The Journal of the
BRITISH INSTITUTION OF RADIO ENGINEERS

FOUNDED 1925 INCORPORATED BY ROYAL CHARTER 1961

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

VOLUME 24 AUGUST 1962 NUMBER 2

NOTICE OF ANNUAL GENERAL MEETING

NOTICE IS HEREBY GIVEN that the FIRST ANNUAL GENERAL MEETING of the Institution since Incorporation by Royal Charter will be held on WEDNESDAY, 24TH OCTOBER 1962, at 6 p.m. at the London School of Hygiene and Tropical Medicine, Keppel Street, Gower Street, London, W.C.1.

AGENDA

1. To note the Minutes of the 36th and Final Annual General Meeting of the Institution Incorporated under the Companies Act 1929 and 1948, held on 24th January 1962 (reported on pp. 94–96 of the February 1962 *Journal*).
2. To receive the Minutes of the first Special General Meeting of Corporate Members of the Chartered Institution held on Wednesday, 23rd May 1962 (as reported on page 116 of this *Journal*).
3. To receive the Annual Report of the Council for the year ended 31st March 1962. (The Report will be published in the September issue of the *Journal*.)
4. To receive the Auditors' Report, Accounts and Balance Sheet for the year ended 31st March 1962. (The Accounts and Balance Sheet will be published in the September issue of the *Journal*.)
5. To confirm the election of the Council for 1962/63.

In accordance with Bye-law 43 notice of Council's nominations was circulated to Corporate Members on 27th June 1962 and was published in the June issue of the *Journal*. As no other nominations have been received under Bye-law 44, a ballot will not be necessary and the following will be elected:—

President:

Admiral of the Fleet the Earl Mountbatten of Burma,
K.G., P.C., G.C.B., G.C.S.I., G.C.I.E., G.C.V.O., D.S.O., LL.D., D.C.L., D.Sc.

Vice-Presidents:

L. H. Bedford, C.B.E., M.A., B.Sc., F.C.G.I.
A. A. Dyson, O.B.E.
W. E. Miller, M.A.
Colonel G. W. Raby, C.B.E.
J. L. Thompson
Professor Emrys Williams, Ph.D.
Professor Eric E. Zepler, Ph.D.

Honorary Treasurer: G. A. Taylor (Member)

Ordinary Members of Council:

MEMBERS

Brigadier L. H. Atkinson, O.B.E., B.Sc.
J. R. Brinkley
H. E. Drew
R. H. Garner, B.Sc.
I. Maddock, O.B.E., B.Sc.
A. St. Johnston, B.Sc., A.C.G.I.
H. F. Schwarz
A. G. Wray, M.A.

ASSOCIATE MEMBERS

W. A. Gambling, Ph.D., B.Sc.
Squadron Leader W. L. Price, O.B.E., M.Sc.
D. L. Leete, B.Sc.

COMPANION: J. N. Toothill, C.B.E.

6. To appoint Auditors and to determine their remuneration. (Council recommends the appointment of Gladstone, Jenkins & Co., 42 Bedford Avenue, London, W.C.1.)
7. To appoint Solicitors. (Council recommends the appointment of Braund & Hill, 6 Grays Inn Square, London, W.C.1.)
8. Awards to Premium and Prize Winners.
9. Any other business. (*Notice of any other business must reach the Secretary not less than forty-two days prior to the meeting.*)

INSTITUTION NOTICES

The First London Meeting of the 1962–63 Session

The opening meeting of the 1962–63 Session is being sponsored by the Television Group Committee. It will be held on Wednesday, 26th September at 6 p.m. at the London School of Hygiene and Tropical Medicine, Keppel Street, Gower Street, London, W.C.1, and will take the form of a discussion on the technical implications of the report of the Committee on Broadcasting. Full details of the meeting, including the names of the opening speakers, will be circulated to members of the Television Group a few days before the meeting.

British Nuclear Energy Society

The Institution has been accepted as a constituent member of the British Nuclear Energy Society whose main object is the advancement and dissemination of knowledge concerning nuclear energy and ancillary subjects.

In accordance with the constitution of the British Nuclear Energy Society, members of the Brit.I.R.E. may now be admitted as members of the Society provided that, at the time of their application, they are engaged in professional, scientific, or technical study of nuclear energy and ancillary subjects. Membership of the Society does not imply any additional technical qualification but enables the member to attend any meeting of the Society or constituent body.

Institution Meetings 1962–63

The usual programme card giving details of meetings of all Sections of the Institution in Great Britain for the first half of the 1962–63 Session will be sent to members towards the end of September. The Council would like to emphasize that a member may attend any meeting of the Institution, irrespective of his present professional activities or his normal place of residence.

Details of Group meetings are however only sent to those members who have requested that they should be affiliated to particular Groups. Similarly details of Local Section meetings are only circulated to members normally resident within the geographical area covered by the Section.

Index to Volume 23

The index to Volume 23 of the *Journal* (the first half of 1962, January to June) has now been prepared and copies are being sent with this issue to all members and subscribers.

Members are reminded that they may send their *Journals* (six issues plus index) to the Institution for binding. The charge for this service is 16*s.* 6*d.*, postage extra (Great Britain 3*s.*; other countries 4*s.*).

Institution Prize for R.A.F. Apprentices

The Council of the Institution has established a Brit.I.R.E. Prize, value £5, for the most outstanding apprentice passing out of each course at the R.A.F. No. 1 Radio School, Locking.

The first prize has been awarded to Sergeant Philip Frederick William Hutchins who was the most outstanding apprentice in the 92nd Entry. He was also awarded a cadetship to the R.A.F. Technical College, Henlow; after completing his training at that establishment he will be able to take examinations for a university degree or a Diploma in Technology.

Sergeant Hutchins gained a total of five prizes, which is quite an outstanding achievement.

Circulation of the *Journal*

The Audit Bureau of Circulations has recently announced that the average monthly circulation of the Institution's *Journal* for the first six months of 1962 was 9580 copies per issue. This represents an increase of 669 copies compared with the previous six months and 933 copies compared with a year ago. The present rate of increase comes not only from the increase in Institution membership but also from research establishments and universities throughout the world who subscribe to the Institution's *Journal*. It is anticipated that the A.B.C. figure for the second half of 1962 will reach a minimum of 10,000 a month, being the *Journals* actually sent to members and subscribers during those six months.

Engineering Institutions Part 1 Examination

The Council of the Institution has applied for membership of the Engineering Institutions Part 1 Examination Committee. This Committee brings together all the engineering Institutions' interest in holding jointly an examination common to the requirements of the engineering Institutions in basic engineering subjects. The Brit.I.R.E. has for long maintained that this is a very desirable course and the Council is therefore pleased to have the opportunity of joining the Institutions of Civil, Mechanical, Electrical, Municipal and County, Marine, Gas, Production, and the New Zealand Institution of Engineers in this joint endeavour.

A further announcement will be made in the autumn but the present intention is that candidates intending to qualify for Graduate or higher grade membership in the Institution will take the Engineering Institutions Part 1 Examination instead of the present Brit.I.R.E. Section A. The schedule of exempting qualifications will be in common agreement with the Engineering Institutions Part 1 Committee.

U.D.C. 621.37/8: 362.41

Print Recognition Apparatus for Blind Readers

By

J. H. DAVIS, Ph.D.†

Presented at the Symposium on "Practical Electronic Aids for the Handicapped" in London on 28th March 1962.

Summary: The facilities offered both to designer and reader by the auditory and tactile types of reading machine are compared. Methods used for solving the problem of producing an electrical signal equivalent to each of a set of printed characters are discussed and some actual machines briefly described. The types of machine required by the blind and facilities to be given by the ultimate designs are set out. The problems of recognition of the electrical signal and storage and accessibility of the sound tracks in the auditory machine are discussed. Techniques required, some existent, others not yet available, are considered. Suggestions are offered as to the methods which may be ultimately used to solve some of the problems.

1. Introduction

After Milton became too blind to read for himself, he taught his three young daughters to read to him in Latin, a language which they did not understand. As they eventually refused to continue this service, which they must have found supremely irksome, one wonders whether any of them conceived a wish for some kind of "engine" (in 17th century parlance) which would read to their father and release all three from what they considered to be drudgery.

At whatever period the idea of a machine to read printed matter aloud to the blind originated, no attempt to realize it could hope to succeed until after certain discoveries had been made in physics, and until these had been applied in the development of apparatus.

Apart from the fundamental phenomena of electromagnetism, the most important of these discoveries and subsequent developments were:

(1) The discovery of the interchange of energy between visible light and electricity, leading to the development of the photo-electric cell.

(2) The development of the principle of scanning.

(3) The invention of the process of recording sound upon wire, and, later, tape.

All these events occurred before 1900 and to them we should add, dating from 1905:

(4) The invention of the triode wireless valve.

The most important work on the problem of producing a reading machine has so far been limited to attempts to design a machine reading figures or letters only. All the devices described in the literature depend upon two or more of the four developments mentioned above. A machine to read words has not yet appeared, and this paper will show that its arrival in a practicable form must be preceded by two or three developments still to come.

† Department of Physics, Bradford Institute of Technology.

2. The Idea of a Print-Recognition Machine

The basic idea of a print-recognition machine is of a machine which will perform a particular action in response to the stimulus of a particular character printed upon a sheet of paper.

Such a machine will perform a limited number of different actions, each of which is selected when a particular character appears. Thus, each of the twenty-six letters of the alphabet may individually cause the machine to "speak" the name of that letter aloud. Alternatively, the machine may emboss the Braille equivalent of the letter on paper. In certain print-recognition machines now available commercially, and used in accountancy, the actions consist of the setting up of digits in a computer, or accounting machine of some kind.

In the present paper we are concerned solely with the print-recognition machine as a means of enabling the blind to read a normal printed book.

