the sampling time is then displayed as the mean deviation. As the modulating frequency increases the number of cycles within one peak deviation sweep get fewer, such that, in the limit for a modulation index of 1 no complete cycle of any frequency within deviation range 0 to the peak deviation frequency can occupy one sweep of the modulating frequency from its zero to peak amplitude. The picture as seen on the oscilloscope in this condition is one where the waveform is an indiscriminate null occupying the full scan. Clearly, the counter will read in error in this case. The effect of low modulation index is considered in Appendix 3 and Fig. 13 shows a plot of maximum error related to modulation index using the results obtained in the Appendix. Clearly, for even 3% certainty the method is not applicable for modulation indexes < 50.

Summarizing on this method then it appears that fairly accurate determination of frequency deviation

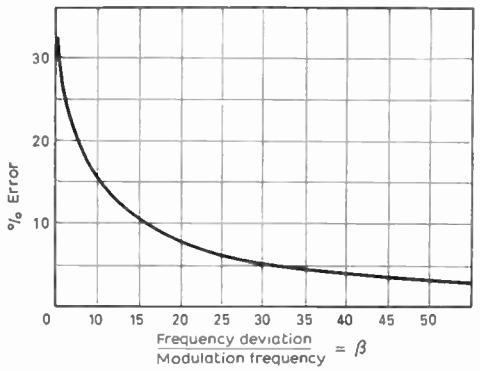


Fig. 13. Variation of maximum percentage error related to modulation index.

is possible for modulation indexes greater than 50 provided distortion is not significant. As far as distortion is concerned, odd harmonic distortion of the deviation waveform is worse than even harmonic and the effect decreases as the order increases. Tuning the local oscillator to the centre frequency is more convenient and is the usual procedure, in which case the counter reads a minimum and the scope, though a useful monitor anyway, is unnecessary. At high carrier frequencies and moderate deviations a highly stable local oscillator is necessary and in all cases a low-pass filter is necessary in the mixer output to prevent direct interference of carrier and local oscillator on the counter reading both of which can be inconvenient.

7. Discriminator Method

The principle here is the conversion of frequency change into voltage change in a circuit with a linear input frequency/output voltage characteristic. The usual discriminator response curve is S-shaped and the middle quasi-linear region is designed to extend over the whole peak-to-peak frequency deviation range. The carrier frequency is set at the centre of the characteristic, which is the zero output voltage position; and frequency excursions each way produce positive or negative voltages depending upon the direction of frequency change. Provided the linear region is greater than the peak-to-peak deviation the output voltage is directly proportional to input frequency over all the deviation range.

Various types of discriminator circuits have been devised. A common one which is the basis of a commercial deviation meter[3] uses a pulse count discriminator, as is shown in Fig. 14. The input is wideband, typically a few MHz to about 1 GHz, and is frequency changed to an i.f. near the low-frequency end of the band. The i.f. after amplification is limited to eliminate a.m. and provides the input

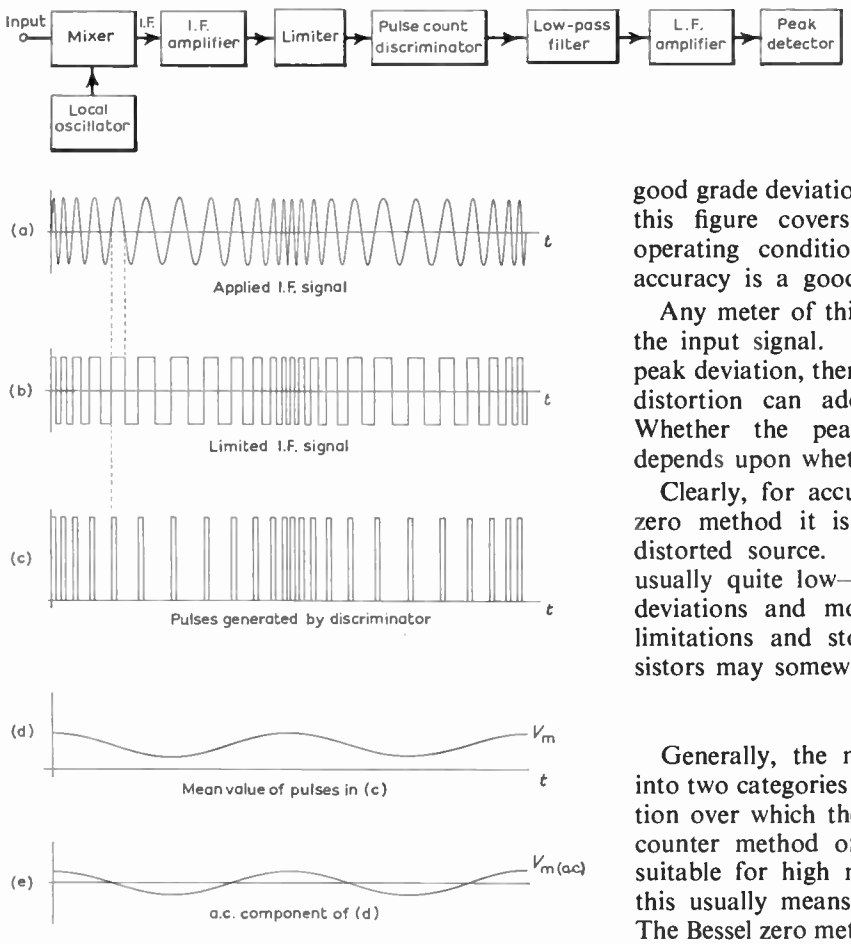


Fig. 14. Pulse count discriminator type deviation meter.

to the pulse count discriminator. This is merely a Schmitt trigger circuit whose constant rise-time output is differentiated and used to drive a pulse generator. The resultant pulses are filtered to remove h.f. components and the output peak detected. The mean value of the pulses varies as the p.r.f. varies with the deviation, and the detected alternating component of the resultant waveform is proportional to the peak deviation.

Since the output amplitude is dependent on the spacing between pulses, the response curve unlike the conventional discriminator response curve, is intrinsically linear. This will apply over an extremely wide frequency range—in fact up to where the pulses interfere with each other.

The method being indirect relies for standardization upon one of the basic methods already described—usually the Bessel zero method. It can by such calibration be used over a wide range of deviation, e.g. 5 kHz to 500 kHz, with modulating frequencies from about 30 Hz to 150 kHz. The accuracy of a good grade deviation meter is about $\pm 3\%$. However, this figure covers extremes of performance and operating conditions and for most purposes the accuracy is a good deal better.

Any meter of this type is affected by distortion in the input signal. Since the reading is the detected peak deviation, then in the 'worst case' the percentage distortion can add or subtract from the peaks. Whether the peak-to-peak deviation is affected depends upon whether the distortion is even or odd.

Clearly, for accurate calibration using the Bessel zero method it is important to have a very low distorted source. Distortion in the meter itself is usually quite low—typically 0·1%—though at high deviations and modulation frequencies, bandwidth limitations and storage-time in discriminator transistors may somewhat increase the distortion.

8. Conclusions

Generally, the methods considered may be split into two categories according to the range of modulation over which they are intrinsically accurate. The counter method of measuring average deviation is suitable for high modulation index and in practice this usually means deviations greater than 10 kHz. The Bessel zero method is suitable for low modulation index and can be used over a wider deviation range from below 1 kHz. The counter and oscilloscope methods require a highly stable local oscillator since the measured deviation is directly affected by drift. These methods require no control over the source and can therefore be used for measuring unknown deviations, as for example radiated signals. In the Bessel zero method the input frequency is changed usually with a crystal oscillator and drift on the final tuned oscillator is less significant. Even then unless this drift is so bad that the position of the null on the display drifts right off the screen during the time of one measurement, it will not affect the accuracy of the measurement. Since control over the source is necessary this is more suitable for setting desired deviations. With a good modulating signal this method is perhaps the most accurate as a standard over a wide range of deviations where the modulation index is low.

The pulse count discriminator method is indirect and is the basis of a commercial instrument. It can be calibrated accurately to cover a wide range of modulation index and provides a convenient and accurate monitor of deviation for most applications.

9. References

1. P. Broderick, 'Effect of distortion on the Bessel-zero method of frequency-deviation measurement', *Proc. Instn Elect. Engrs*, **113**, No. 5, p. 740, May 1966.
2. I. Godier and P. S. Christensen, 'New method of measuring f.m. deviation uses electronic counter', *Canadian Electronics Engng*, **6**, No. 7, p. 38, July 1962.
3. V. F. Arnold, 'F.m./a.m. frequency meter', *Marconi Instrumentation*, **10**, No. 3, p. 34, December 1965.

10. Appendix 1

The resultant waveform shown in Fig. 11(b) is given by

$$y_s = a_1 \sin\theta + a_2 \sin(2\theta - \pi/2) \quad \text{where } \theta = 2\pi f_m t$$
$$= a_1(\sin\theta - k\cos 2\theta) \quad \text{where } a_2 = ka_1$$

The intersections with the time axis in Fig. 11(c), namely, t_1 and t_2, are given by:

$$\sin\theta_{1,2} = \frac{-1}{4k} \pm \frac{(1+8k^2)^{1/2}}{4k}$$

where $\theta_{1,2} = 2\pi f_m t_{1,2}$.

For small values of θ, $\sin\theta \simeq \theta$ rad. This applies for θ near to zero which is the case for reasonably low distortion

$$\theta = -\frac{1}{4k}[(1 \mp (1+8k^2)^{1/2})]$$
$$= k \text{ (taking the positive solution)}$$

Therefore,

$$\theta_1 = k$$
$$\theta_2 = \pi - k$$

The average value of y_s is given by:

$$\text{Av}\,|y_s|_{0\to\pi} = \frac{1}{\pi}\left[\int_{\theta_1}^{\pi-\theta_1} y_s\,dy_s - 2\int_0^{\theta_1} y_s\,dy_s\right]$$
$$= -\frac{a_1}{\pi}\left[\cos\theta + \frac{k}{2}\sin 2\theta\right]_k^{\pi-k} + \frac{2a_1}{\pi}\left[\cos\theta + \frac{k}{2}\sin 2\theta\right]_0^{k}$$
$$= \frac{2a_1}{\pi}[2\cos k + k\sin 2k - 1]$$

also

$$\text{Av}\,|y_s|_{\pi\to 2\pi} = \frac{2a_1}{\pi}$$

Therefore,

$$\text{Av}\,|y_s|_{0\to 2\pi} = \frac{a_1}{\pi}[2\cos k + k\sin 2k]$$
$$= \frac{2a_1}{\pi}\cos k[1 + k\sin k]$$
$$\simeq \frac{2a_1}{\pi}(1-k^2)^{1/2}(1+k^2)$$

i.e.

$$\text{Av}\,|y_s|_{0\to 2\pi} \simeq \frac{2a_1}{\pi}\left(1 + \frac{k^2}{2}\right) = \text{measured mean value}$$

Also

$$\text{Peak}\,|y_s| = a_1(1+k) = \text{true peak}$$

$$\text{Deduced peak} = \frac{\pi}{2} \times \text{mean} = a_1\left(1 + \frac{k^2}{2}\right)$$

% Error in deduced peak

$$= \frac{a_1(1+k) - \left(1 + \frac{k^2}{2}\right)a_1}{a_1(1+k)} 100$$
$$\varepsilon \simeq 100k(1 - \tfrac{3}{2}k)\%$$

Thus we obtain eqn. (10) mentioned in Sect. 6.1.2.

11. Appendix 2

Consider Fig. 15 where T is the counter sampling time which covers $(n+k)$ half-cycles of the modulating frequency f_m. The mean value of the 'effectively full-wave rectified' waveform is given by:

$$M = \frac{n}{\pi(n+k)}\int_0^{\pi} a\sin\omega t\,d(\omega t) + \frac{1}{\pi(n+k)}\int_0^{k\pi} a\sin\omega t\,d(\omega t)$$
$$= \frac{a}{\pi(n+k)}[2n + 1 - \cos k\pi] \qquad \text{.....(12)}$$

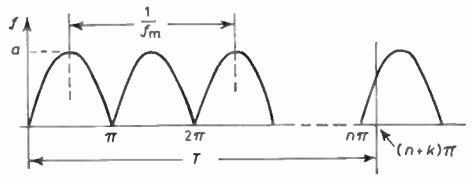


Fig. 15. Effective input signal to counter with a sampling-time T.

Error in the mean value occurs when the sampling time is such that k lies between 0 and $\frac{1}{2}$. The greatest error will be at that value of k which makes M a minimum (a maximum for k between $\frac{1}{2}$ and 1).