3. Methods available for Presenting the Information to the Blind Person

The problem here is that the blind person requires a machine which will automatically convert the printed matter on the paper into a form which can be easily appreciated by one of his remaining senses. Of these only the senses of hearing and touch form practicable channels for conveyance to his brain of the information contained in the printed word. This gives rise to machines of two kinds, the one giving an audible output and the other a tactile one.

The particular forms which the output from each type of machine may take may be listed as follows:

I. Tactile machines.

(1) Braille, Moon, or other commonly used raised characters embossed on tape of paper or metal foil.

(2) Braille simulated by plungers raised momentarily through a metal plate.

(3) Plungers raised under the hand in positions used in the "alphabetic glove".

(4) Normal alphabet embossed in either Roman or coded characters.

II. Auditory machines.

(1) Single notes and chords of notes.

(2) Letters and words spoken from a sound-recording.

(3) Synthetic speech.

Nearly all these outputs have been cited in machines already patented.

4. Critical Comparison of Auditory and Tactile Machines

The sense of touch is used by the blind in reading Braille and Moon signs which can be embossed on suitable paper. Reading by touch is normally a process of reading single signs or small groups of signs, and is comparatively slow, although a good Braille reader can achieve the same speed as a sighted person reading aloud.

For persons who are both deaf and blind a machine with a tactile output delivered letter by letter is the only one possible. In this case the output could take one of the four forms mentioned above and which will now be discussed.

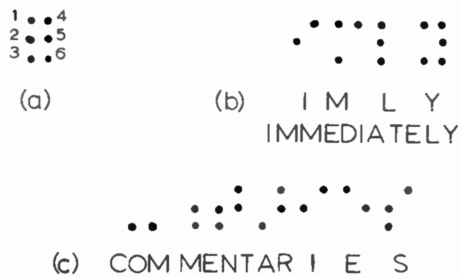


Fig. 1. Standard English Braille. (*a*) Numbering of Braille dots. (*b*) and (*c*) Two contracted words.

First, a paper tape could be embossed in Braille, or Moon by the machine as it reads.

Second, a metal plate pierced with six holes of the standard Braille dimensions and spacing would be used. Plungers raised momentarily through the plate would simulate Braille letters, one at a time (Fig. 1).

This type of output would be tiring to read for long periods, and perhaps slightly frustrating, being presented a letter at a time. The extended output on a long tape would be easier and less tiring, but expensive in paper. It might be possible to use metal foils in continuous loops, rolling out the impressions after reading, but it is questionable whether this would afford much reduction in cost or weight. Paper tapes, and simulation by plungers have both been used in a type of blind-deaf communicating (not reading) machine (the "Arcaid").

Unfortunately, most blind-deaf persons are unable for one reason or another to read Braille. Communication with some of these people is made possible for sighted persons by use of "alphabetic gloves". These are thin gloves marked in particular places with the letters of the alphabet. Pressure of the speaker's fingers at the appropriate points communicates words, letter by letter, to the blind-deaf person. For the blind-deaf who do not read Braille the tactile output from the reading machine would thus take the third form consisting of the raising of large plungers under the hand outstretched on a metal plate through which the plungers would rise.

The special output of the Visagraph is referred to later, and forms the fourth possible output.

For the person who is blind, but not deaf, an audible output in the form of words would obviously be best in every way. But, as will be apparent later, this is the most difficult type of output to achieve, and would require a comparatively large machine. An audible output in letter-form is technically simpler but makes reading slower. A machine giving this type of output might be made small enough to be portable. This would be an obvious advantage of this type of output.

A tactile machine giving letter by letter output could probably be made smaller and more easily portable than a similar type of auditory machine. A tactile machine has an additional advantage which would be important in many cases. Some blind persons, when in quiet surroundings, feel the atmosphere to be menacing because they are receiving no significant external impressions through either eye or ear. To relieve this they often work with a window open to admit traffic noise or turn on the radio quietly. A tactile reading machine would leave the ears of the reader exposed to room-noise, so putting his mind at ease as far as the external world was concerned. Further, he would be capable of having his attention attracted by his friends through his hearing in the natural way, instead of by the laying of a hand on his arm which is more likely to make him apprehensive.

Practically all auditory machines so far proposed or built have delivered their outputs, not in the form of the spoken names of the letters, but as tones, chords, or coded notes of some kind. Machines of this kind are very difficult for the user, and may be impossible if he lacks a "musical" ear. In one case of a machine which, in itself, worked well—the "Optophone"—it appears that only an extremely small number of people were able to use it successfully. One modern machine giving this kind of output was due to Flory and Zworykin and it is claimed that it was so small as to be effectively portable. One may speculate that a person who enjoyed listening to music and who was a good reader of, say, the Optophone, would find his enjoyment impaired by his acquired habit of associating letters with particular chords. This is not as surprising as it may, at first, appear to be because it is known that telegraphists used to reading Morse Code find themselves trying to read such "signals" as the noise of escaping steam.

"Synthetic speech" outputs seem to be out of the question at present because of the large volume of apparatus required. This type of output would also raise the problem of changing the acoustic equivalents of the vowels and certain groups of letters according to the preceding consonants, as in such groups of words as "bone", "done", "gone", and "cough", "rough", "ought", etc. The same difficulty occurs if it is proposed to build up the spoken words from the phonemes of the language.

It is appropriate now to describe two early reading-machines—one auditory and the other tactile.

5. Two Early Reading Machines

5.1. *The "Optophone" of Fournier d'Albe*

The first successful reading machine of any kind to be described was demonstrated by the inventor E. E. Fournier d'Albe[1] at the British Association meeting of 1913 and was called the "Optophone" (Fig. 2).

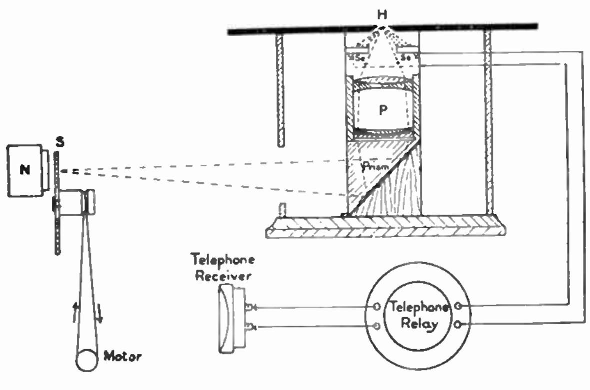


Fig. 2. The "Optophone" of E. E. Fournier d'Albe.

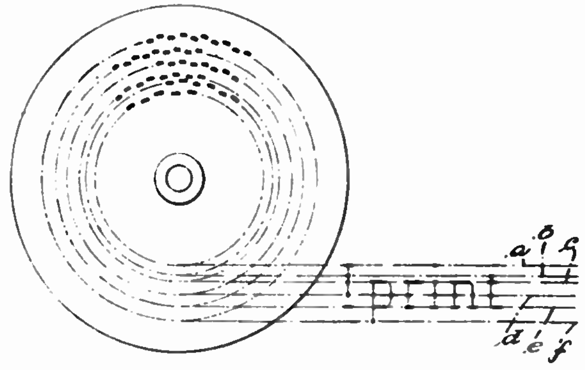


Fig. 3. The "Visagraph" of R. E. Naumburg.

This machine scanned each letter, H, in turn by means of a rotating disc, S, pierced with eight concentric circles of small holes, the larger circles having more holes than the smaller. An electric lamp, N, with a straight-line filament was arranged behind the disc with the filament located radially. The spacing of the holes was such that only one hole at a time was illuminated. The prism and lens, P, projected an image consisting of a vertical line of eight periodically illuminated holes on to the letter to be scanned. As the disc revolved the eight circles of holes interrupted the light at eight different frequencies. Light was well reflected from the white paper but much less well from the black parts of the letters. The reflected light was received by two selenium cells, Se, the output from which, heard in a telephone receiver, consisted of a chord of eight notes when white paper was scanned. When a letter lay under the line of spots, one or more notes corresponding to the blacks would be missing. Thus each letter produced its own characteristic chord. The printed matter was traversed steadily under the scanner and it will be apparent from the asymmetry of the letters that the characteristic chord must have varied as the letter moved.

5.2. *The "Visagraph" of R. E. Naumburg*

The first tactile machine was patented by Naumburg[2] in 1932 under the name of the "Visagraph". In this machine (Fig. 3), the scanning disc was provided with six concentric circles of short slots spaced equally. Thus, each letter was scanned by six spots of light regularly interrupted at different frequencies as in the Optophone. But in the Visagraph the output from the six selenium cells was fed to six amplifiers. Each amplifier controlled a solenoid which caused a plunger to emboss a dot upon thin aluminium strip. When blank white paper was scanned all six plungers were held back. When one of the light spots fell on black the associated solenoid allowed its plunger to emboss a dot on the aluminium and this dot was elongated (as the strip moved), for as long as the particular light-spot remained on black. Thus a crude representation of the letter scanned was embossed.

Figure 4 shows letters "o", "f", "t" as embossed by the Visagraph.

These two machines were successful as far as actual operation was concerned. They were crude, but succeeded in "reading" letters and figures. The Visagraph must have been slow in use, although probably easier to read than the Optophone.

In both these machines the use of a scanning disc rotating at a fairly high speed is objectionable for several reasons. But, easy though it is to criticize these inventions, they made an important initial attack on the problem of devising a machine of each type.

A number of other patents exist for machines producing an output either in Braille or in the form of audible tones.

Butler-Burke's[3] patent (1925) was in principle ingenious, but internal evidence indicates that it is unlikely that it was ever actually constructed in the form described. The output was to be delivered by an electric typewriter or a Braille machine.

Flory's[4] machine required the input to be in the form of a paper tape punched in a special manner, but he claimed an output of synthetic speech. Mauch's[5] machine (1959) used a specially prepared 16 mm ciné film for the input. The output was in the form of a continually changing chord.

From an examination of these and other patents it is apparent that the problem which has presented the greatest difficulty so far is that of deriving the information from the printed paper, and converting it into an electrical or mechanical form.

Let us now consider in more detail methods so far tried for solving this problem.

6. Methods Proposed for Deriving the Information from the Printed Paper

The following methods have so far been proposed:

(1) Scanning by Nipkow or similar disc, or cathode-ray tube, using
   (a) small number of spots with separate output corresponding to each spot,
   (b) full scan by single spot or one spot at a time.

(2) Use of matrix with shapes of characters or parts of characters stencilled or cut through, images of the printed characters being fitted on to the cut-outs.

(3) Formation of an image of the character on a matrix of photo-electric cells, using

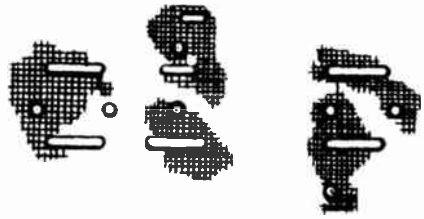


Fig. 4. Letters "o", "f", "t", as embossed by the Visagraph.

   (a) two-dimensional matrix,
   (b) one-dimensional, i.e. straight line matrix.

(4) Scanning by flying-spot scanner of oscillating mirror type in conjunction with an oscillator of variable pitch.

Other methods have been devised for use in special cases. For instance, machines feeding the input to accounting machines in some cases use magnetic-ink spots incorporated in the printed character. The E.M.I. machine FRED uses a flying-spot scanner with specially shaped figures. Obviously, devices of this kind do not help in the problem of reading already existent printed books.

Patents also exist for machines to read figures only, in one particular type-face printed on, say, a particular type of index-card, or bank-cheque, only. Restrictions such as these make more practicable the use of the two-dimensional matrix method.

7. Some Recent Print-Recognition Machines

The actual operation of these different methods of converting the information conveyed by the printed character into an electrical form is best appreciated by considering some reading machines which have used or proposed to use them. (In the paragraphs following the references are to the list of the previous section.)

The Optophone and the Visagraph are examples of machines using rotating-disc scanners in which the character is swept by a fixed number of light-spots (1a), the output from the passage of each spot being then separately used to control some device. In the Visagraph the device is one of a set of embossing plungers each operated by a solenoid. In the Optophone the notes of the chord are separated by the brain of the listener.

Zworykin, Flory and Pike[6] also used eight spots but generated them on the face of a cathode-ray tube (1a). The spots traversed the letters in turn and the photocell produced eight separate signals called by the inventors "interruption-patterns", which were then fed sequentially to eight "channel-counters", the output from these being then analysed in a type of computer.

A full scan by Nipkow disc (1b) with a relatively large number of lines is cited in Shepard's[7] patent. In this case it was proposed to record the signal from the photocell on magnetic tape, and then to compare this with standard recordings of the characters on a magnetic drum. A similar method using 24 spirals of square apertures and a memory-store is the basis of the I.B.M. 1418 Character Reader.

It has been proposed to use rotating discs called matrices with the shapes of the characters cut or stencilled through the metal (2). Alternatively, the disc is pierced with short slots and holes in positions corresponding to certain limbs of the characters.

In the first case, an image of the character is focused on a square aperture, and as the disc rotates, the particular cut-out which, when in position centrally over the aperture passes no light from the white parts of the character, reduces the photocell current to a minimum. This condition then causes selection of the corresponding output signal. This method was used by Handel.[8] Tauschek[9] used it for figures only, which were stencilled through a cylinder.

The main part of Shepard's[7] patent cites the second matrix-disc method. As the disc rotates nine different combinations of slots and holes are placed in turn over the image of the character. For each combination which gives a minimum photo-cell current, a corresponding relay is operated. Then the particular combination of relays operated is uniquely representative of the particular letter being examined.

The greatest difficulty of this type of recognition lies in the fact that extreme accuracy of location of the image relative to the matrix is essential at the instant at which recognition is attempted. This difficulty is very great and, in fact, the British Tabulating Company's[10] patent cites the use of a matrix carrying a dozen or more stencils of the same character, each stencil having a slightly different registration.

A similar principle to the use of a matrix is that in which an image of the character is formed on a rectangular array of small photocells (3a). This again requires very accurate registration but appears in general to be a little less exacting in this. The method has been used by Jannopoulo[11] and others.

Glaubermann[12] used a set of photocells arranged in a single straight line (3b), for figure-reading only. A magnified image of the figure is projected on to the cells, the outputs from which are sequentially examined. The first examination checks that registration of the image is correct as indicated by the two end-cells receiving full white. After this, as the figure is traversed, the outputs are examined for length of pulse; that is, the horizontals of the figures, which give long pulses, are picked out separately from the verticals, which give short pulses. Finally, a number consisting of two digits, and characteristic of each figure, is produced. The first digit gives the number of short-blacks and the second the number of long ones.

Several workers have used flying-spot scanning generated by an oscillating mirror (4). One of these was Flory in an improvement of a method originally due to Zworykin. In this machine each time the spot is traversed up the letter an oscillator produces an audio-frequency signal of increasing frequency. The frequency is returned to its minimum value on the flyback of the spot to the foot of the character. This machine is described in more detail in Section 8.1.

A brief comparison of the above methods of obtaining the information from the printed page is given later in considering the optimum design.

8. The Optimum Solution of the General Problem

The various types of output which existing patents specify are determined by the different purposes for which the machines are intended. The different inputs cited show, above all, the difficulty of producing the device which has the function of actually recognizing the printed character. In most cases, in which an actual printed input, rather than tape or film, is cited, this is limited to the ten numerals. In view of this situation, the writer is of the opinion that it would be helpful to prepare a statement of the types of reading-machine required for the use of blind persons, with the facilities to be given by each.

8.1. *Types of Reading-machine Required*

(1) A high-accuracy machine to reproduce printed matter in Braille and in Moon for the automatic production of the sheets for Braille and Moon books and periodicals.

(2) A small machine to reproduce the same printed matter letter by letter in some form of embossed character on tape for reading by blind-deaf persons in their own homes.

(3) A small portable machine reproducing letter by letter with either auditory or tactile output for use in the office or work-room, and when away from home.

(4) A machine to read books and newspapers printed in any normal type-face, reproducing the text word by word for the blind person in his own home.

Let us consider these machines separately in more detail.

The high-accuracy machine is required to produce automatically the sheets for Braille books and periodicals from an input of the printed text in a normal book or periodical. The method used at present is a hand-process and is, of course, slow and expensive. Such a machine would be even more useful if it were the means of producing a Braille daily newspaper, or a Braille edition of each of the printed daily papers. At present the blind are limited to a small number of monthly or weekly magazines, and the Braille *Radio Times*.

The Braille used today is said to be "contracted", that is to say, a considerable number of common words, and combinations of letters, as "but", "from", "go", and "ch", "er", "con", are each represented by a number of Braille signs smaller than the number of letters. This is done in accordance with an elaborate system of rules. It implies that the high-accuracy machine will have to contain a switching mechanism which will have the function of recognizing combinations of letters and of substituting the proper "compound-sign" for the multiplicity of letters of the word. Recognition of the contracted words could be carried out by arranging for the initial letter to engage the switching mechanism and for this to become active on particular combinations of letters following. Techniques for accomplishing this type of operation are well-known in automatic telephony and need present no difficulty. The number of high-accuracy machines required will not be large, so that the considerable volume of apparatus required for contraction will be less objectionable. The same mechanism could also carry out such operations as insertion of the numerical sign before numbers, and indicating when the text contained an unrecognized letter from a non-Roman alphabet, or any other strange character.

Once the actual print-recognition problem itself is solved it should not be difficult to produce such a machine as this, together with such refinements as automatic turning of the pages of the printed book, and changing of the Braille sheets after embossing.

The second machine required is merely the simplest possible form of the high-accuracy machine, and would simply transliterate from letterpress into uncontracted Braille on tape. The Visagraph is probably the nearest approach so far to this.

The third type of machine which is needed is a small portable one with either auditory or tactile output delivered letter by letter. This machine would have the object of reading the short passages which occur in bills, notices, the telephone directory, and so on. Such a machine should extend the range of work which the blind can do in industry beside sighted people.

Zworykin and Flory[13] have produced a device which, according to their paper, may be taken as an auditory form of this machine. A flying-spot scanner of the oscillating mirror type is used, and the light reflected from the paper is reflected through a lucite rod to a photo-cell, this part of the apparatus having the form of a probe which is held in the hand and so traversed over the paper. As the spot moves up the letter, an oscillator produces an audio-frequency signal of rising pitch. The blind person uses a telephone receiver to which is delivered a note of frequency dependent upon the distance of the spot from the start of its vertical movement and of intensity which is reduced when the spot lies on black. It is stated that the listener actually hears a series of thumps and clicks which he has to interpret. The weight of the apparatus with batteries is given as $5\frac{1}{4}$ lb and the authors claim that a reading speed of ten words per minute has been attained.

The fourth machine may be said to represent the ultimate objective in reading machines for the blind. It is intended to recognize not only letters but words and to simulate as nearly as possible the reading of a sighted person. The nature of such a machine is best indicated by means of a formal "schedule of facilities" which the writer suggests may be as follows:

8.2. *Facility Schedule of the Ideal Reading Machine*

The machine should offer the following facilities:

(1) To recognize 26 capital letters, 26 small letters, 9 numerals, the question and exclamation marks, blank spaces of one "em" in length and of more than two "ems", and the additional letters used in certain foreign languages, and to produce an electrical signal uniquely characteristic of each symbol to be recognized.

(2) To recognize the characters when printed in black on white or cream paper and in any typeface normally used for the text of a printed book or newspaper.

(3) To recognize whole words when printed as in facility (2), and to produce an electrical signal uniquely characteristic of the word recognized.