Hence

$$\frac{dM}{dk} = 0 \quad \text{when } (n+k)\pi\sin k\pi + \cos k\pi = 2n+1$$

and

$$n = \frac{k\pi\sin k\pi + \cos k\pi - 1}{2 - \pi\sin k\pi} \qquad \text{.....(13)}$$

As $n \to 1$, it may be shown that $k\pi \to 0{\cdot}62$ rad

As $n \to \infty$, $\sin k\pi \to \frac{2}{\pi}$ and $k\pi \to 0{\cdot}69$ rad

Also, when $k\pi = 0{\cdot}69$ rad,

$$k\pi \sin k\pi + \cos k\pi - 1 = 0{\cdot}210$$

In fact, for $n > 10$,

$$k\pi \sin k\pi + \cos k\pi - 1 \simeq 0{\cdot}21$$

whence

$$\pi \sin k\pi = 2 - \frac{0{\cdot}21}{n} \qquad \text{.....(14)}$$

Now, $(M)_{\min} = a \sin k\pi$ (from eqns. (12) and (13)) and the value of M when $n \to \infty$ is

$$\frac{2a}{\pi} = M'$$

$$\text{Error in the value of } M = \frac{M' - (M)_{\min}}{M'}$$

$$= \left(\frac{2a}{\pi} - a \sin k\pi\right) \Big/ (2a/\pi)$$

Using eqn. (14) we get

$$\% \text{ Error in value of } M = \frac{\frac{2a}{\pi} - \frac{a}{\pi}\left(2 - \frac{0{\cdot}21}{n}\right)}{\frac{2a}{\pi}} \times 100\%$$

$$\text{Error} = \frac{10{\cdot}5}{n}\% \qquad \text{.....(15)}$$

Hence for less than 1 % error, $n > 11$ and

$$T > \frac{5}{f_m}$$

Thus we obtain eqn. (11) shown in Sect. 6.2.

Thus, if sampling-time is 0·1 then for less than 1 % error the minimum modulation frequency is 50 H

12. Appendix 3

From Section 1 we have the following equations:

$$e(t) = E \cos \theta(t) \qquad \text{.....(1)}$$

$$\theta(t) = \omega_c t + \frac{\Delta f_{\max}}{f_m} \sin \omega_m t$$

$$e(t) = E \cos \left[\omega_c t + \frac{\Delta f_{\max}}{f_m} \sin \omega_m t\right] \qquad \text{.....(5)}$$

If we tune the local oscillator in Fig. 7 for f_c then the mixer output will be given by

$$e(t) = E \cos (\beta \sin \omega_m t) \qquad \text{.....(16)}$$

As t varies from 0 to $1/f_m$, $e(t)$ goes through a complete cycle from E through zero to $-E$ and back to $+E$ again. It will be assumed that the counter will register a count every time its trigger level is crossed in the 'positive-going' direction. For simplicity the counter is assumed to be a.c. coupled and the trigger level to be zero. It is therefore required to find solution of $e(t) = 0$ in a period of f_m whence the number of counts in one second, or the counter sampling time, is obtained.

For 1 cycle of f_m

$$0 < t < 1/f_m$$

i.e.

$$0 < \omega_m t < 2\pi$$

and $e(t) = 0$ when

$$\beta \sin \omega_m t = \pm \pi(n + \tfrac{1}{2})$$

where $n = 0, 1, 2 \ldots n_1$ and n_1 is the largest integer such that

$$\frac{\pi}{\beta}(n_1 + \tfrac{1}{2}) \leqslant 1 \qquad \text{.....(17)}$$

For any value of n, $\beta \sin \omega_m t = +\pi(n+\frac{1}{2})$ has two solutions in the range $0 < \omega_m t < 2\pi$. Similarly $\beta \sin \omega_m t = -\pi(n+\frac{1}{2})$ has two solutions in the same range.

The total number of solutions is therefore four for each value of n. Since now n can take on all values from 0 to n_1 the total number of solutions of $e(t) = 0$ is $4(n_1 + 1)$.

From eqn. (17), n_1 is the largest integer such that

$$(n_1 + \tfrac{1}{2}) \leqslant \beta/\pi$$

i.e.

$$n_1 = \left[\frac{\beta}{\pi} - \frac{1}{2}\right]$$

where the square brackets denote the largest integer not greater than the value of the expression inside the brackets.

The number of solutions of $e(t) = 0$ is then given by

$$4\left[\frac{\beta}{\pi} + \frac{1}{2}\right] \quad \text{within } 0 < \omega_m t < 2\pi$$

The number of counts recorded in this interval will be half the number of solutions of $e(t) = 0$, namely,

$$2\left[\frac{\beta}{\pi} + \frac{1}{2}\right]$$

Since this is the number of counts in one cycle of f_m then the number of counts per second is given by:

$$N = 2f_m \left[\frac{\beta}{\pi} + \frac{1}{2}\right]$$

Now the deviation $\Delta f'$ deduced from this measurement is given by:

$$\Delta f' = \frac{\pi}{2} N$$

$$= \pi f_m \left[\frac{\beta}{\pi} + \frac{1}{2}\right]$$

Substituting $\beta = \Delta f / f_m$, where Δf is the true deviation:

$$\Delta f' = \Delta f \frac{\pi}{\beta} \left[\frac{\beta}{\pi} + \frac{1}{2}\right]$$

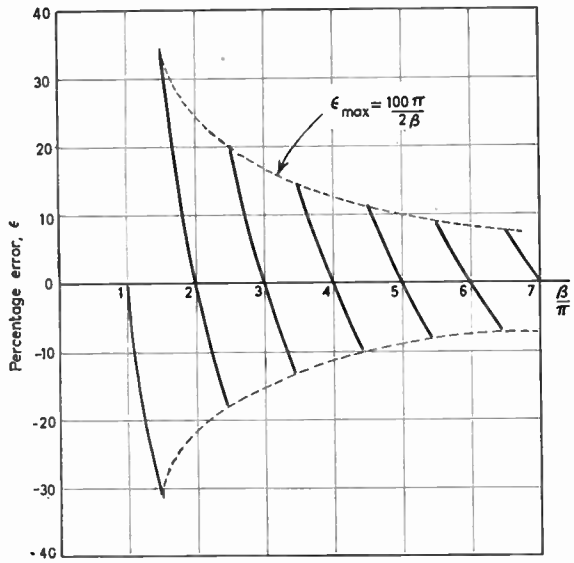


Fig. 16. Variation of percentage error as a function of modulation index.

The error in the measurement ε is

$$\frac{\Delta f' - \Delta f}{\Delta f} \times 100\%$$

i.e.

$$\varepsilon = \frac{\pi}{\beta}\left\{\left[\frac{\beta}{\pi} + \frac{1}{2}\right] - \frac{\beta}{\pi}\right\} 100\% \qquad \text{.....(18)}$$

Equation (18) is shown plotted in Fig. 16. The maximum error deduced from eqn. (18) by applying the limits

$$-\frac{1}{2} < \left[\frac{\beta}{\pi} + \frac{1}{2}\right] - \frac{\beta}{\pi} \leqslant \frac{1}{2}$$

is given by

$$\varepsilon_{max} = \frac{\pi}{2\beta} 100\% \qquad \text{.....(19)}$$

This is shown in Fig. 13.

Manuscript first received by the Institution on 26th September 1967 and in final form on 12th October 1967.
(Paper No. 1188/IC1.)

U.D.C. 621.374.32

An Extension of the Use of Karnaugh Maps in the Minimization of Logical Functions

By

K. J. DEAN, M.Sc., F.Inst.P., C.Eng., F.I.E.E., F.I.E.R.E.†

Summary: Karnaugh maps are used as a graphical technique for minimizing logical functions. They are usually employed where there are up to four literals, or with some difficulty for a maximum of six literals. It is proposed that the use of these maps can be extended to eight literals by arranging that each cell of the map be divided to form a subsidiary Karnaugh map.

1. Introduction

Recent papers have described the design of some electronic counters[1] and cyclic code generators.[2] One technique for the minimization of the logic functions involved in these designs is due to Karnaugh.[3] These functions are necessary to steer the bistable elements which store the various sequential states, from each state to the succeeding one. In this type of logical design the Karnaugh maps are completed with reference to a steering table which is appropriate to the type of bistable element which is being employed. Table 1 shows a steering table for a J–K flip flop. This steering table should not be confused with the truth table from which it is derived. The truth table shows the effect of certain stated input levels on the state of the flip-flop. The steering table shows what input levels are required to steer the flip-flop between two given states. The derivation of the steering table was discussed in two earlier papers.[1, 4]

Table 1

Steering table for J-K flip-flop

Steering		Inputs	
from	to	J	K
0	0	0	x
0	1	1	x
1	0	x	1
1	1	x	0

x denotes an immaterial state

2. Use of Four-literal Map

The use of a Karnaugh map may be illustrated by the design of a cyclic generator which has six states as follows:

A	B	C	D
0	0	0	1
0	1	0	1
0	1	1	0
1	0	0	1
1	1	0	0
1	1	1	0

It may be assumed that other states do not occur, although some precautions may be necessary to ensure this. Two Karnaugh maps are required for each of the four bistable elements (flip-flops), the maps for element D being as shown in Fig. 1.

Minimization is usually carried out by looping groups of cells. However, minimization will be shown here by shading cell groups, since this method is easier in some of the cases which will be examined here. In order to recognize more quickly the groups of cells in each minimization, different types of cross hatching will be used. Thus, in this example, $D_J = C$ and $D_K = A + B$ if all immaterial and unmarked cells can be assigned either 0 or 1 as is convenient.

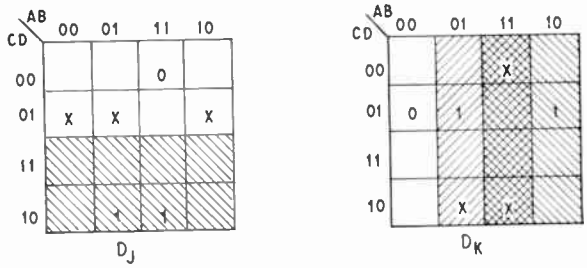


Fig. 1. Four-literal map

The limitation of the method is often stated to be that it cannot handle more than four literals whilst maintaining the property that adjacent cells differ in only one literal. To extend the method to five literals two maps have been used to replace each map, so that four maps are now needed for each flip-flop. However this can be avoided if each cell of the maps is divided into two phases, so that minimization can be carried out on either phase or by including both of them.

† Department of Science and Electrical Engineering, Letchworth College of Technology, Letchworth, Herts.

3. Five-literal Maps

Consider the following example in which nine states must occur sequentially. The conditions for flip-flop E are examined by two maps of this kind. (Fig. 2.) Here again the steering table (Table 1) is used to determine the states entered in the cells.

A	B	C	D	E
0	0	0	1	0
0	0	1	0	1
0	0	1	1	1
0	1	0	0	0
0	1	0	0	1
0	0	0	0	1
1	0	0	1	0
1	1	0	0	0
1	1	0	0	1

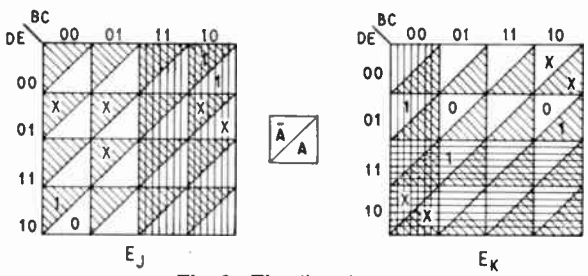


Fig. 2. Five-literal maps

Examination of the maps shows that $E_J = \bar{A}+B$ and $E_K = A+D+B\bar{C}$. (There are some possible alternatives to these.)

4. Six-literal Maps

When six literals are involved it is usual to consider each map as a solid figure with 64 cells forming a cube. It is suggested here that this can be avoided if each cell of a 16-cell map is divided into four phases, each cell containing in effect a two-literal map. This is illustrated in the example in Fig. 3 of part of a sequential design where the conditions for flip-flop F are deduced.

From the maps $F_J = BD+E$ and $F_K = C+D$. Clearly, this could not have been completed so readily had all the cells been filled and had there been no immaterial cells, but there are certainly some occasions when the method reduces the labour involved. In addition, the method depends to a large extent on visual pattern recognition, which is often more rapid than mathematical manipulation.

5. Extension to Eight Literals

Now it will be seen that the method consists of placing a subsidiary map within each cell of the main map. In this way there will be eight phases in each cell if three additional literals are needed beyond the basic four literals (i.e. a total of seven literals). Similarly, using 16 phases, eight literals can be manipulated. Two such maps are shown in Figs. 4 and 5 with their minimized solutions.