(4) To contain a vocabulary of sound-recordings of the words and names of the letters to be recognized, and to reproduce audibly from such vocabulary each word recognized and the name of each letter when printed separately.

(5) To reproduce audibly letter by letter any printed word for which the vocabulary contains no corresponding sound-recording.

(6) To give no audible signal in response to the question and exclamation marks and any other characters not required in reading by electro-mechanical means.

(7) To reproduce the words at a normal average reading speed of four words per second.

(8) To pause on reaching the blank space at the end of each sentence.

(9) To traverse the line of printed characters laterally before the photosensitive mechanism, and to traverse it in the perpendicular direction line by line when the longer blank space at the end of each line is recognized.

(10) To give an audible signal on reaching the blank line or a horizontal black line at the foot of the printed page or section thereof.

(11) To focus and align automatically the line of printed characters to be read.

(12) To be capable of being set in operation, adjusted as necessary, and stopped by a blind person.

It is unnecessary, of course, to stress that, in addition to the above, it is desirable to keep size, weight, power consumption and cost of production to a minimum, and to make the machine as reliable, and as simple to maintain as possible.

8.3. *Techniques now available which may be applied in the ultimate design*

It seems highly probable that, in the ideal machine, some form of flying-spot scanning will be used. Arrays of photocells are bulky and inclined to be expensive. Rotating discs, whether used as matrices or as Nipkow scanning discs are undesirable for several reasons and the associated apparatus is massive. It appears therefore that the large machines may use a cathode-ray tube scanner and the smaller, perhaps, an oscillating-mirror type. Scanning by either a single spot, or a small number of spots also allows the use of a larger and more sensitive type of photocell than does the rectangular array in which the smallest possible size is essential.

The flying-spot type of scanner takes all the information available from the paper. In fact, the amount of information it conveys is embarrassingly large, and the problem arises of sorting out what, above noise-level, is useful, that is, characteristic of the letter scanned, from what is not uniquely characteristic of that particular letter. This, and the problem of locating the letter exactly within the raster, are the two problems met in using this type of scanner.

When, by any means, an electrical signal uniquely characteristic of the letter to be recognized has been generated, several methods have been used to identify the character:

(1) Magnetic recording of the signal, and comparison with the tape recordings of standard letters.

(2) Use of a computer to count and assess numbers of intersections, or lengths of verticals and horizontals.

(3) Searching for energized and non-energized cells in either a physical array of photocells or an electrical analogue of this set up from the scanning waveform within the apparatus.

(4) It has also been proposed to use division of the alphabet of small letters into risers (as "b", "d"), descenders (as "p", "q") and bodies (as "a", "m"), and to subdivide thereafter on a basis of area.

Of these methods, any requiring a computer will be too large and expensive. The use of risers and descenders does not give unique discrimination. The array of cells seems clumsy and over-elaborate for what accuracy it gives. The magnetic recording method appears to be neater and more accurate for simpler characters, but this method is unlikely to be of use for discrimination between the complicated waveforms given on scanning whole words.

To provide reasonably continuous reading without the necessity for spelling out too many words not in the vocabulary, it will be necessary for the vocabulary to contain at least 20 000 words. This arises partly because all tenses of each verb must be provided and nouns must be present in both single and plural numbers. The problem of recognizing each of these words as a separate entity seems at present to be almost insoluble because of the high resolving power which would be required. If, however, the machine recognized reliably all the capital and small letters of the alphabet it might be possible to use a type of switching mechanism to select the corresponding sound-recording by a step-by-step process. This would be a modification of the method used in automatic telephony but the ten numerals would be replaced by the twenty-six letters of the alphabet. The volume of apparatus required would be large, but the type of circuitry used in the electronic type of exchange should help to reduce this.

For the reproduction of words and letters audibly as they are spoken it has been suggested that the "phonemes" of the spoken language might be recorded and stored and subsequently selected in order as required. This method would not give the actual pronunciation which would be provided by a recording of the spoken word. It seems therefore that a store of recordings of the spoken words is inevitable. At least 20 000 will have to be individually accessible within two or three seconds. The problem of access seems to make any type of recording other than magnetic impracticable. In the case of magnetic recording, the problem of the number and size of the sound-heads then arises.

8.4. *Problems for which no practicable solution is at present apparent*

From the foregoing sections it is obvious that there are three main problems requiring solution:

(1) Either the devising of an accurate method of discriminating quickly between a total of 20 000 different electrical signals; or, the design of a suitable switching mechanism to select a particular sound-recording on the successive recognition of the letters of the word.

(2) The storing of 20 000 sound-tracks.

(3) The devising of a method of quick access to each sound-recording in the vocabulary by some means not involving switching.

For none of these problems does there appear to be available a solution of the kind which could be incorporated in a reading-machine which could be placed in a small living-room. It seems that some completely new technique will be required for the solution in each case.

9. Conclusion

The early machines, the Optophone, the Visagraph, and others, could never have appeared in advance of the invention of the Nipkow disc and selenium cell. The considerations of the preceding section indicate that the completion of the ideal reading machine for the blind must await further progress in electronics, and perhaps even in physics.

Nevertheless, the progress which has been made encourages the belief that the complete solution may be less than ten years away. Meanwhile, the interest of one's blind friends provides a sufficient spur to further research.

10. Acknowledgments

Acknowledgment, with thanks, is made to the Royal Society for permission to reproduce the diagram of the Optophone and to H.M. Patents Office for permission to reproduce the two diagrams illustrating Naumburg's Visagraph.

11. References

1. E. E. Fournier d'Albe, "On a direct-reading optophone", *Proc. Roy. Soc.*, A. **90**, p. 373, 1914.
2. R. E. Naumburg, "Scanning and Reproducing Printed Matter in Embossed Form". U.K. Patent No. 383,835 (1932).
3. J. B. Butler-Burke, "Means for controlling the Operation of Type-writing, Type-setting and other Machines". U.K. Patent No. 244,138 (1925).
4. L. E. Flory, "Reading Aid for the Blind". U.S. Patent No. 2,517,102 (1950).
5. H. A. Mauch, "Reading Machine", U.S. Patent No. 2,907,833.
6. V. K. Zworykin, L. E. Flory and W. S. Pike, "Letter reading machine", *Electronics*, **22**, No. 6, pp. 80–6, June 1949.
7. D. H. Shepard, "Apparatus for Reading". U.S. Patent No. 2,663,758 (1951).
8. P. W. Handel, "Statistical Machine". U.S. Patent No. 1,915,993.
9. G. Tauschek, "Reading Machine". U.S. Patent No. 2,026,329.
10. British Tabulating Machine Co., "Optical Means for Reading Printed, Stencilled or like Records". U.K. Patent No. 504,719.
11. C. Jannopoulo, "An Image-detecting Photoelectric Control Device". U.S. Patent No. 2,518,694 (1950).
12. M. H. Glaubermann, "Character recognition for business machines", *Electronics*, **29**, No. 2, pp. 132–6, February 1956.
13. V. K. Zworykin and L. E. Flory, "Reading aid for the blind", *Electronics*, **19**, No. 8, pp. 84–7, August 1946.

Manuscript received by the Institution on 14th February 1962 (Paper No. 744/MBE5).

U.D.C. 621–52 : 681.33

A Transportation Lag Circuit for Analogue Computation

By

R. E. KING, Ph.D., M.Sc.†

Summary: An analogue transportation lag based on the expansion

$$e^{-2s\tau} = \lim_{n\to\infty}\left(\frac{1-s\tau/n}{1+s\tau/n}\right)^n$$

is described. The unit uses n simple phase inverters in cascade, the time delay τ and maximum frequency of operation f_{max} being related by

$$\tau = \frac{n}{6{\cdot}3\,f_{max}} = 2nRC$$

where RC is the phase-shifting time constant.

By means of a simple cross-over unit and high gain d.c. amplifier the delay unit is designed to have extremely low drift. An example is shown which has a transportation lag of 32 ms and a maximum frequency of operation of 150 c/s.

1. Introduction

Situations in which transportation lags occur in physical systems are numerous. In feedback control systems they are often met with to the detriment of stability. They are in general undesirable and no effort is spared to reduce them and their effects if at all possible.

In simulating physical processes on electronic analogue computers it is often therefore necessary to incorporate such transportation lags. In the current research programme into the adaptive control of an analogue of a distillation process[1] it has been necessary to incorporate a transportation lag to simulate a heat exchanger in the plant. Here the unit is characterized by a liquid flow response $L(t)$ which is ideally identical to the input vapour flow $V(t)$ but delayed from it by a time τ, i.e.

$$L(t) = V(t-\tau)$$

In terms of the Laplace operator s, the transfer function can be stated as

$$Z^*(s) = \frac{L(s)}{V(s)} = e^{-s\tau} \qquad \ldots\ldots(1)$$

A number of methods have been described towards this end. Some examples of techniques used in obtaining time delays for analogue computing use

(a) magnetic tape

(b) magnetic drums

(c) capacitors

all these methods relying on mechanical movement,

(d) operational amplifier configurations

for approximate solutions of the ideal transfer function of a time delay. These take the form of expansions of expression (1), a very common one being the Padé approximation on which there exists considerable published work. An extensive bibliography on transportation lags, circuit techniques for their implementation, their application and analysis is given by Weiss.[2]

The method described below falls into the last category and makes use of the expansion‡

$$e^{-2s\tau} = \lim_{n\to\infty}\left\{\left(\frac{1-s\tau/n}{1+s\tau/n}\right)^n\right\} \qquad \ldots\ldots(2)$$

This series does suffer from the fact that it is very slowly convergent and consequently a large number of terms n must be taken for any accuracy. It is here that the more rapid convergence of the Padé approximation makes it more generally attractive for computer simulation. It has, however, been used here because of the relative simplicity in obtaining a transfer function of the type given by expression (2) using simple valve circuitry.