A	B	C	D	E	F
0	0	1	0	1	1
0	0	1	1	0	0
0	1	1	0	1	0
0	1	1	0	1	1
0	1	1	1	0	0
1	0	0	0	0	1
1	0	0	1	0	1
1	1	1	0	0	0
1	1	1	0	1	0

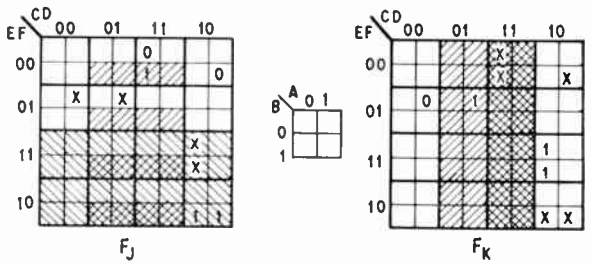


Fig. 3. Six-literal maps

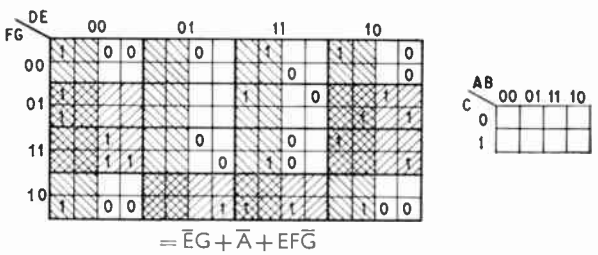


Fig. 4. Seven-literal map

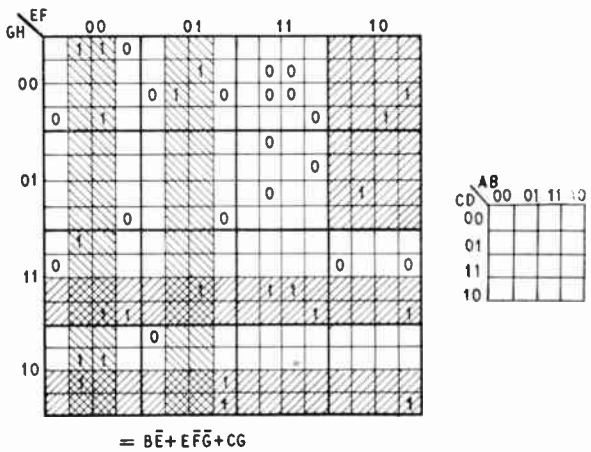


Fig. 5. Eight-literal map

6. Conclusions

Since this is a graphical method it is not suggested that a true (best) minimization will necessarily be recognized but the deduced results may be obtained more quickly than with alternative methods, such as those due to McCluskey[5] and Hirst[6] within the limits of the complexity described. The method may also be compared with that described by Phister[7] for the extension of Veitch diagrams. Although there are some points of similarity, the present method retains many of the advantages of the Karnaugh system of numbering, both so far as the overall map is concerned, and within each cell of the overall map, although not between the cell maps (i.e. the phases of any cell) and the overall map.

7. Acknowledgment

This work has been carried out in preparation for a research programme at Letchworth College of Technology.

8. References

1. K. J. Dean, 'The design of parallel counters using the map method', *The Radio and Electronic Engineer*, **32**, No. 3, p. 159, September 1966.
2. K. J. Dean, 'The design of variable ratio cyclic generators using JK flip-flops', *Proc. Instn Radio Electronics Engrs (Aust.)*, **28**, p. 53, February 1967.
3. M. Karnaugh, 'The map method for synthesis of combinational logic circuits', *Commun. Electronics*, 9, p. 539, November 1953.
4. K. J. Dean, 'Code Converters and Associated Systems', M.Sc. Thesis, University of London, 1967.
5. E. J. McCluskey, Jr., 'Minimization of Boolean functions', *Bell Syst. Tech. J.* **35**, No. 6, p. 1417, November 1956.
6. P. D. Hirst, 'A jig for minimization of Boolean functions', Conf. on Logic Design, University of Reading, 1967.
7. M. Phister, Jr., 'Logical Design of Digital Computers', p. 84 (Wiley, New York, 1958).

Manuscript first received by the Institution on 8th November 1967 and in revised form on 11th January 1968.
(Contribution No. 103/CC7.)

U.D.C. 621.373.431.1

A 5-MHz Switching Multivibrator using a Complementary Pair of Transistors

By

B. D. RAKOVICH,
Dip.Eng., Ph.D.†

AND

S. L. TESIC,
Dip.Eng.†

Summary: The principle of operation of a novel free-running transistor multivibrator is presented. The circuit is based on the application of a complementary switch, the operation of which is controlled by a single timing network. This circuit provides a suitable means of generating rectangular pulses having rise- and fall-times in the nanosecond range.

The circuit exhibits some other useful features which include excellent thermal stability and considerable loading capabilities for both resistive and capacitive loads, and less dependence on power supply variations.

The experimental model which was built to verify the design operated properly in the frequency range extending from very low frequencies to frequencies over 5 MHz when only the timing capacitor was changed.

1. Introduction

The available literature on transistor astable multivibrators provides many useful circuit designs of collector- and emitter-coupled multivibrators[1-5] or multivibrators using serial connection of the transistors,[6,7] most of which are analogous to the previously described circuits with thermionic valves. The conventional type of astable multivibrator with the collector coupling capacitors (Fig. 1), is most widely used but in some applications it does not yield entirely satisfactory results, primarily because of the integrating effect of the cross-coupling capacitors which lengthens the rise- or fall-times of the output pulses. Frequency stability of this circuit over the ambient temperature range from 20–60°C can be hardly made better than 1% even if silicon planar transistors are used. Starting the oscillations may also constitute a real problem. Under certain conditions the circuit may fail to start oscillating with both transistors remaining in the 'on' state. To prevent this it is often necessary to apply collector voltage before applying base voltage when the circuit is turned on.

Various modifications of the basic collector-coupled multivibrator have been described in which the afore-mentioned deficiencies are greatly reduced or completely removed.[8,9] However, this can be achieved only at the expense of much greater complexity of the circuit. In most of these designs four transistors are used instead of only two in the basic circuit.

Other possibilities for the generation of rectangular pulses with short rise- and fall-times consist in using emitter-coupled or serial connection circuits. The best known and probably the most efficient emitter-coupled multivibrator is shown in Fig. 2. In this circuit the collector of the transistor TR2 is not directly included in the feed-back path so that no integrating effect is present at that collector. Hence,

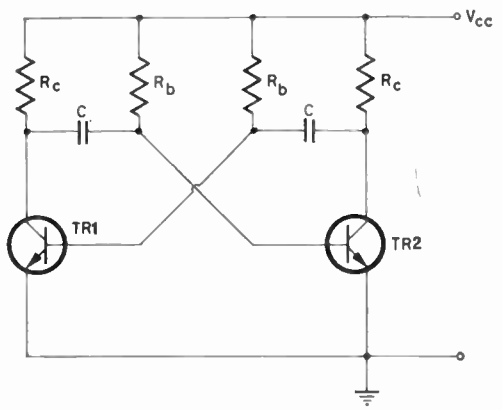


Fig. 1. Basic arrangement of collector-coupled multivibrator.

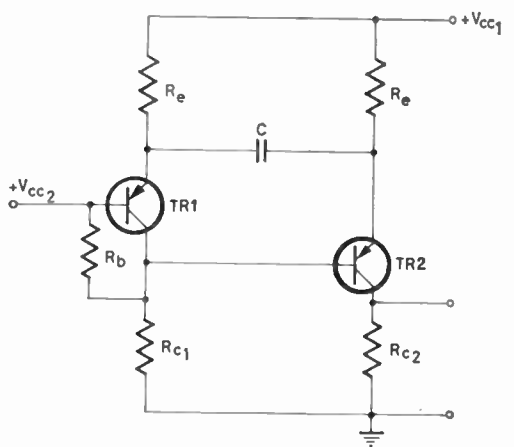


Fig. 2. Emitter-coupled multivibrator.

† Faculty of Electrical Engineering, University of Belgrade, Belgrade, Yugoslavia.

the collector of TR2 makes an ideal point from which to obtain the output pulses with very short rise- and fall-times. The effect of ambient temperature on frequency is very much reduced in this circuit especially if silicon planar transistors are used. The frequency of oscillation is almost independent of resistive loads, while capacitive loading affects both rise- and fall-times equally. The main disadvantage of this circuit results from the fact that the transistor TR2 must not saturate when conducting; otherwise the multivibrator will not be inherently self-starting. Consequently, the output pulses suffer from a distinct tilt of the lower level, and their magnitude, which is a rather small fraction of the supply voltage, is greatly dependent on resistive loading.

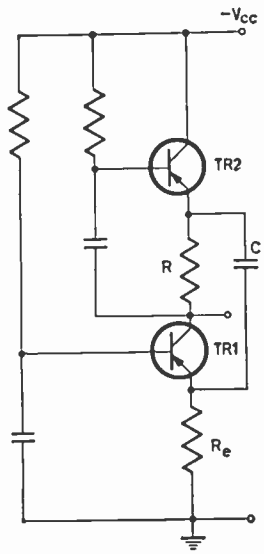


Fig. 3. A serial multivibrator.

Serial multivibrators, a variant of which is shown in Fig. 3, yield fairly good shaped output pulses but their timing period is strongly affected by resistive load and by changes in supply voltage. Due to a relatively high dynamic output impedance, the circuit cannot be loaded capacitively to any appreciable extent.

This paper presents a novel astable multivibrator which has no counterpart in valve technique. Two complementary transistors are used to form a regenerative switch, the operation of which is controlled by a simple timing network. Unlike all other multivibrators of the afore-mentioned types, both transistors of this circuit are either simultaneously in the 'on' or in the 'off' state. Symmetrical or asymmetrical rectangular waveforms with very short transition times are obtained at one collector. This circuit provides small static and dynamic output impedances so that it can be loaded appreciably both with resistive and capacitive loads. Other important features of this circuit such as the temperature stability and the constancy of the period of oscillation against the changes in supply voltage are also discussed.

A silicon controlled switch (p-n-p-n transistor) can also be used in this circuit to replace complementary transistors. When the speed attainable with these devices is adequate, they offer the advantage of using a single active device and a minimum of components.

2. Description of the Circuit

2.1. *The Complementary Switch*

It is known that a p-n-p and an n-p-n transistor connected as shown in Fig. 4 generate a voltage-current characteristic similar to that of the silicon controlled switch. By breaking the circuit at the emitter of TR1 (Fig. 5) the voltage-current characteristic of the switch can be examined as a two terminal device. The essential piecewise-linear characteristic, shown in Fig. 6, exhibits three regions. The negative resistance characteristic (region II) is generated while both transistors are active. Since the collectors and bases of TR1 and TR2 are cross-connected they both come 'on' and both saturate together. Therefore, in the region III both transistors are saturated and in the region I they are both in the 'off' state. The complete characteristic can be defined by analysing the negative resistance region only since its end points can be easily obtained.

We shall start the analysis by writing the following equation for the circuit in Fig. 5:

$$V_N - V_{be1} + R_{c2} i_2 = 0 \qquad \text{.....(1)}$$

$$i_2 = \beta_2 i_{b2} - (1-\alpha_1) i_N = \left[\alpha_1 \beta_2 \frac{R_i}{R_i + R_{c1}} - (1-\alpha_1)\right] i_N \qquad \text{.....(2)}$$

where R_i is the input resistance of TR2.

Substituting i_2 from eqn. (2) in eqn. (1) and neglecting V_{be1} yields for the negative resistance R_N

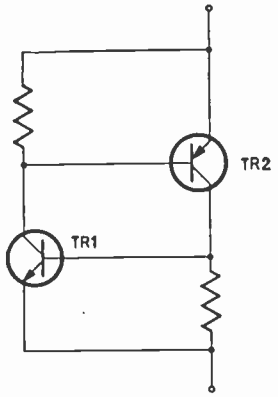


Fig. 4. A regenerative connection of a p-n-p and an n-p-n transistor.

$$R_N = \frac{V_N}{i_N} = \left[1-\alpha_1\left(1+\beta_2\frac{R_i}{R_i+R_{c1}}\right)\right]R_{c2}$$
$$= (1-\alpha_1-A_S)R_{c2} \qquad \text{.....(3)}$$

where

$$A_S = \frac{i_{c2}}{i_N}$$

From Fig. 5 we find

$$V_{cc1}-\alpha_1\frac{R_{c1}R_i}{R_{c1}+R_i}i_N - V_{cb1} - V_{be1} + V_N = 0 \quad \text{...(4)}$$

For $i_N = 0$, $V_{cb1} = V_{cc1}$ and $V_{be1} = -V_{beo1}$ (base-to-emitter cut-in voltage) so that the upper limit of the negative resistance region is: $i_N = 0$ $V_N = -V_{beo1}$. For the lower limit of the active region we have $i_N = I_p$, $V_N = -(V_{cc1}+V_{ces2}-V_{bes1})$.

The behaviour of the circuit in the active region is thoroughly analysed in the literature and will not be repeated here. However, it is obvious that this circuit is capable of performing the fast switching function since the regeneration can be made the highest

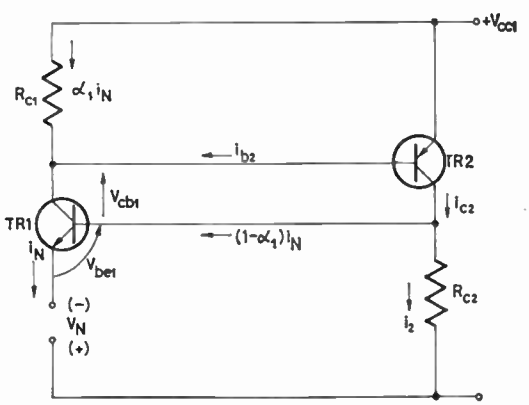


Fig. 5. Circuit for computing the voltage-current characteristic of the complementary switch.

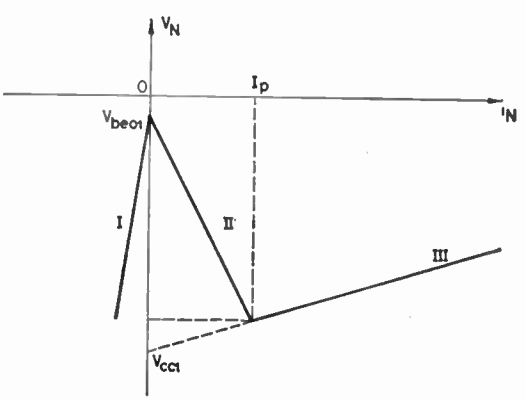


Fig. 6. Current-controlled characteristics of complementary switch.

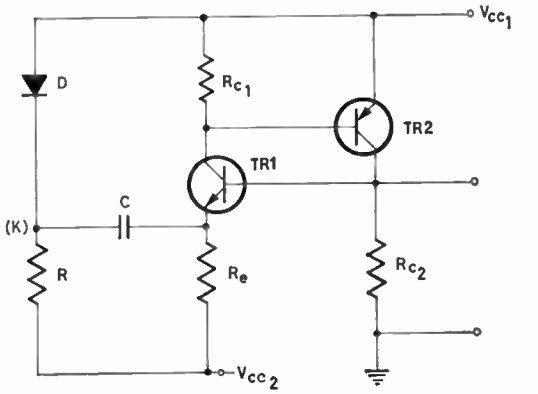


Fig. 7. Multivibrator using p-n-p–n-p-n switch.

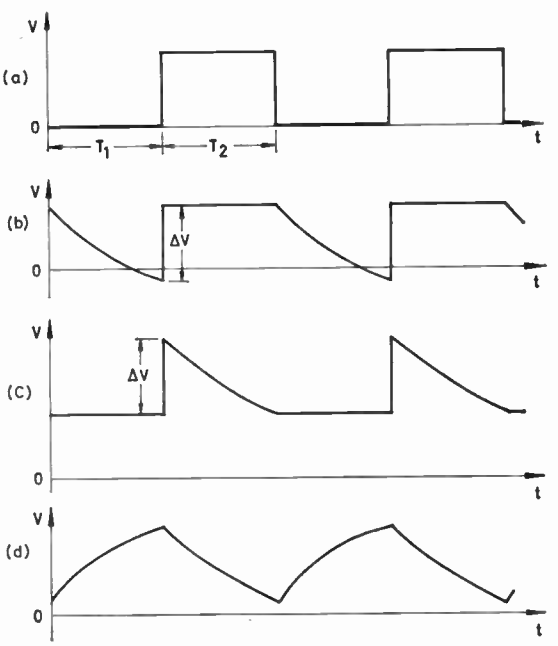


Fig. 8. Waveforms of the circuit in Fig 7.

(a) collector TR2;
(b) emitter TR1;
(c) cathode side of diode D;
(d) across timing capacitor C.

possible for any way of connecting together two transistors. In fact, several bistable and monostable trigger circuits have already been described using this type of regenerative switch.[10-12]

2.2. *Astable Operation*

The astable multivibrator, obtained by adding an appropriate timing network to the regenerative switch, and the waveforms at different points are presented in Figs. 7 and 8. During the quasi-stable state in which the timing capacitor charges through

the diode D and the emitter resistance R_e both transistors are in the 'off' state since the base-to-emitter junction of the transistor TR1 is inversely biased by the positive voltage drop across the emitter resistor R_e. The charging of the timing capacitor terminates at the instant when the emitter-base junction of TR1 becomes forward biased. Then a regenerative process starts bringing both transistors to saturation in a very short time. After switching to the second quasi-stable state, when the emitter of TR1 has risen due to the increased voltage drop across R_e, the cathode of D will also rise, thus cutting-off the diode current. The timing capacitor starts to discharge through R and the complementary switch. This quasi-stable state will terminate after the voltage drop across the timing resistor R has decreased to a value at which the cathode of the diode D is clamped to V_{cc1}. At this moment both transistors desaturate and the loop gain greatly exceeds unity causing both transistors to turn off by regeneration.

In order to ensure the astable operation of the circuit we shall require that the driving-point impedance $Z(p)$ of the equivalent circuit in Fig. 9 has a positive real zero in the complex frequency plane $p = \alpha + j\omega$

$$Z(p) = \frac{r_d\left(R + \frac{R_e R_N}{R_e + R_N}\right)Cp + \frac{RR_eR_N}{R_e+R_N}Cp + R + r_d}{\left(R + \frac{R_eR_N}{R_e+R_N}\right)Cp + 1} \quad \ldots\ldots(5)$$

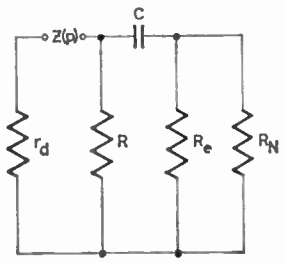


Fig. 9. Low-frequency equivalent circuit for computing the driving-point impedance.

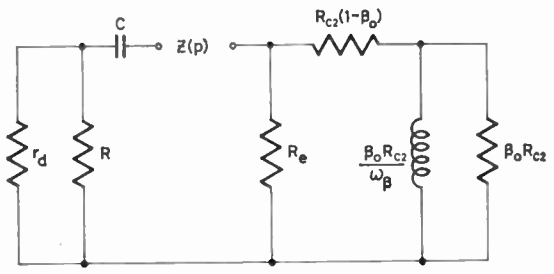


Fig. 10. High-frequency equivalent circuit for computing the driving-point impedance.

where r_d is the resistance of the positively biased diode D. Hence,

$$p = -\frac{(R+r_d)(R_e+R_N)}{C(r_dRR_e + r_dRR_N + r_dR_eR_N + RR_eR_N)} \quad \ldots(6)$$

It follows from eqn. (6) that the zero of $Z(p)$ will lie on the positive real axis in the p-plane if

$$R_e + R_N > 0 \quad \ldots\ldots(7)$$

$$r_dRR_e + r_dRR_N + r_dR_eR_N + RR_eR_N < 0 \quad \ldots(8)$$

from which we obtain

$$r_d < \frac{R_eRR_{c2}(A_S + \alpha_1 - 1)}{RR_e - (R+R_e)R_{c2}(A_S + \alpha_1 - 1)} \quad \ldots\ldots(9)$$

$$\frac{RR_e}{R+R_e} > R_{c2}(A_S + \alpha_1 - 1) \quad \ldots\ldots(10)$$

The conditions (9) and (10) can be expressed in terms of current loop-gain of the circuit by stating that the current loop-gain can exceed unity if, and only if, both transistors and the diode are conducting and the current amplification A_S is greater than unity.

At high frequencies a more rigorous model of the transistor equivalent circuits than the simple current-generators must be used when computing A_S. This would lead, however, to a rather complicated expression for the input impedance of the complementary switch. The calculation can be greatly simplified if we assume that the collector resistor of TR1 is much higher than the input impedance of TR2 ($R_{c1} \gg Z_i$) and that the only frequency dependent transistor parameter is the current amplification factor β_2 of TR2. These assumptions seem to be reasonable for not too small values of R_{c1} since the Z_i decreases as ω increases.

Now, neglecting the effect of R_{c1} and substituting

$$\beta = \frac{\beta_o}{1 + j\frac{\omega}{\omega_\beta}}; \qquad \alpha_1 = \alpha_2 = \alpha$$

in eqn. (3), we have

$$R_N = [1 - \alpha(1+\beta)]R_{c2} = (1-\beta)R_{c2}$$

$$= \left(1 - \frac{\beta_o}{1 + j\frac{\omega}{\omega_\beta}}\right)R_{c2}$$

$$R_N = \left(1 - \frac{\omega_\beta^2\beta_o}{\omega_\beta^2 + \omega^2}\right)R_{c2} + j\omega\frac{\omega_\beta\beta_oR_{c2}}{\omega_\beta^2 + \omega^2} \quad \ldots(11)$$

The equivalent circuit for computing the driving-point impedance $Z(p)$ is shown in Fig. 10. The inductive component in the input impedance of the complementary switch is due to the second term on the right side of eqn. (11). The driving-point impedance

$Z(p)$ is found to be

$$Z(p)=\frac{C(R'R_e+R'R_{c2}+R_eR_{c2})p^2+[R_{c2}+R_e+\omega_\beta C(R'+R_e)R_{c2}(1-\beta_o)+\omega_\beta CR'R_e]p+\omega_\beta R_e+\omega_\beta R_{c2}(1-\beta_o)}{pC[p(R_{c2}+R_e)+\omega_\beta R_e+\omega_\beta R_{c2}(1-\beta_o)]} \quad \ldots\ldots(12)$$

where

$$R'=\frac{Rr_d}{R+r_d}$$

For astable operation both zeros of $Z(p)$ must have positive real part and for square-wave output waveform one might postulate that they lie on the positive real-axis in the p-plane. Therefore, for astable operation the coefficient of p in the numerator of eqn. (12) must be negative, from which we find

$$C>\frac{R_{c2}+R_e}{\omega_\beta(R'+R_e)\left[R_{c2}(\beta_0-1)-\dfrac{R'R_e}{R'+R_e}\right]} \quad \ldots(13)$$

The last expression provides a useful piece of information on the smallest values of the timing capacitor for which both zeros of the driving-point impedance lie in the right half of the p-plane. However, with these values of C, the zeros of the driving-point impedance are complex conjugates and are located near the imaginary-axis so that the output waveform loses its square corners and tends to become sinusoidal. Increasing the timing capacitor, the zeros move towards the positive real axis thus improving the shape of the output pulses. Alternatively, transistors with higher cut-off frequency may be used. These conclusions are in a reasonably good agreement with the results obtained experimentally. For example, in an experimental circuit using BSX28 and BSX29 for TR1 and TR2 respectively and $R_{c2}=100\Omega$, oscillation at approximately 10 MHz was produced with the value of the timing capacitor of only 20 pF. When a transistor (V435) having more than three times lower current gain-bandwidth rating than BSX29 was used to replace BSX29 it was necessary to increase the timing capacitor to 50 pF to start oscillation.

2.3. *Multivibrator Period*

The duration of the timing period can be determined from a consideration of the waveform diagram Fig. 8, and the equivalent circuit for computing the charging period of the timing capacitor (Fig. 11), from which the following equations can be written:

$$V_{cc1}-r_d(i_1+i_2)-R_ei_1-\frac{q_1}{C}+V_{cc2}=0 \quad \ldots(14)$$

$$V_{cc1}-r_d(i_1+i_2)-Ri_2+V_{cc2}=0 \quad \ldots(15)$$

Subject to the initial condition $V_{co}=V_{bes1}+V_{ces2}-v_d$ at $t=0$, where V_{bes1}, V_{ces2} are saturation collector-to-emitter and base-to-emitter voltages of TR1 and TR2 respectively and v_d is the voltage across the forward biased diode, the solution to these equations is

$$i_1=\frac{V_{cc1}+V_{cc2}-V_{co}\left(1+\dfrac{r_d}{R}\right)}{R_e+r_d\left(1+\dfrac{R_e}{R}\right)}\exp\left(-\frac{t}{\tau_1}\right) \quad \ldots(16)$$

where

$$\tau_1=C\left(R_e+\frac{Rr_d}{R+r_d}\right)$$

is the charging time constant. The voltage at the emitter of TR1 is

$$v_{e1}=-V_{cc2}+R_e\frac{V_{cc1}+V_{cc2}-V_{co}\left(1+\dfrac{r_d}{R}\right)}{R_e+r_d\left(1+\dfrac{R_e}{R}\right)}\exp\left(-\frac{t}{\tau_1}\right) \quad \ldots\ldots(17)$$

The quasi-stable state in which both transistors are in the 'off' state terminates immediately after the emitter voltage v_{e1} has reached the value $v_{e1}=-V_{beo1}$ at which TR1 starts conducting. Substituting $v_{e1}=-V_{beo1}$ and $t=T_1$ in eqn. (17) we find

$$T_1=\tau_1\log\frac{1}{1+\dfrac{r_d}{R_e}+\dfrac{r_d}{R}}\times \times\frac{V_{cc1}+V_{cc2}-(V_{ces2}+V_{bes1}-v_d)\left(1+\dfrac{r_d}{R}\right)}{V_{cc2}-V_{beo1}}$$

$$\simeq\tau_1\log\frac{V_{cc1}+V_{cc2}-V_{ces2}-V_{bes1}+v_d}{V_{cc2}-V_{beo1}} \quad \ldots\ldots(18)$$

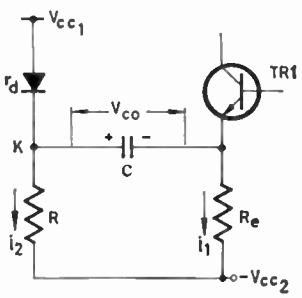


Fig. 11. Equivalent circuit for computing the charging period of the timing capacitor.