2. Theory of Operation

The requirements for the transportation lag unit simulating the heat exchanger were that it should have unity gain, a time delay of 32 ms, a highest frequency of operation of approximately 150 c/s and be accurate to within 5%. The cost of producing such a delay unit using either of the first three methods described above

† Formerly with the Autonomics Division, National Physical Laboratory; now at Electrical Engineering Department, The Queen's University of Belfast.

‡ Since $\lim_{n\to\infty}\left\{1+\frac{x}{n}\right\}^n = e^x$

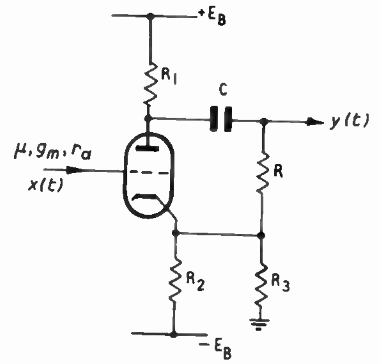


Fig. 1. The basic lag unit

was prohibitive. A unit making use of the Padé relationship necessitates a fourth or fifth order approximation entailing the use of a large number of operational amplifiers which were not available.

It was consequently decided to use the properties of the phase inverter and phase shifting arrangement shown in Fig. 1. This stage gives a good approximation to the non-minimum phase transfer function.

$$C(s) = A_L\left\{\frac{1-sT}{1+sT}\right\} = \frac{y(s)}{x(s)}$$

where $\quad T = RC$ and $A_L = \left|\frac{y(j\omega)}{x(j\omega)}\right|$

The overall transfer function of n such stages in cascade is therefore

$$A_L^n\left(\frac{1-sT}{1+sT}\right)^n$$

which is of the form (2) required. The value of n, i.e. the number of cascaded stages, is determined from considerations of the highest frequency of operation for a given error.

The frequency dependent part of the above transfer function is

$$G(j\omega) = \left(\frac{1-j\omega T}{1+j\omega T}\right)^n$$

where $\quad |G(j\omega)| = 1$

and has a phase shift for any angular frequency ω of

$$\phi(\omega) = -n \arctan\left(\frac{2\omega T}{1-\omega^2T^2}\right) \quad \ldots\ldots(3)$$

which must be compared with that of the ideal transfer function (1)

$$\phi^*(\omega) = -\omega\tau = -2n\omega T$$

since $\quad \tau = 2nT$

A comparison of these phase shifts in Fig. 2 shows a phase error in the unit of 5% at $\omega T = 0{\cdot}57$, resulting in a maximum allowable phase shift of

$$\phi_{max} = 55n \text{ degrees}$$

ϕ_{max} can be expressed in terms of the maximum frequency of operation, f_{max}, as

$$\phi_{max} = 360\, f_{max}\, \tau \text{ degrees}$$

whence $$n = 6{\cdot}3\tau f_{max} \quad \ldots\ldots(4)$$

Thus for a time delay of τ = 32 ms and $f_{max} \simeq 150$ c/s the number of cascaded stages to the nearest even integer is 32. It is of interest here to note the increase to the maximum frequency of operation for each additional stage. From (4)

$$\Delta f = \frac{1}{6{\cdot}3\tau} \simeq 5\text{ c/s}$$

In cascading a large number of stages of the type shown in Fig. 1, a certain amount of drift is inevitable and a unit built up in this manner would not be entirely satisfactory. In order to minimize the adverse effects of drift the n cascaded lag stages are separated into two sections by an amplifier A1 (see Fig. 5). If all stages drift similarly therefore the inversion introduced by A1 will, to some degree counteract the drifts of the two halves. To eliminate the slow drift altogether a simple first-order cross-over unit having time constant T' is added and this is shown schematically in Fig. 3. Here the input signal $V(t)$ is separated into two channels α and β. The β channel transmits only the very low frequency components of $V(s)$ (ideally zero frequency components only for which the

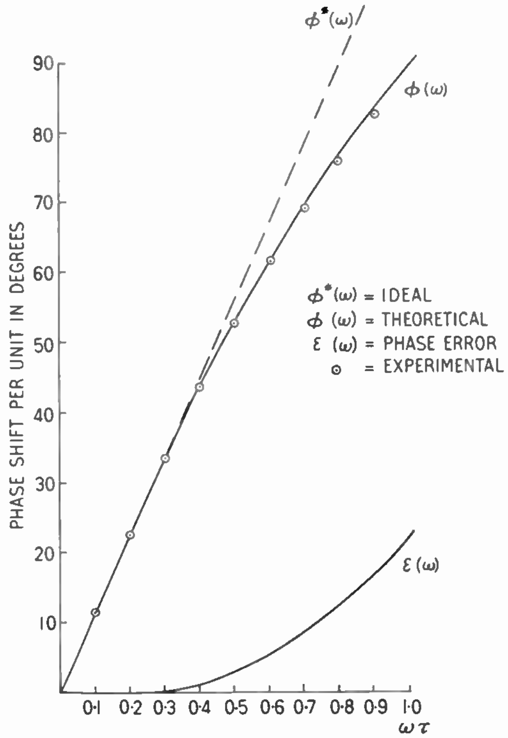


Fig. 2. Frequency phase shift characteristics.

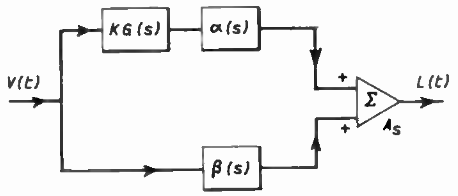


Fig. 3. Schematic diagram of transportation lag unit.

phase shift is zero, (3)) and the α channel the a.c. components only. The reconstituted final output $L(t)$ is the sum of these two signals, i.e.

$$Z(s) = \frac{L(s)}{V(s)} = \beta(s) + K \,.\, \alpha(s).G(s) \qquad \ldots\ldots(5)$$

A necessary requirement here is that the poles of transfer functions $\alpha(s)$ and $\beta(s)$, at $s = -1/T'$ must be sufficiently displaced from the poles of $G(s)$ at $s = -1/T$ for little interaction. This can be ensured by making the cut-off frequencies of $\alpha(s)$ and $\beta(s)$ very low compared with f_0 $(=1/2\pi\tau)$.

In practice networks $\alpha(s)$ and $\beta(s)$ can be incorporated in the final summing unit, an operational amplifier having very low drift. This is shown in Fig. 4 where

$$\alpha(s) = \frac{L_1(s)}{U_1(s)} = \frac{mT's}{1+T's}$$

$$\beta(s) = \frac{L_2(s)}{U_2(s)} = \frac{1}{1+T's}$$

$$T' = R'C'$$

Network $\alpha(s)$ having a high-pass characteristic effectively blocks the very low frequency drift components of the lag unit from appearing at the final output.

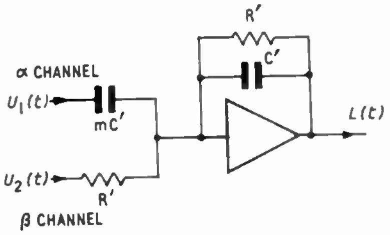


Fig. 4. The summing amplifier, $\alpha(s)$ and $\beta(s)$ functions.

The method of obtaining $\alpha(s)$ by coupling the n cascaded amplifier stages and amplifiers forming $KG(s)$ to the final summing amplifier by means of capacitor mC' may normally lead to the phenomenon of "paralysis" where the first stage of the summing amplifier may be momentarily cut off if the rate of change of the input signal $U_1(t)$ is too great. For this application, however, such high rates cannot in fact occur due to the "effective" low-pass filtering nature of $G(s)$.

Noise generated in the lag unit $G(s)$ is reduced at the output by making the $\alpha(s)$ transfer function eventually fall off at high frequencies. This is simply done by adding a small resistance r in series with mC', resulting in

$$\alpha'(s) = \frac{mT's}{(1+T's)(1+T''s)}$$

where $T'' = rmC'$, and $T'' \ll T'$ so that the phase error introduced by this high-frequency fall-off is of no consequence in the operating range.

3. The Practical Circuit

The complete circuit of the transportation lag unit is shown in Fig. 5.

Each lag stage has equal anode and cathode resistances as far as the a.c. signal is concerned. In terms of the standing (or d.c.) signal the cathode resistance must be approximately twice the anode resistance R_1 since the current flowing through both must be equal, i.e.

$$R_1 = \frac{R_2R_3}{R_2+R_3} = 100\,\text{k}\Omega$$

for a standing current of 1·5 mA. Thus

$R_2 = R_3 = 200\,\text{k}\Omega$ or $R_2 = 180\,\text{k}\Omega$ and $R_3 = 220\,\text{k}$

The gain per stage for small signal operation can easily be shown to be

$$A_L = \frac{\mu}{2+\mu+\dfrac{r_a}{R_1}}$$

With a mean anode current of 1·5 mA the valve parameters for an ECC81 are $\mu = 50$ and $r_a = 30$ kΩ. Hence with $R_2 = 2R_1 = 200\text{k}\Omega$, $A_L \simeq 0{\cdot}97$.

As the unit was required to accept a peak input of 100 V it was necessary to attenuate the input signal $V(t)$ prior to feeding it to the first lag stage to ensure small signal operation. The grid swing was limited to 10 V peak to peak by a 20 : 1 input attenuator.

Under the stated operating conditions the cathodes sit at approximately 2 V positive with respect to their grids. To reduce this cumulative effect cathode follower buffer stages are introduced after every eight cascaded stages. A total of four such cathode followers is therefore necessary.

The overall gain of 32 cascaded stages and input attenuator is approximately 1/80. To increase this gain to a level more suitable for coupling to the final summing amplifier and in order not to use a large

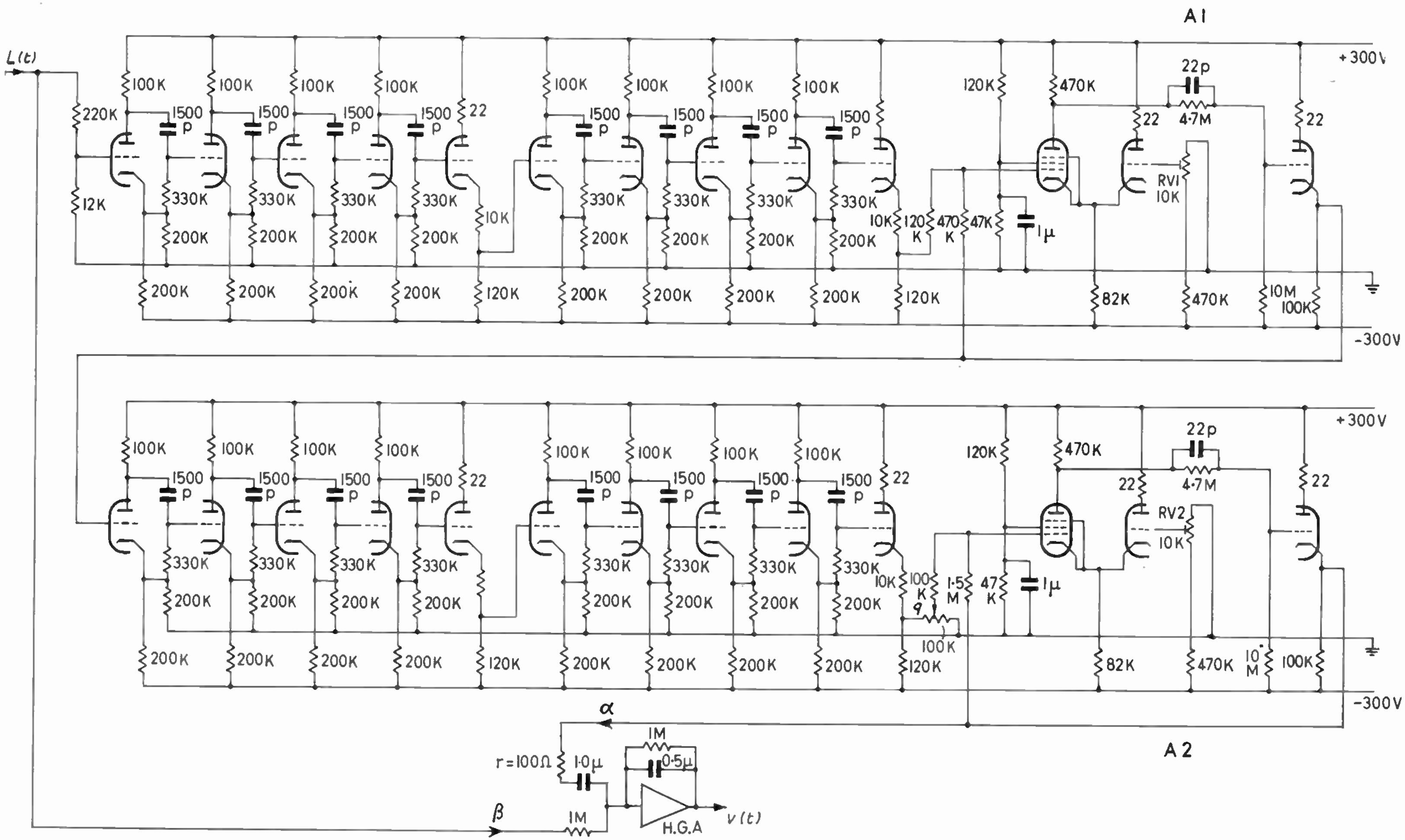


Fig. 5. The complete circuit of the analogue lag circuit.

value of mC', two low gain (open loop gain approx. 200) amplifiers A1 and A2 are included. These amplifiers do not require drift correction and two are necessary in order to get the correct signal sense at the input to the final summing amplifier As. The closed loop gain of A1 is 4 (this amplifier also acting as a partial drift corrector) and that of A2 is 15. Thus the overall gain of the α channel is

$$K = \frac{1}{20} A_{cf}^{4} A_{l}^{32} A_1 A_2 q \qquad \ldots\ldots(6)$$

where A_{cf} ($\simeq 0{\cdot}9$) is the gain of the cathode followers and q the output potentiometer (RV_0) gain-setting.

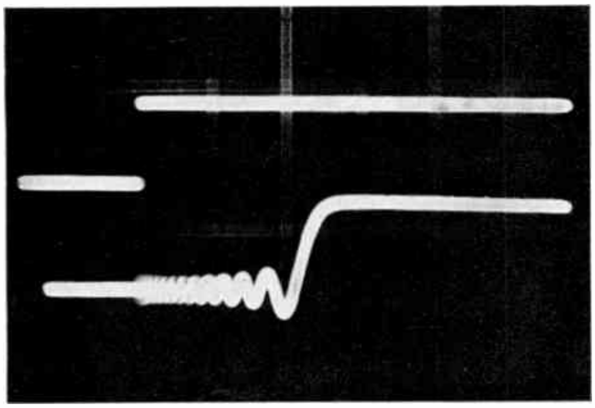


Fig. 6. Transient response.

In setting up the complete unit the value of K is made equal to $1/m$ so that the effective gains to both the α and β channels are equal and unity. This is done by feeding a known input to the unit and adjusting q to satisfy the condition $K = 1/m$. A value of $m = 2$ and a time constant $T' = 0{\cdot}5$ second were taken, this being some 100 times greater than the time constant of the individual lag stages. Potentiometers RV1 and RV2 are pre-set to give approximately zero output for d.c. zero input at the output of A2.

Using selected high stability preferred values of $R = 330$ kΩ and $C = 1500$ pF the resulting transportation lag is 32 ms and the exact maximum frequency for a 5% error in phase is $f_{max} = 160$ c/s.

The step response of the unit is shown in Fig. 6.

Analytically the step-response can be derived using the Residue theorem which gives it as

$$V(t) = \frac{L\delta_{-1}(t)}{(n-1)!}\frac{d^{n-1}}{ds}\left[(1 - sT)^n \frac{e^{st}}{s}\right]_{s=-\frac{1}{T}}$$

where $L\delta_{-1}(t)$ is the input step. For high orders of n the solution is obviously very cumbersome.

4. Acknowledgments

The author wishes to acknowledge the help and advice of Dr. M. J. Somerville[4] of the Electrical Engineering Department, Manchester University, for the basis of this paper.

The work described above has been carried out as part of the research programme of the National Physical Laboratory, and this paper is published by permission of the Director of the Laboratory.

5. References

1. R. E. King and N. W. Rees, "Optimization of an electronic analogue of a simple distillation process". (To be published.)
2. R. Weiss, "Transportation lag—an annotated bibliography", *Trans. Inst. Radio Engrs (Automatic Control)*, AC-4, No. 1, pp. 56–64, May 1959.
3. G. A. Korn and T. M. Korn, "Electronic Analog Computers". (McGraw Hill, New York, 1952.)
4. M. J. Somerville, Ph.D. Thesis, Manchester University, 1958.

Manuscript received by the Institution on 7th December 1961 (Contribution No. 53/C41).

SPECIAL GENERAL MEETING

Minutes of Special General Meeting of Corporate Members held on Wednesday, 23rd May 1962, at the London School of Hygiene and Tropical Medicine, Gower Street, London, W.C.1, commencing at 6 p.m. At the commencement of the meeting 66 Corporate Members had signed the attendance book.

The Chairman, Mr. W. E. Miller (Vice-President) asked the Secretary to read the notice convening this first Special General Meeting of the Chartered Institution, which was circulated to Corporate Members on 26th April 1962.

Mr. Miller then reminded members that under Article 18 of the Institution's Charter the Corporate Members were required to make the Bye-laws under which the Institution would in future be governed. The draft Bye-laws which had been compiled by a specially appointed Committee, in consultation with the Petitioners, the Institution's Counsel and Solicitors, had been circulated to all Corporate Members.

Subject to the Corporate Members approving these Bye-laws, they would then be submitted for final approval by the Lords of Her Majesty's Privy Council. Mr. Miller continued:

"I have here the printed document which is laid before this meeting and which I have identified as being precisely in the same terms as the document which accompanied the letter dated 26th April last convening this meeting.

"We have been greatly helped by observations made by various Corporate Members. I hold 212 proxies vested in myself or in other Officers of the Institution who are present in support of the two Resolutions. Any questions on the second Resolution will be dealt with later, but before I move the First Resolution I would like to receive any comments from the members present."

After replying to questions raised by three Corporate Members, the Chairman formally moved the adoption of the first Resolution, namely:

"That the bye-laws contained in the printed document laid before this meeting and subscribed for identification by the Chairman be and the same are hereby approved and made the Bye-laws of the Institution."

The Resolution was passed without any amendment.

Referring to the second Resolution, Mr. Miller stated:

"We now come to the second Resolution mentioned in the notice of meeting and in explanation I should point out that it is possible that the Lords of Her Majesty's Privy Council may require some minor alteration or amendment to the Bye-laws we have made. If there is any radical alteration proposed by Their Lordships then the Council will call another Special General Meeting.

"The Resolution does, however, empower the Council to act on your behalf and I formally ask if you will give the Council this authority, it being clearly understood that any amendments accepted will be duly circulated to all members of the Institution."

Mr. Miller then formally moved the second Resolution:

"That the Council of the Institution be and is hereby authorized to exercise its discretion to approve such amendment, alteration or addition to the Bye-laws of the Institution as may be required by the Lords of Her Majesty's Privy Council." The Resolution was passed unanimously.

After thanking Corporate Members for their attendance and for their confidence in the Council, Executive Committee and legal advisers, the Chairman declared the Special General Meeting at an end.