The duration of the other quasi-stable state in which both transistors are conducting can be readily computed by noting that the initial voltage rise at the cathode of D is equal to

$$\Delta V = V_{cc1} - V_{bes1} - V_{ces2} + V_{beo1}$$

Now

$$(V_{cc1} - v_d + \Delta V) \exp\left(-\frac{T_2}{\tau_2}\right) = V_{cc1} + V_{cc2} - v_d \qquad \text{.....(19)}$$

from which we have

$$T_2 = \tau_2 \log \frac{2V_{cc1} + V_{cc2} - V_{ces2} - V_{bes1} - v_d + V_{beo1}}{V_{cc1} + V_{cc2} - v_d} \qquad \text{.....(20)}$$

where the discharging time-constant τ_2 is approximately equal to CR.

2.4. *Design Consideration*

It can be concluded from the foregoing analysis that the d.c. operating points of the transistors may be situated almost anywhere in the active region since, with timing network disconnected, the circuit is absolutely stable in the negative resistance region (eqn. (10)). However, the best performance of the circuit with regard to the loading capability and the temperature stability is obtained if d.c. operating condition of the complementary switch are arranged so that both transistors are near the saturation region or even slightly saturated. In the latter case, care must be taken not to oversaturate the transistors under d.c. operating condition, otherwise A_S may be very much reduced and the circuit will not oscillate when the timing network is connected.

At the beginning of the quasi-stable state in which the output pulses are generated at the collector of TR2 both transistors are driven further in saturation by the discharging current of the timing capacitor. As the discharging current decreases the current loop-gain increases, but the output pulse will not terminate before the point K has been caught by diode clamp. At this moment the emitter resistor R_e becomes effectively shunted by the timing resistor and a relatively small resistance of the forward biased diode and the loop-gain exceeds unity thus enabling the regenerative process to start.

The complete design procedure of the multivibrator involves the following steps:

(i) Select transistors TR1 and TR2 taking into account that rise- and fall-times are determined primarily by the alpha-cut-off frequency of TR1 and TR2. Preferably, they should have low saturation voltage.

(ii) If supply voltages are not given, higher values of V_{cc2}, say, $V_{cc2} \geq 2V_{cc1}$, would enhance the temperature stability of the circuit.

(iii) Select the d.c. operating point of TR2 and find the value of R_{c2}:

$$R_{c2} = \frac{V_{cc1} - V_{ces2}}{I_2}$$

V_{ces2} can be found from the detailed characteristics of the transistor as one half of the base-to-emitter saturation voltage V_{bes2} for which $I_c : I_b = 10$. If these characteristics are not available it is safe to assume for silicon switching transistors $V_{ces2} = 0{\cdot}5$–$0{\cdot}8$ V.

The operating point corresponding to smaller values of the collector current ($I_{c2} = 10$–20 mA) should be used unless a heavy capacitive load is expected at the output terminals.

(iv) The collector of TR1 must supply quite a small current

$$\frac{V_{bes2}}{R_{c1}} + \frac{I_{c2}}{\beta_2}$$

so that the current through TR1 may be chosen in the range 2 to 4 mA. Now.

$$R_e \simeq \frac{V_{cc2} + V_{cc1}}{I_{c1}} \quad \text{and} \quad R_{c1} = \frac{V_{bes2}}{I_{c1} - I_{b2}}$$

Since TR2 is slightly saturated the current gain β_2 may be reduced even by a factor of 2. However, it is better to use the typical value of β as indicated on the transistor data sheets (or a somewhat higher value) thus allowing for the possible variation of β between different samples of TR2. In this way, if the transistor with β at its upper tolerance limit is used in the circuit, TR2 will saturate under d.c. operating condition while, with β at its lower tolerance limit, the d.c. operating point of TR2 will be in the active region. This does not prevent TR2 from being driven to saturation during the quasi-stable state in which both transistors are conducting, since the discharging current of the timing capacitor increases the current through TR1. This quasi-stable state terminates when the discharging current has decreased to approximately

$$\frac{V_{cc1} + V_{cc2}}{R}$$

(v) Calculate the value of C and R by using eqns. (18) and (20).

(vi) Check that condition (10) is fulfilled, if not, R_e must be increased and R_{c1} decreased.

Using this procedure a number of symmetrical and asymmetrical multivibrators were designed and quite a good agreement between the calculated and measured values were found.

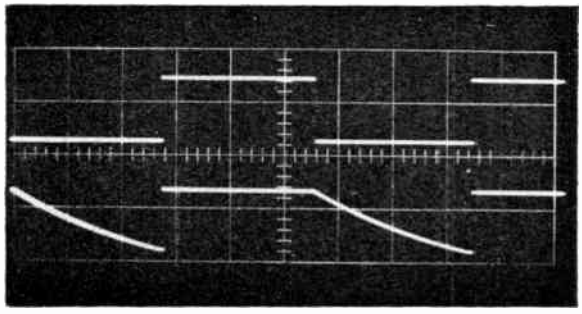


Fig. 12. TR2-collector (circuit of Fig. 11) and TR1-emitter waveforms at $f = 1$ kHz. (Horizontal scale: 0·2 ms/div. Vertical scale: 5 V/div.).

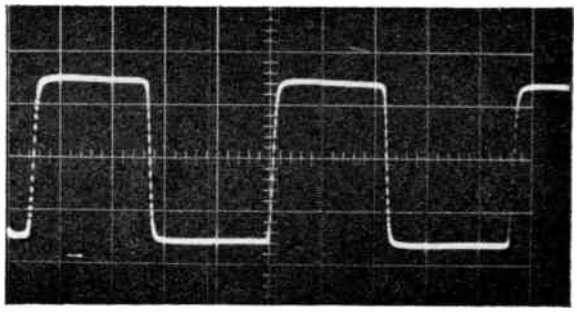
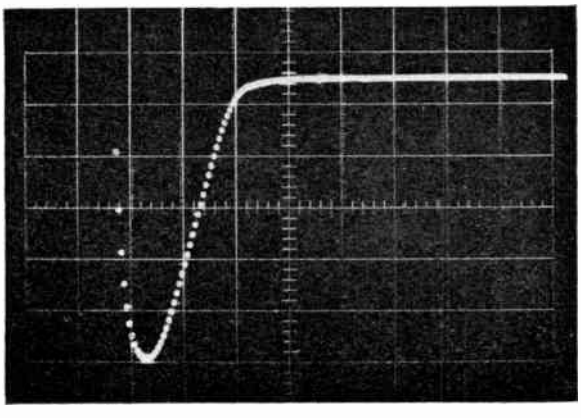


Fig. 13. Output pulses at $f = 4$ MHz (Fig. 11). (Horizontal scale: 50 ns/div. Vertical scale: 2 V/div.).

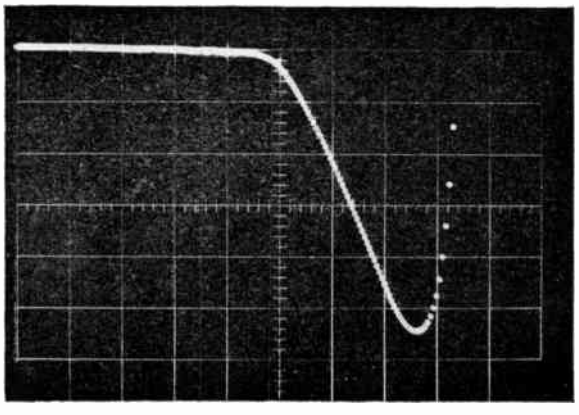
It is worth noting that the proposed circuit can also be designed in such a way that the duration of the output pulses is not controlled by the diode clamp. Moreover, in this design the diode D can be removed from the circuit. The current loop gain is smaller than unity while the transistors are in saturation and this quasi-stable state terminates soon after the discharging current has decreased to the value at which the base currents of TR1 and TR2 cannot keep them in saturation. This mode of operation is, however, very much inferior when compared with the one previously described especially in respect to loading capability and temperature stability. Since the base current at which TR1 comes out of saturation depends on transistor β, the pulse width will vary considerably with different samples of the transistors if this mode of operation is used. Besides, the adjustment of the circuit is rather critical and the duration of the 'on' state cannot be easily predicted.

3. Experimental Results

The experimental model of the circuit, shown in Fig. 7, was constructed using high-speed silicon planar transistors BSX28 and BSX29 for TR1 and TR2 respectively, and the diode type EA828. Other circuit parameters were as follows: $R = 15$ kΩ, $R_e = 10$ kΩ, $R_{c1} = 1$ kΩ, $R_{c2} = 100$ Ω, $C = 0{\cdot}1$ μF, $V_{cc1} = 6$ V, $V_{cc2} = -12$ V. In Fig. 12 the output pulses of approximately 6 V (peak-to-peak) and the frequency of oscillation $f = 1$ kHz are shown together with the waveform at the emitter of n-p-n transistor. In Fig. 13 the output pulses at $f = 4$ MHz are shown while in Fig. 14 the leading and trailing edges of these pulses are displayed on 5 ns/div sweep. Rise- and fall-times of the output pulses are not affected by the value of the timing capacitor and they can be decreased to approximately 10 ns if the smaller values for R_{c1} and R_e are used.



(a) Leading edge of the output pulses at $f = 4$ MHz.

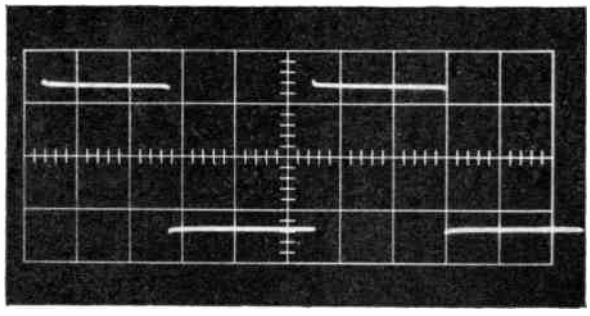


(b) Trailing edge of the output pulses at $f = 4$ MHz.

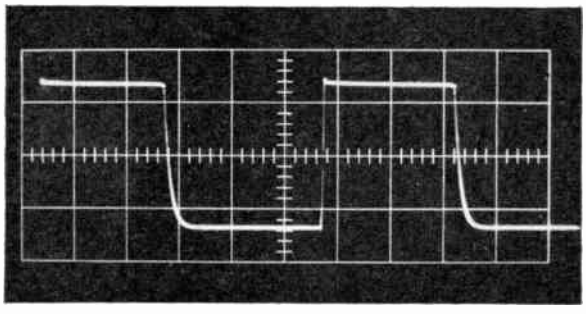
Fig. 14. (Horizontal scale: 5 ns/div. Vertical scale: 1 V/div.).

The circuit is inherently self-starting and owing to the small output impedance it can be loaded considerably both with resistive and capacitive loads. A loading resistor $R = 1\ k\Omega$ connected between the collector of TR2 and ground produced a change of the period of approximately 0·5%, while a 10% change of the oscillation period was recorded with $R = 100\ \Omega$. The effect of capacitive loading can be seen in Fig. 15. Figure 15(a) shows 10 kHz output pulses with the unloaded output while in Fig. 15(b) the effect of a capacitive load ($C = 0{\cdot}01\mu F$) connected across the output terminals is shown.

Since the duration of the 'on' state is controlled primarily by the clamping action of the diode D, the



(a) Output terminals unloaded.



(b) Output terminals loaded with $C = 0{\cdot}01\ \mu F$.

Fig. 15. Output waveforms at 10 kHz in circuit of Fig. 11. (Horizontal scale: 20 ns/div. Vertical scale: 2 V/div.).

frequency of oscillation is almost independent on transistor current amplification factor, β. A change in the frequency of oscillation of only 0·1% was measured with different samples of BSX28 used for TR1 having direct current gain of 30 and 90 respectively. However, the frequency variation of 2% was recorded with different samples of BSX29, used for TR2 having direct current gain of 40 and 140 respectively.

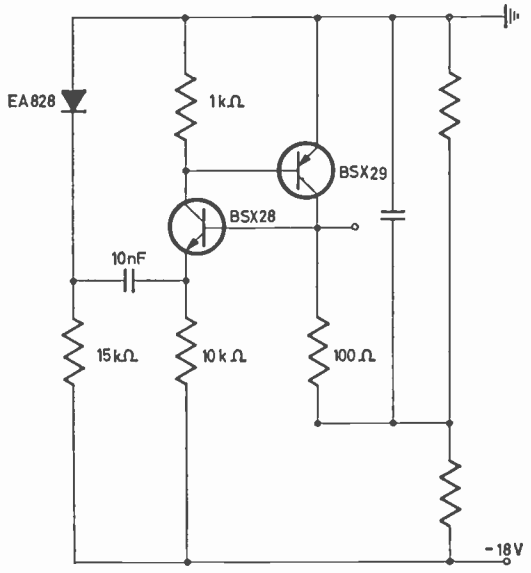


Fig. 16. Circuit for measuring the influence of supply voltage variations.

The frequency dependence of the circuit at a given frequency with respect to the power supply variation was checked with the circuit shown in Fig. 16 in which only one supply is used while the other supply voltage is derived from a resistive divider. It was found that a −50% change in supply voltage varied the frequency of oscillation by only 2%. This result is slightly better than that obtained with the multivibrator circuit in Fig. 2. The claim that for the latter circuit the frequency variation is less than 1% for a ±50% change in supply voltage[4] could not be confirmed during the experiment even in the case when both transistors are allowed to saturate in the 'on' state.

The variations of the frequency of oscillation with temperature were checked over the temperature range of 25°C to 100°C. The experiment was carried out with the timing network and biasing resistors outside the oven. Increasing the temperature increases the frequency of oscillation, but the maximum change of frequency was inside the limit of 0·2% over the temperature range of 25°C to 100°C, and only 0·03% over the temperature range of 25°C to 60°C. The experiment was repeated with the diode D outside the oven and the changes in the oscillation frequency of −0·3 and −0·4% were recorded corresponding to temperature variations from 25°C to 60°C and 25°C to 100°C respectively. This indicates the important role of the diode D in reducing the temperature sensitivity of the circuit in the temperature range where the influence of the inverse currents of TR1 and TR2 is negligible.

It can be seen from eqns. (18) and (20) that the temperature stability of the oscillation frequency can be further increased by using the highest permissible values for the supply voltages. In fact, this is true for all transistor multivibrators since in this case the timing periods are less influenced by the temperature dependent transistor parameters. With respect to the temperature stability it is also advisable to arrange d.c. operating conditions at smaller current levels where the collector-to-emitter saturation voltage is less dependent on temperature variation.

4. Conclusion

A free-running transistor multivibrator using complementary transistor has been described. The unusual feature of this circuit is that the transistors are either both in saturation or reverse biased state. The durations of quasi-stable states are controlled by the charging and discharging of one timing capacitor. At one collector, symmetrical or unsymmetrical rectangular waveforms are generated. The shape of the output pulses is not affected by the timing network and if high-speed switching transistors are used, rise- and fall-times of the output pulses of only 10 ns can be obtained at the operating frequencies extending from a small fraction of 1 Hz to several MHz.

The temperature stability of the oscillation frequency was checked and found that the total frequency variation due to the influence of the transistors and the diode was 0·03% over the ambient temperature range of 25°C to 60°C.

The circuit is inherently self-starting and provides small static and dynamic impedances at its output terminals. Therefore, it can be loaded considerably both with resistive and capacitive loads.

5. References

1. A. E. Jackets, 'A method for sharpening the output waveform of junction transistor multivibrator circuits', *Electronic Engng*, **30**, p. 371, June 1958.
2. F. Rozner, 'Transistor multivibrator circuit', *Electronic Engng*, **29**, p. 455, September 1957.
3. K. Ehlers, 'Einfache Kompensation der temperatur-abhangigen Frequenzdrift eines Multivibrators', *Internat. Elektronische Rundschau*, **10**, p. 553, October 1965.
4. P. I. Bénéteau and A. Evangelisti, 'An Improved Emitter-coupled Multivibrator', Application Report 19, Fairchild Semiconductor Corp., April 1962, also:

   P. Alderisio and A. Evangelisti, 'A Fast Switching Emitter-coupled Multivibrator', Application Report 170, *ibid.*, August 1966.
5. E. C. Bell and D. Robson, 'Use of multivibrators in small telemetry systems', *Proc. Instn Elect. Engrs*, **114**, p. 327, March 1967.
6. J. H. Smith, 'Multivibrator circuits', *Electronic Engng*, **35**, p. 46, January 1963.
7. V. A. Ristic, 'A new type of free-running multivibrator', *Electronic Engng*, **36**, p. 232, April 1964.
8. D. T. Jovanovic, 'Multivibrator circuit using p-n-p and n-p-n junction transistors', *Electronic Engng.*, **31**, p. 301, May 1959.
9. G. H. Stearman, 'Transistor astable multivibrators for general purpose logic elements', *Electronic Engng*, **37**, p. 812, December 1965.
10. N. F. Moody and C. D. Florida, 'Some new bistable elements for heavy duty operation', *Trans. Inst. Radio Engrs on Circuit Theory*, **CT-4**, p. 241, September 1957.
11. N. C. Hakemian, 'PNP-NPN circuits: New look at a familiar connection', *Electronics*, **35**, p. 42, 24th November 1962.
12. B. D. Rakovich, 'A high-efficiency monostable circuit using complementary transistors', *Electronic Engng*, **39**, p. 384, June 1966.
13. M. V. Joyce and K. K. Clarke, 'Transistor Circuit Analysis', Chapter 15, (Addison-Wesley, London 1961).

Manuscript first received by the Institution on 16th October 1967 and in final form on 7th February 1968.
(Paper No. 1189/CC8).

U.D.C. 621.317 : 621.396.96

The Use of Noise Measurements in Radar Receiver Analysis

By

E. W. HOUGHTON,†

R. S. PETERS†

AND

M. W. SINCLAIR†

Reprinted from the Proceedings of Joint I.E.R.E.–I.E.E. Conference on 'R.F. Measurements and Standards' held at the National Physical Laboratory, Teddington, on 14th–16th November 1967.

Summary: The characteristics of band-limited random noise are employed to measure the signal transmission properties of non-coherent pulse radar receiving systems used for evaluating back-scatter signals from the ground, sea and precipitation clutter targets. A scheme using a calibrated microwave noise generator source and amplitude distribution and spectrum analysing equipment is employed, as both clutter and noise signals can be characterized by their amplitude distributions and spectra. Theoretical and measured results for a radar receiving system fitted with a logarithmic amplifier are compared and the effect of distortions explained.

1. Introduction

The back-scatter signals from wanted and unwanted radar targets usually fluctuate in amplitude. Because target signal fluctuations are complex and different for different types of target, it is necessary to measure them by experimental methods. These methods must include statistical tests, because the properties of fluctuating waveforms can only be described by probability theory. A knowledge of the magnitude and rate of these fluctuations, if used when selecting radar parameters, can lead to improvements both in target detection performance and discrimination between wanted and unwanted targets.

The term 'clutter' can be used to describe an important class of unwanted radar back-scatter from targets often covering a region much larger than a radar resolution cell. However, within each resolution cell, clutter may consist of many individual targets moving independently of each other, such as flocks of birds, vegetation, raindrops, and the surface of the sea. The resultant signal from these randomly and quasi-randomly phased targets is a fluctuating one. Clutter of this nature and random noise fluctuations have been linked from the earliest days of radar because of superficial resemblances between their appearance on an A-scope (after being fed through a band-pass receiver limiting their spectrum) and also due to the similarity of their mathematical treatment. There is, however, a closer relationship between fluctuating clutter and noise signals, which has an important bearing on the work described in this paper. The amplitude probability distributions of clutter generated by a large number of independent scatterers and of noise differ only by a multiplicative constant at the receiver second detector output.[1] Though the characteristics of some clutter are akin to those of random noise, the characteristics of other types of clutter can be quite different, and for this reason it is necessary to make clutter measurements. The radar receiving and recording system described in this paper converts received signal amplitude into digital data. The digital data are processed by using a computer to extract statistical information in the form of the amplitude distribution, which describes the magnitude of signal fluctuation, and the autocorrelation function (or its Fourier transform, the normalized power spectrum), which describes the rate of signal fluctuation.[2]

Before a receiver system can be used for measuring clutter it must be calibrated and checked for distortion. Because of the presence of the logarithmic receiver, the properties of the overall receiving system cannot be tested by using c.w. signal generator methods. A modulated test signal is required whose characteristics are similar to those of fluctuating clutter. This requirement may be satisfied by using a noise signal whose fluctuating characteristics are known. As previously stated, a randomly fluctuating waveform cannot be measured in the same way as a steady recurring waveform, but must undergo statistical checks. This is conveniently done in this case by processing the test signal in the same way as the clutter signals. However, unless an on-line computer is available, the time-delay involved in processing test data is, in general, far too long, and clutter trials are held up pending the outcome of computer results. Furthermore the very expensive radars used for collecting clutter data may only be allocated for short measuring periods. It is therefore essential to make a check on system performance as quickly as possible before collecting a mass of digital information, which

† Royal Radar Establishment, Great Malvern, Worcestershire.

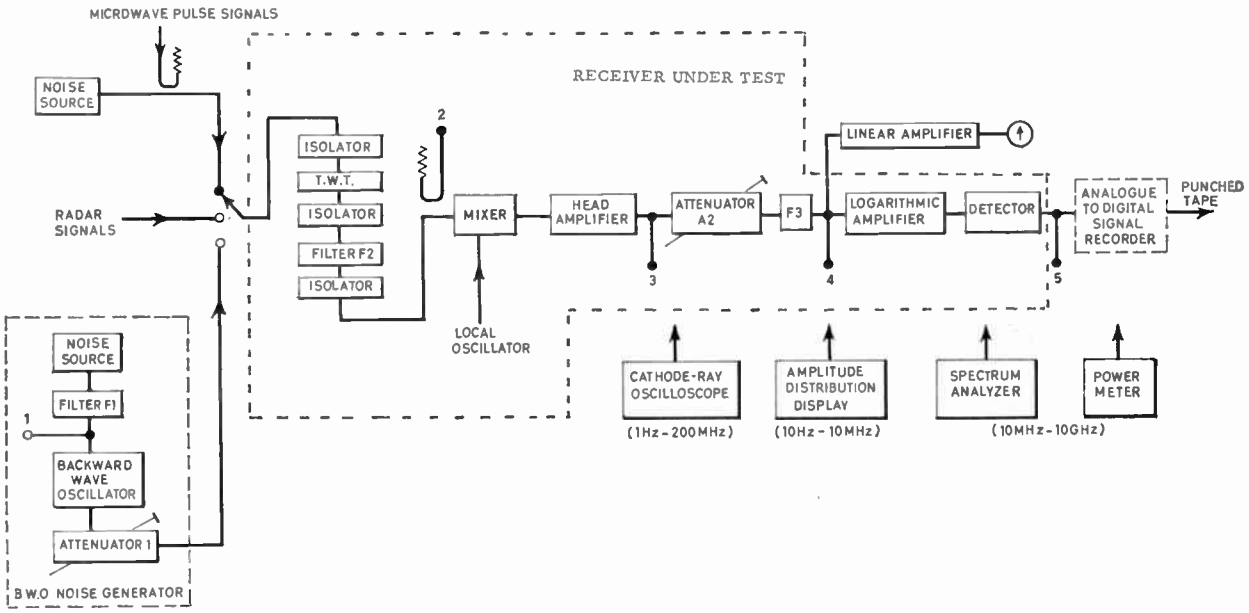


Fig. 1. Block diagram of signal measuring receiver channel and receiver performance measuring system.

will be worthless if the system has been running after being incorrectly set-up or with a fault condition.

This paper describes an experimental system for making such a check. The clutter measuring receiver is designed to yield data which, after computer processing, will define the target in terms of its amplitude density distribution and power spectrum. It is convenient and logical, therefore, to check for receiver distortion by measuring amplitude density distributions and power spectra on c.r.t. analysers.

2. Receiver Performance Measuring System

Five kinds of performance information are required for correctly operating a clutter measuring radar system:

(i) Confirmation that the radar is transmitting the correct waveform and spectrum, and no spurious signals.

(ii) Confirmation that the transmitted signal spectrum is correctly placed within the pass band of the radar receiver and the pulse shape is correct.

(iii) Check of the overall receiver noise temperature.

(iv) Calibration of signal level range (before and after trial) in discrete steps from a known and constant reference level (e.g. a gas discharge tube noise source).

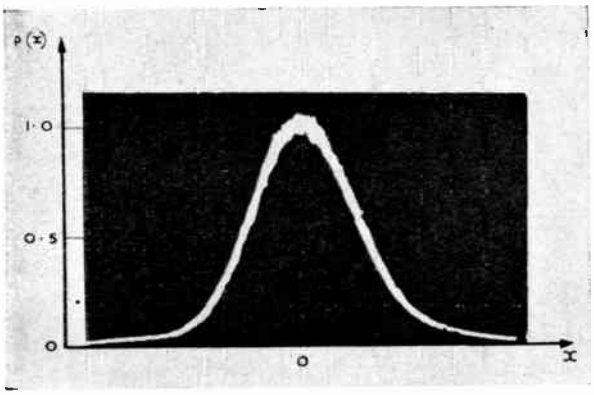
(v) Confirmation that when the radar receiver is fed by a known fluctuating signal, the expected amplitude distribution and spectrum functions are produced with the minimum distortion.