For the information of Corporate Members

The Lords of the Privy Council have made observations on Bye-laws 48, 52, 60 and 79. In pursuance of the authority given to the Council at the Special General Meeting, the draft Bye-laws have been altered or amended as follows:

Bye-law 48: After clause (*d*) add "(*e*) if at a General Meeting of the Institution it be resolved by a majority of not less than three-fourths of those present in person or by proxy and voting that he cease to be a member of Council. Provided that such member shall have had not less than twenty-one days notice in writing of the intention to move such a resolution and an opportunity to reply to it either personally or in writing".

Bye-law 52: A further sentence has been added: "No business shall be transacted at any meeting of the Council unless a quorum of six members of the Council is present at the time when the meeting proceeds to business."

Bye-law 60: A final paragraph has been added after Clause (*c*): "A copy of the Balance Sheet and Auditors' Report that is to be presented to an Annual General Meeting shall be sent at least twenty-one days before the meeting to all members who are entitled to receive notice of the meeting."

Bye-law 79: The last sentence now reads: "If a quorum be not present within a quarter of an hour from the time appointed for holding the meeting the meeting shall be adjourned to such date and time as the Council shall decide."

Corporate Members may wish to amend accordingly the Bye-laws sent to them with notice of the meeting held on 23rd May 1962. Advice will be sent to all Corporate Members when the Bye-laws, including the above amendments, are allowed by Her Majesty's Privy Council.

U.D.C. 681.32: 621-52

The Application of the Hierarchy System to On-Line Process Control

By

J. F. ROTH, B.Sc.†

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

Summary: With the acceptance of a computer for the control of separate on-line processes, consideration must now be given to extending such a control system to complete plants. The use of a single composite computer unit for this purpose is not generally practical, and an alternative approach to the problem is discussed. This uses a number of self-contained control units which are interconnected to form a hierarchy system. The pattern of control so formed is analogous to a conventional management structure.

To illustrate the principles involved and demonstrate their application an integrated control and data handling system for a steel works is described.

1. Introduction

The use of computer systems to assist in the solution of the complex problems associated with the operation of a process plant is now an established technique. The many successful installations have demonstrated that the computer can provide a practical and economic system for process control. It would be valuable to have adequate information about these installations in order to fully assess them, but unfortunately due to various reasons, mainly commercial secrecy, very little is available. However, from the few systems which have been described[1–5] it is evident that a major step forward in plant control and operation has been achieved. The secrecy maintained by so many operators supports this impression, for management are obviously reluctant to tell their competitors how they have improved their production methods. Of course, similar reasoning can be applied had the system proved unsatisfactory, but enough is known to indicate that these form a very small minority. In these cases failure can generally be attributed to insufficient initial system analysis and not to the performance of the installation.

Of the computer controlled installations at present functioning[6] almost all are concerned with a single process or operation. As the complexity of the process increases, the size of the computer control system required can become extremely large, with consequent problems for maintaining an adequate availability figure. This figure provides an indication of the reliability of the system, and is defined as the ratio of the useful time to the total time during which operation is required. This latter period is equal to the useful time plus the overall down-time due to a fault.

† Elliott Bros. (London) Ltd., Process Computing Division, Borehamwood, Hertfordshire.

The magnitude of the plant and its control problems dictate the size of the computing system and thus the number of components used. The reliability of such devices as resistors, capacitors, transistors and soldered connections in this type of application has been investigated[7] and although the expected life of single components is high, the life of a complete equipment is inversely proportional to the number of components involved. In process control work an availability figure of 98–99% is normally acceptable, and quite evidently if the size of the computer system is increased, a point will be reached where this figure can no longer be maintained.

2. Integrated Plant Control Requirements

Having successfully shown that computer control can be applied to a single process, industry is now investigating similar systems which can be used for the integrated control of a complete plant. Before this task can be tackled a successful solution of the problem of reliability is required. To build a single large system which is capable of dealing with the whole plant would obviously be the wrong solution, as its sheer size would not give an adequate availability figure. Coupled with this is perhaps the even greater practical problem of how the plant could be organized so that it would be able to carry on safely when the control system fails. Assuming that this could be overcome, some means must then be provided for telling the control system what has happened during the breakdown period so that it is able to "catch up" with the actual process. One further point, which is perhaps the most important, is that in order to accept all the information regarding plant conditions, perform the required calculations and produce necessary control information, an extremely fast computer would be required. Even some of the modern ultra-fast computers may have difficulty in dealing with such a vast number of operations.

3. The Hierarchy System

3.1. *General Arrangement*

As so often happens with complex problems, the solution is frequently much simpler than it at first appeared. If any control problem is examined it will follow a general pattern in which a graded organization or hierarchy can be seen. The lowest level will be the most widely based, and will produce information to adjust plant controls, and deal with the acquisition of data from transducers which measure various plant parameters. This level can be divided up into a number of sections each dealing with specific aspects of the overall plant. Each section will thus accept a number of inputs, perform simple calculations, or organize the data and present the results in a form to be used as inputs by the next higher level. The second level will accept input data from a number of lower level sections, and will operate in a similar manner on this data. In its turn it will supply information to the next higher level. The number of levels and the sections in each level will depend on the magnitude of the control problem, and in general terms the higher the level the more complex are the calculations that have to be performed. In this manner the essential information regarding the plant operation and performance is passed up through the various levels, and each level receives only that information which it requires to perform its specified task.

To illustrate the principles involved, a section of a refinery will be considered. In a crude-oil distillation column, one of the lowest level sections would—amongst many other tasks—calculate from measurements of the temperature, and pressure drop across an orifice in the feed pipe the instantaneous rate of flow of crude oil, corrected to n.t.p., passing into the pipe-still. The next level would use this value to calculate the mass of oil passing into the pipe-still during a given period of time. For this purpose the elapsed time and the density of the crude oil is required. The latter would normally be obtained from laboratory experiments, and entered from a manual keyboard either directly at this level or possibly at some other, if more convenient, and then transferred to this level. It would of course be possible theoretically to monitor the density continuously and use this as a further input. The raw material input so calculated is required by the next higher level for the calculation of the material balance which in its turn is necessary for the determination of the overall economy of the plant.

So far only upward movement of information has been considered, but in a similar manner control instructions will be transferred down through the levels. Thus, an order from a customer would be entered at one of the higher levels, and the necessary instructions would be generated to modify the plant performance to meet the new demand.

3.2. *Role of the Operator*

At each section of every level facilities are provided for the intervention of the human operator. This is essential in order to cope with fault conditions or situations which are unexpected and require verification. Alternatively the experience of the operator may be necessary in order to solve a particular problem about which insufficient is known to allow the computer system to be programmed to make the necessary decisions. For these reasons the operator can never be dispensed with, although it may well be possible to increase the number of sections of the plant under his control.

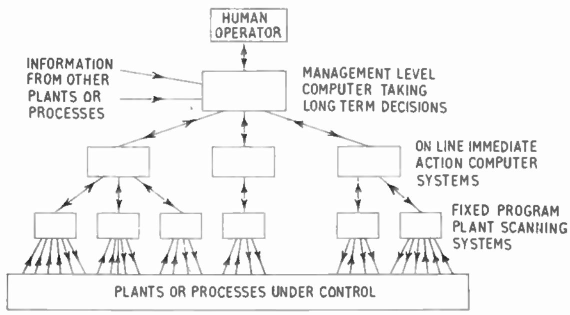


Fig. 1. Management system of control.

A diagrammatic representation of the Hierarchy System is shown in Fig. 1. Supervising the whole installation is of course the human operator, for at this uppermost level there will always be many functions which cannot be left to a machine. The form of this diagram is very similar to that of a management structure and a comparison of the two will show that the mode of operation is also similar. In both systems information is fed up through the various levels and operating instructions, which can be initiated at any of the supervisory levels, come downwards. At each level a man is responsible for his section, but in the management structure he must also approve and initiate all message transfers thus involving him in a large amount of routine work. However, in the hierarchy system being described, the message transfers take place automatically and the man responsible for the section is only involved when abnormal conditions occur or when his attention is definitely required. This form of control is termed "management by exception" for it relieves the operator of routine functions thus allowing him more time to deal with those problems which require his experience and intelligence for their correct solution.

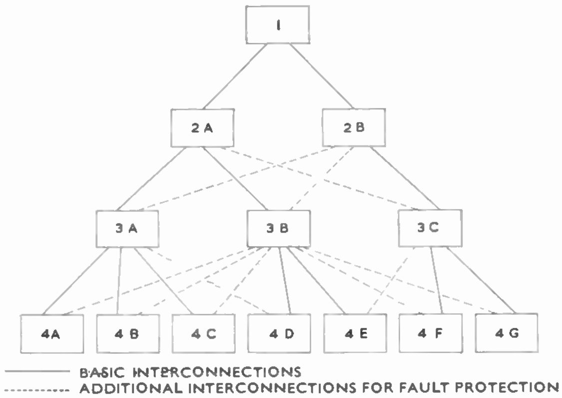


Fig. 2. Inter-section message paths.

3.3. *Safety*

As previously described, the Hierarchy System is based on the division of the overall control system into a number of self-contained sections which are interconnected. Thus if a failure occurs in a given section it will only be that one which will be affected, the rest of the control system carrying on normally. However, this is not quite true as the failed section will have some effect on those sections immediately adjacent to it. This can be covered by arranging for the tasks of the failed section to be performed by other sections, using the various interconnecting paths which are available. To enable this to be achieved additional interconnections to those which are basically necessary will be required. In addition the sections must keep a watching brief over their neighbours in order to take over their functions in the event of failure, and be able to signal to other sections what has happened. The arrangement is illustrated in Fig. 2.

The lowest level sections are operating in relation to real-time and are thus "on-line". Should one of these sections fail essential plant conditions may be lost, yet due to the physical equipment needed to monitor plant operation, there are practical difficulties in providing alternative arrangements. For this reason it is necessary to keep these on-line sections as small as possible so that a failure will cause the least possible disturbance. The availability figure for these sections will thus be high due to their small size, and as the other sections are provided with failure protection, the overall availability figure for the complete integrated control system will, to a first approximation, be equal to that for the largest on-line section. In this way the availability figure will be far superior to that which would be obtained with an equivalent single unit system.