2.1. *Layout of Receiver Channel and Measuring System*

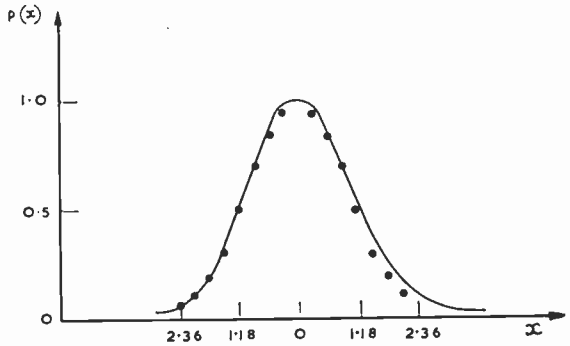
A block diagram of the receiver system employed for measuring clutter signals and of the experimental receiver performance measuring system is illustrated in Fig. 1. Radar signals at microwave frequencies are fed through the switch, isolator, travelling-wave tube and band-pass image rejection filter F2. The signals are converted to i.f. by a mixer and amplified by the head-amplifier. A matched i.f. attenuator A2 is inserted between the head-amplifier and narrow band-pass i.f. filter F3. The output of this filter is fed into a logarithmic i.f. amplifier, detected and then passed to the analogue-to-digital signal recorder. The output of the recorder in the form of punched tape, can be used, after processing, for signal analysis.

Alternatively, the switch can be connected to a coupler and a gas discharge noise source tube. The noise is used in measuring the noise temperature of the receiver and for setting the lowest calibration level. Microwave pulse signals of similar characteristics to those radiated by the radar transmitter are fed into the coupler and adjusted to provide a series of calibrated steps over the receiver signal range. A linear amplifier and meter, following the receiver i.f. filter, is provided for measuring noise temperature.

The third input switch position connects the radar receiver to a microwave backward wave oscillator (b.w.o.) noise generator, which is used as a source of fluctuating test signals both for the longer method of analysis using the analogue-to-digital recorder and computer, and the quick checking method to be described.



(a) Amplitude distribution display.



(b) Calculated points plotted on tracing from Fig. 2(a). Points are calculated from $p(x) = \exp(-x^2/2)$.

Fig. 2. Video output distribution of b.w.o. noise generator modulator.

3. Band-limited Random Noise Measurements

A detailed description of the measurements made when the input switch is connected to the third position is given in this section. A b.w.o. is used as a source of random noise of known characteristics, because it is capable of simulating the wide range of signal levels produced by clutter echoes. Such a test source is required both for the quick analyser method and the slower off-line computer analysis method of measurements. The generator output is fed into the receiver and then examined at monitor points 1, 2, etc., for spectrum and amplitude distribution distortions. Spectrum and amplitude distribution analysers, fitted with display graticules calibrated with known signal and signal distortion limit curves, are compared with the actual signal displayed by photographing with a Polaroid (instant) camera. Under normal operating conditions, after setting-up power levels, etc., it will only be necessary to look at the spectrum and amplitude distributions at the output of the receiver, intermediate points being used for tracing distortion and fault conditions.

3.1. *The Microwave B.W.O. Noise Generator*

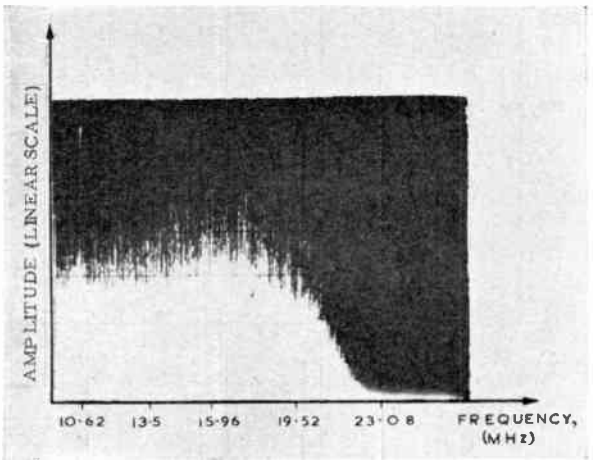
A noise source consisting of a number of paralleled Zener noise diodes generates a frequency spectrum flat to within 3 dB over a bandwidth of approximately 20 MHz. The amplitude probability distribution of the noise fluctuation follows a Gaussian law to within $\pm 10\%$. This video noise is amplified by a linear amplifier whose bandwidth is restricted to 8 MHz at the 3-dB points by a five-element Butterworth low-pass filter. The amplitude density distribution function of the noise output of the filter taken at monitor point 1 is shown in Fig. 2(a). The curve, indicating the relative portions of time the fluctuating voltage spends near specific values, is a photograph of an amplitude distribution meter display.[3] A tracing of Fig. 2(a) was made, and calculated points superimposed on it as shown in Fig. 2(b). The calculated values were based on the Gaussian curve equation

$$p(x) = \exp(-x^2/2)$$

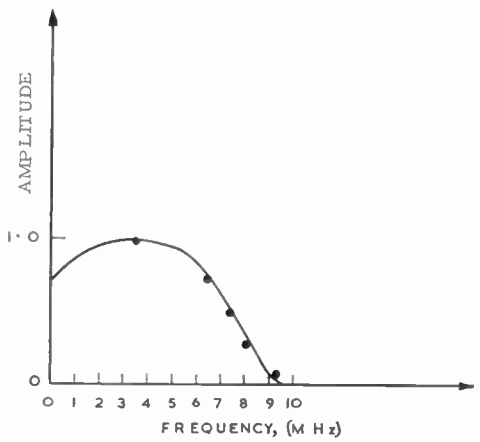
the vertical scale being normalized to the peak of the tracing and the horizontal scale normalized such that the tracing fitted the theoretical curve at the half-amplitude points. The subsequent points plotted with these scaling factors follow the tracing to within $\pm 10\%$.

The spectrum at monitor point 1 (Fig. 1) was measured using a spectrum analyser.[4] The frequency band of the analyser, as indicated in Fig. 1, commences at 10 MHz, and consequently, before it could be used to measure the video noise spectrum, it was necessary to shift the spectrum to a band above 10 MHz. The mixing and subsequent filtering were done at a centre frequency of 13·5 MHz. Figure 3(a) shows a photograph of a portion of the output spectrum together with a frequency marker of 10·6 MHz. The lower sideband of the output falls partly outside the spectrum analyser range, and measurements were therefore confined to the upper sideband. A tracing of the upper sideband envelope is shown in Fig. 3(b). The frequency scale was fixed using a marker in conjunction with a knowledge of the analyser frequency scale. Points were plotted on the tracing based on the product of the Butterworth low-pass filter response and the post mixer i.f. amplifier response. There is a good agreement between the calculated values and the practical curve except for the point between 9 and 10 MHz. The fall-off of the spectrum at the lower end of the curve is due to the combined effects of an intentional fall-off of the video amplifier and a 10% ripple on the post mixer amplifier frequency response.

The noise output of the video filter is fed on to the cathode of a microwave backward wave oscillator,



(a) Frequency spectrum.



(b) Calculated points plotted on tracing from Fig. 3(a).

Theoretical points calculated from:

$$z(\omega) = \frac{1}{(1+x)^{\frac{1}{2}}} \times z(\omega)$$

For practical b.w.o. 8 MHz modulation filter

$$x = \frac{1}{\delta^6}\left(\frac{f}{f_0} - \frac{f_0}{f}\right)^6$$

Fig. 3. Output of noise modulator low-pass filter.

whose cathode voltage/frequency characteristic is approximately linear. For sufficiently high modulation indices it has been shown by Middleton and Mullen[5] that frequency modulation of a carrier by a random stationary noise process results in a spectrum, which is approximately proportional to the first order probability density function of the modulating process. This has been shown to be a Gaussian approximation in this case. Stewart[6] has shown that if the frequency deviation of an f.m. signal is larger than the bandwidth of the modulating waveform, the shape of the f.m. spectrum is practically independent of the bandwidth and shape of the modulating spectrum. He shows that for f.m. produced by a modulating function having a rectangular power spectrum and for large values of D/B, the normalized power spectrum for f.m. approximates to the Gaussian expression

$$W_F(\Delta\omega) = \frac{A_0^2/2}{(2\pi D^2)^{\frac{1}{2}}} \exp(-\Delta\omega^2/2D^2)$$

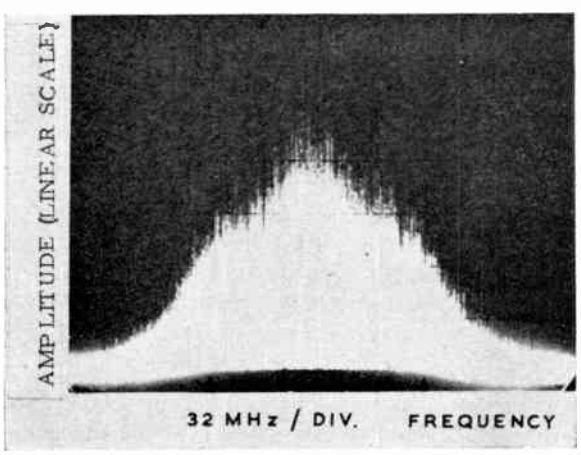
where D is the r.m.s. (angular) frequency deviation and is proportional to the square root of the average power of the modulating noise;

B is the bandwidth of the modulating waveform power spectrum in rad/s;

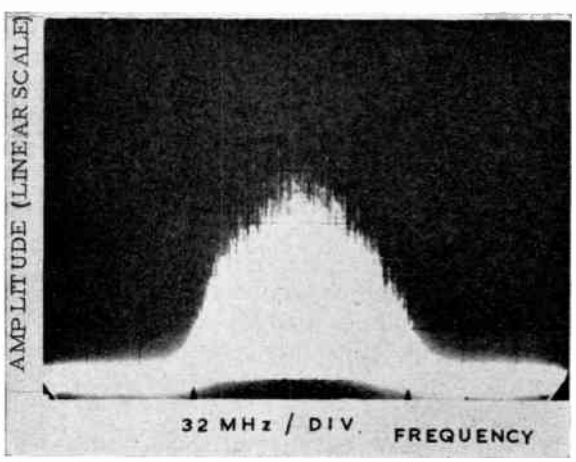
β is the modulation index ($= D/B$);

$\Delta\omega$ is the difference frequency from the modulated carrier;

A_0 is the peak amplitude of the f.m. wave.

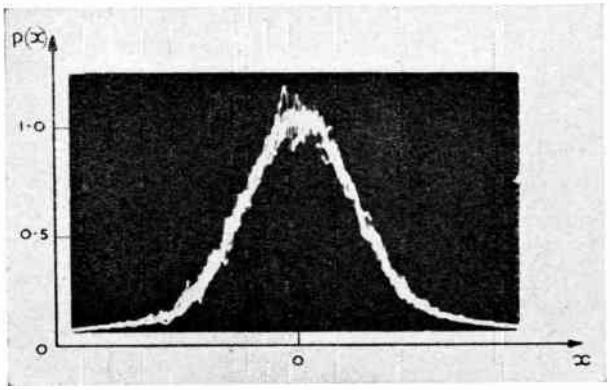


(a) B.w.o. noise generator microwave output.

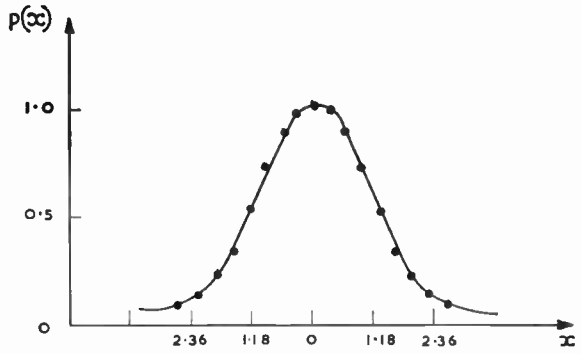


(b) Microwave bandpass filter output.

Fig. 4. Spectrum analyser display.

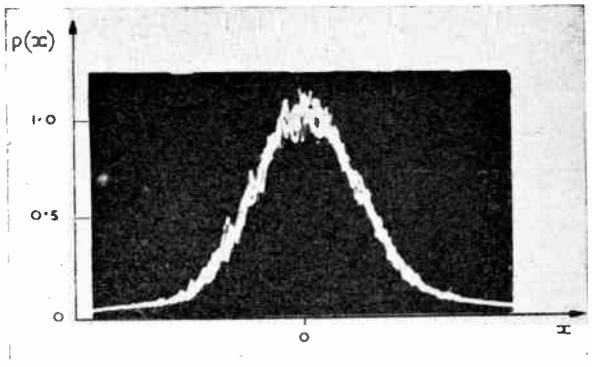


(a) Amplitude distribution display.

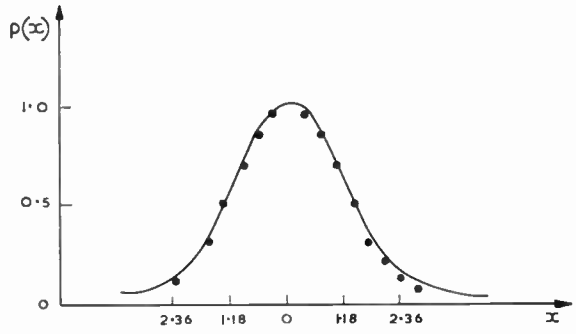


(b) Calculated points plotted on tracing from Fig. 5(a). Points are calculated from $p(x) = \exp(-x^2/2)$.

Fig. 5. Output distribution of i.f. narrowband filter (b.w.o. noise generator input).



(a) Amplitude distribution display.



(b) Calculated points plotted on tracing from Fig. 6(a). Points are calculated from $p(x) = \exp(-x^2/2)$.

Fig. 6. Output distribution of i.f. narrowband filter (t.w.t. noise source input).

Though it is possible to find a portion of the voltage/frequency characteristic on most b.w.o. tubes where an approximately Gaussian-shaped f.m. spectrum can be generated, complex ripples on the characteristic tend to generate departures from the ideal spectrum. The photograph in Fig. 4(a) illustrates a typical spectrum with a r.m.s. deviation of about 57 MHz and a peak to peak variation of 160 MHz. The effect of a discontinuity in the voltage/frequency characteristic shows up at 50 MHz to the right of the centre frequency. This spectrum hole is due to a b.w.o. fault, but similar ones can be produced by impedance mismatches in the output external circuit. The hole is well clear of the centre frequency and the required narrow receiver spectrum, and does not affect the use of the generator in this case.

The output power of the b.w.o. is monitored and a continuously variable waveguide attenuator is available, providing 110 dB$\pm$1 dB attenuation over the desired microwave band.

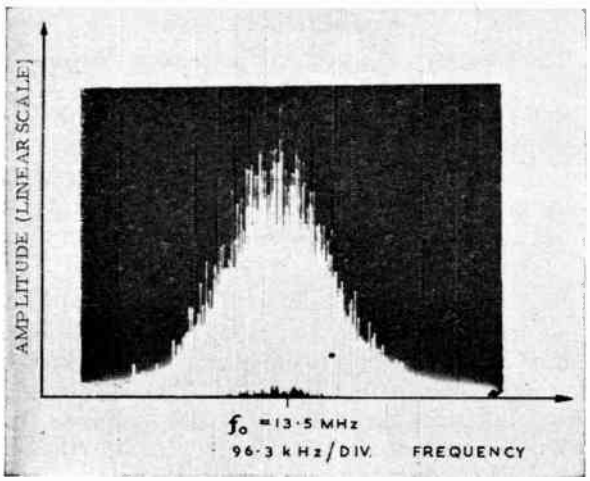
3.2. *Experimental Results*

The output spectrum of the b.w.o. noise generator is limited by the microwave band-pass filter F2 and the spectrum at the output of the filter is 60 MHz at the 1 dB points. This spectrum, monitored at point 2 (Fig. 1), is shown in Fig. 4(b). The receiver noise level fed into the narrow-band i.f. filter is monitored at point 3.

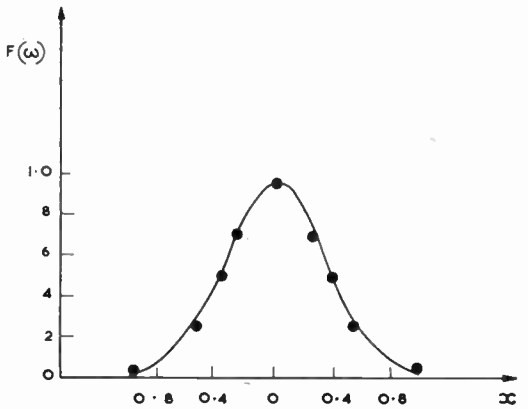
The amplitude distribution at the output of the narrowband i.f. filter when the receiver is fed by a b.w.o. noise generator is shown in Fig. 5(a) and this curve is compared with points plotted for the Gaussian function

$$p(x) = \exp(-x^2/2)$$

in Fig. 5(b). The errors lie well within $\pm 10\%$ for all parts of the curve. A photograph of the spectrum is shown in Fig. 7(a) and in Fig. 7(b) a comparison is made between a tracing from this curve and points taken from the calculated response of the narrow band



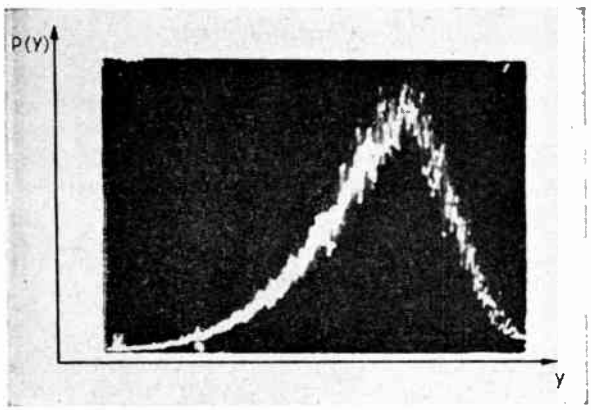
(a) Spectrum analyser display.



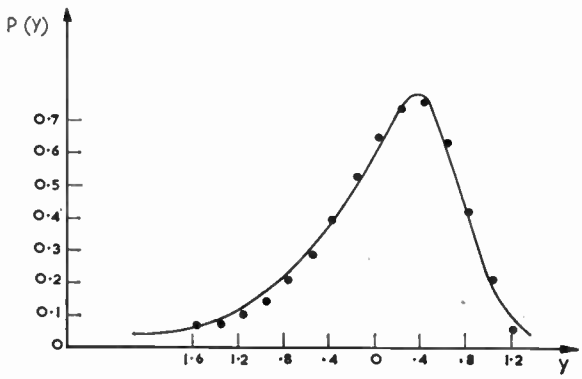
(b) Calculated points plotted on tracing from Fig. 7(a).
Points are calculated from

$$|F(\omega)| = \left(\frac{1}{\sqrt{1+x^2}}\right)^n; \quad x = Q\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right); \quad n = 3$$

Fig. 7. Output of narrowband i.f. filter.

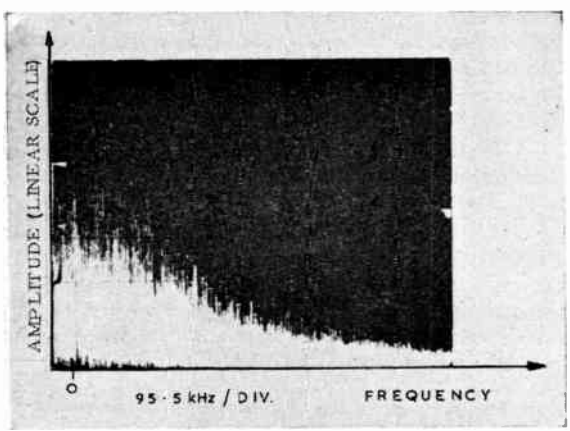


(a) Amplitude distribution display.

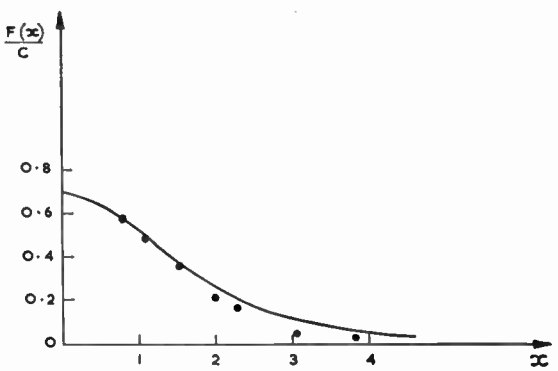


(b) Calculated points plotted on tracing from Fig. 8(a).
Points are calculated from $p(y) = e^{2y} \exp\left[-\frac{e^{2y}}{2}\right]$

Fig. 8. Video output distribution from cascaded i.f. logarithmic receiver.



(a) Spectrum analyser display.



(b) Calculated points plotted on tracing from Fig. 9(a).
Points are calculated from:

$$F(x) = 0{\cdot}61[c \exp -0{\cdot}35x^2 + 0{\cdot}132c \exp -0{\cdot}175x^2 + \\ + 0{\cdot}40c \exp -0{\cdot}117x^2 \ldots + \ldots]$$

Fig. 9. Video output spectrum of logarithmic receiver.

filter. In terms of the bandwidth of this filter (approximately 200 kHz at 3 dB points) the microwave spectrum after the microwave filter can be considered as approximately constant in amplitude.

As clutter signals are generated by an a.m. source it is necessary to check at the output of the narrowband filter that the b.w.o. source can replace an a.m. source. This check was made by replacing the b.w.o. noise generator by a gas discharge noise source followed by a wideband travelling-wave tube amplifier. There was no change in the amplitude density distribution function, as can be seen in Fig. 6, or in the spectrum.

The output of the narrowband filter is fed into a logarithmic i.f. amplifier whose input/output characteristic has an 80 dB dynamic range within ± 1 dB of an ideal logarithmic curve. The amplitude density distribution function at the video output of this amplifier is shown in Fig. 8(a). The theoretical function

$$p(y) = \exp(2y) \exp[-\exp(2y/2)]$$

for the distribution of an ideal logarithmic amplifier has been evaluated by Croney,[7] and this is shown plotted on the tracing from the photograph in Fig. 8(b). Agreement between the curve and points is better than $\pm 10\%$, except at the right-hand tail of the curve, where there are instrument limitations. The distribution function must remain constant throughout the logarithmic characteristic. In these experiments the only effect on the result shown in Fig. 8(a) was to increase the 'noisiness' of the trace in the region of the peak of the curve (this being due to the properties of logarithmic noise, the rate of change of amplitude with time becoming less at the peak as more noise is fed into the amplifier). In order to achieve results as shown in Fig. 8(b) it is also essential to ensure that the level of the input signal is well clear of the bottom bend of the logarithmic characteristic. The output video spectrum is converted into i.f., in the same manner as the video modulation for the b.w.o., before being applied to the spectrum analyser. A photograph of the spectrum together with a zero identifying marker is shown in Fig. 9(a). A tracing from this spectrum is compared with calculated points evaluated from the theoretical output spectrum function of an ideal logarithmic receiver[8] in Fig. 9(b). There is poor agreement between the theoretical points and the tail of the curve, because the theoretical spectrum is based on a Gaussian input spectrum, which has narrower skirts and a faster fall-off of the skirts than the spectrum of the narrow-band filter. A computer program is now being written for the theoretical spectrum, but based on the narrow-band filter input spectrum instead of the Gaussian one, and it is expected that this will give a better fit.

4. General Remarks

The spectrum analyser and amplitude distribution meter provide convenient means whereby the b.w.o. high level random noise source and the overall receiver performance may be checked relatively quickly. The experimental results quoted have demonstrated the feasibility of the method, and work is now proceeding to set up distortion limits which can be drawn on to the display graticules.

In general the presence of gaps in the spectrum and unwanted spurious signals, especially periodic ones, can be easily seen on the analyser, but distortions of the overall narrow-band spectrum due to mistuning effects, etc., are best revealed by photography (this can take about 20 seconds). Loss of gain, setting-up errors, and faults producing changes in the logarithmic slope can be detected using the amplitude distribution analyser. The only errors that matter, of course, are those that take the functions beyond their tolerance limits.

5. Acknowledgments

The authors are indebted to Messrs. W. Chandler, R. Barrett and S. Cobb for helpful discussions and data. Miss S. Weir designed the engineered version of the b.w.o. noise generator and Mr. L. M. Davies was largely responsible for the receiver experimental work.

6. References

1. J. Lawson and G. Uhlenbeck, 'Threshold Signals', M.I.T. Radiation Laboratory Series, Vol. 24, Chapters 6 and 11, (McGraw-Hill, New York, 1950).
2. D. Kerr, 'Propagation of Short Radio Waves', M.I.T. Radiation Laboratory Series, Vol. 13, Chapter 6 (McGraw-Hill, New York, 1951).
3. M. Drayson, 'An amplitude distribution meter', *Electronic Engng*, **31**, No. 380, pp. 578–584, October 1959.
4. H. Halverson, 'A new microwave spectrum analyser', *Hewlett-Packard J.*, **15**, No. 12, August 1964.
5. J. Mullen and D. Middleton, 'Limiting forms of f.m. noise spectra', *Proc. Inst. Radio Engrs*, **45**, p. 874, June 1957.
6. J. L. Stewart, 'The power spectrum of a carrier frequency modulated by Gaussian noise', *Proc. I.R.E.*, **42**, pp. 1539–1542, October 1954.
7. J. Croney, 'Clutter on radar displays: reduction by use of logarithmic receivers', *Wireless Engineer*, **33**, p. 85, April 1956.
8. S. Geldston, 'Probability Distribution at the Output of a Logarithmic Receiver', Research Report No. PIBMR1–1087–62, Electrical Engineering Dept., Polytechnic Institute of Brooklyn, Microwave Research Institute.

Manuscript received by the Institution on 1st September 1967. (*Paper No.* 1190/AMMS 12.)