With the higher level sections, a break in operation may have little visible effect on the complete system, assuming, of course, that the failure is of a reasonably limited duration. In such cases it may well be sufficient to supply adequate buffer storage facilities with each section, so that the messages which should be transferred are temporarily held until they can be dealt with.

The adjustment of the actual plant controls is best achieved by maintaining conventional controllers for the purpose, and arranging for the relevant section of the system to change remotely the position of the set-point. In this way fail-safe operation is ensured, for if the section generating the change instructions should fail, the set-point will remain at its last position. It is then easy for the operator to take-over, and move the set-point manually should he so desire. With this method of operation the controller forms the lowest level of the Hierarchy System.

3.4. *Message Transfers*

The Hierarchy System requires the facility for transferring messages between the various sections. Two problems are involved in this, (a) that the message is correctly transferred over the link, which may be of considerable distance, and (b) that two sections can be coupled for the purpose of transferring a message without interfering with the operation of the sections.

3.4.1. Transmission link

Having divided the control system into a number of sections, these can be separately placed around the plant in their most convenient positions. Thus the "on-line" sections would be best placed on the "production floor", whereas the upper supervisory level would be better positioned in the administrative building. However, the distances between the various sections may be considerable, even as far as many miles and thus a method of accurately transmitting data is required. This problem is similar to those already solved for standard communication channels, except that the interference may well be much greater. With the equipment found in a modern industrial plant (such as mercury-arc rectifiers or high power reversing motors) the noise introduced into signal lines can be very troublesome. To reduce this to a minimum, lines should have a low impedance and be correctly terminated. To prevent the alteration of the message or part of it as a result of the noise, error detecting codes[8] should be used, so that if an error occurs, an automatic request for its repeat can be made. Alternatively, even though a greater band-width is required, it may be considered worth while to use an error correcting code should the noise level be so high as to introduce frequent errors.

3.4.2. Interlinking of the sections

As each section of the control system is self-contained and operates independently from its neighbours, the organization of the message transfers requires some care. If one section has prepared a message for transmission to another, it can indicate this to the receiving section. However, this section may be currently performing a task which cannot be interrupted. In this case the transmitter section may wait until the receiving section is free or alternatively perform some other routine and try again later. It is thus possible that either one section may waste time in waiting for the other to become available, or a fair period may elapse before both sections are capable of being interlinked for the transfer to take place. An improved method of interconnecting two such asynchronous systems would be to use a buffer store in each link, and allow both sections the facility of detecting the conditions of the store. The transmitter section would in this case first test to see if the buffer store is clear. When it finds that it is, the message will be transferred to the buffer, and this section is then free to perform some other task. The receiver section will be arranged to test the buffer store at convenient intervals to see if a message is waiting. If it finds that this is the case, it will then read the store and clear it. In this way each section spends a minimum of time in organizing the transfer.

The quantity of information involved in a single transfer is largely a matter of economics. With single character transfers the buffer store is as simple as it can be, but a complete message now involves each section in a number of operations equal to the number of characters in the message. However, in general terms the characters of a message are separately formed, and on receipt are dealt with one at a time. On the other hand, with a complete message transfer, each section is only involved in one operation, but the buffer store is much more complex as it must be able to sequentially store the separate characters, and then later output them in the correct order.

Generally a single character transfer is satisfactory, with one link for each direction. This arrangement can provide the facility for an easy method of manually modifying or checking the information being transferred. This is a valuable aid for the operator especially under fault conditions or when the system is being tested.

3.5. *Flexibility of the Hierarchy System*

There are two aspects of the Hierarchy System which are most important to its general philosophy, and have a profound effect on the approach to its utilization. These have been considered previously and are:

(a) A buffer store in each message transmission link between the sections.

(b) Each section is self-contained and operates as an independent unit.

As a result of these features it is quite practical to install a single section of any level in the hierarchy. The buffer stores in the communication links provide the method for entering information and obtaining results from the section in a manner similar to the way in which it will be used when finally integrated into the control system. To begin with, the new section would be used in parallel with the part of the existing system which it will finally replace or supplement. In this way its operation can be checked and modifications made until the required performance is achieved.

Further sections can be added without affecting those previously installed. Again when two or more sections are interconnected, the buffer stores in the communication links provide the facility of monitoring the performance and ensuring that the sections are working properly. Constraints can be applied at the buffer store to prevent information passing between the sections moving outside limits set by the operator. Alternatively an alarm may be given so that the operator is warned of the onset of an unexpected condition. As has been found, an automatic control system can often operate a plant stably and successfully in a region where this could not be achieved manually, and thus the control system may attempt to adjust the operating conditions towards a point which would seem wrong to the operator.

By the use of self-contained sections, both digital and analogue techniques can be easily combined. Thus the technique to be employed with any given section can be selected as the most suitable for the particular application without having to be modified as a result of the requirements of the remainder of the system. The only limitation to the interconnection of different types of sections is that they must have a common format and code for the transfer of messages. A similar limitation occurs to the interconnection of equipment made by different manufacturers, which could otherwise be incorporated into a hierarchy system. There is a definite need for such a common language, and the British Standards Institution is at present dealing with this problem.

It has often been stated that before a computer control system can be applied to a process or part of it, an accurate mathematical model is necessary. However, the normal plant operator who at present manages to control the process quite successfully knows little about the equations governing the process. He works from a series of mainly empirical rules and experience which have been developed over a period of time to produce the required operating conditions. A new operator is trained essentially by watching an experienced man and then trying himself, in this way developing a "feel" for the plant. Each action taken by the operator can therefore be reduced to a series of logical deductions. By analysis of what plant operators do, a computer program can be constructed which will behave in a similar manner as the operator. In the same way as a new operator will find that some action of his produces the wrong result and he thus learns to modify his procedure, so a similar method can be applied to the computer program which can be modified accordingly. Different plant operators obtain varying results from the process under their control, and therefore using this method of developing the computer program it will be possible consistently to reproduce the performance of the best operator. In this manner a mathematical model of the process is not required, although in all probability the optimum plant operating conditions will not be achieved. The improvement that results may however be adequate to justify the control system, and the search for even better results by constructing the mathematical model can then be carried out in parallel using the same equipment by means of alternative programming techniques.[9]

Because the sections are joined by communication links, the distance covered by the link can be many miles, and it is quite practical for example to use a telephone line with digital transmission facilities for this purpose. In this way the control system can embrace a number of widely scattered installations and thus a group of establishments may be centrally controlled. The prospects which are in this way laid open for the future reveal untold horizons—yet they can be achieved in easy stages through the gradual planned introduction of the single sections which will finally form a complex hierarchy system of control.

4. Description of a Hierarchy System

The application of the Hierarchy System to the solution of complex on-line process control problems has been discussed in general terms in the preceding sections. To emphasize the principles involved and indicate more fully the scope of the system, a part of an integrated control and data handling Hierarchy System for a steel works will be described. An exhaustive description will not be attempted, but those aspects which are relevant to an understanding of the system will be considered in more detail.

4.1. *Application*

Figure 3 shows the section of the steel works which is covered by the Hierarchy Control System. A brief description of the process for producing strip steel will be given so that the problem can be readily understood.

The production of strip steel starts from the point at which the molten steel is available from the converters. From the customers' orders for steel strip the quality of the steel to be produced is known. The converters are therefore charged in such a manner so as to produce this grade of steel. The molten steel is poured into moulds—technically referred to as "teeming"—the type of mould depending on the grade of steel.

Between 5 and 10 moulds are assembled on a train at least two hours before teeming is to take place. A short period before this occurs, the steel is tested to see if it is of the required grade. Should it not be to specification a different set of moulds may be required, and for this purpose, a spare train is held in readiness.

The teeming operation takes about 20 minutes and after a further two hours the steel has cooled sufficiently for the moulds to be removed, the resultant "block" of steel being called an ingot. The weight of an ingot is between 5 and 20 tons and an average size is 7 ft × 2 ft × 5 ft. This operation—called "stripping"—is not always straightforward, and sometimes the mould cannot be released—a "sticker" —or it may fall over before the ingot can be removed. In either of these cases the ingot cannot be put into the "soaking pit", where it is brought up to the required uniform temperature for the next operation. The soaking period is about 4 hours. In the stripping bay there is a stock of cold ingots as a result, for example, of a later release of a sticker, the spare train having been used, or more likely due to the successive sections of the process having been closed down for the daily routine maintenance. Whenever possible this stock must be used, but a cold ingot may take nearly 16 hours to be brought up to the correct temperature after being placed in the soaking pit.

When the ingots are at the required temperature they are withdrawn from the soaking pit, weighed, and passed to the slabbing mill. Here the ingot is reduced to about 6 in. thick, 50 in. wide and 25 ft. long. The slab is then inspected for imperfections and, as the leading and trailing ends have an irregular shape, these are cropped off on the slab shear.

The slab is then transferred to the slab stock yard. Slabs are sometimes sent to other rolling mills for further processing and similarly slabs from other steel works are sent to this one for rolling into strip.

When required for the strip mill a slab is transferred to the reheat furnace about 2 hours in advance of it being rolled so that it can be brought up to the required temperature. When it is ready, it passes through the roughing mill (where only one set of reducing rollers act on the slab at a time) and is reduced to about 1 to 1½ in. thick by 50 in. wide and 150 ft long before passing to the finishing mill. This consists of six sets of reducing rollers (six-stand mill) closely spaced and on leaving the final stand the strip is moving at about 30–40 miles/hour and is now between 0·064 in. and 0·375 in. thick by 50 in. wide and up to 3000 ft long. The strip can then be coiled,

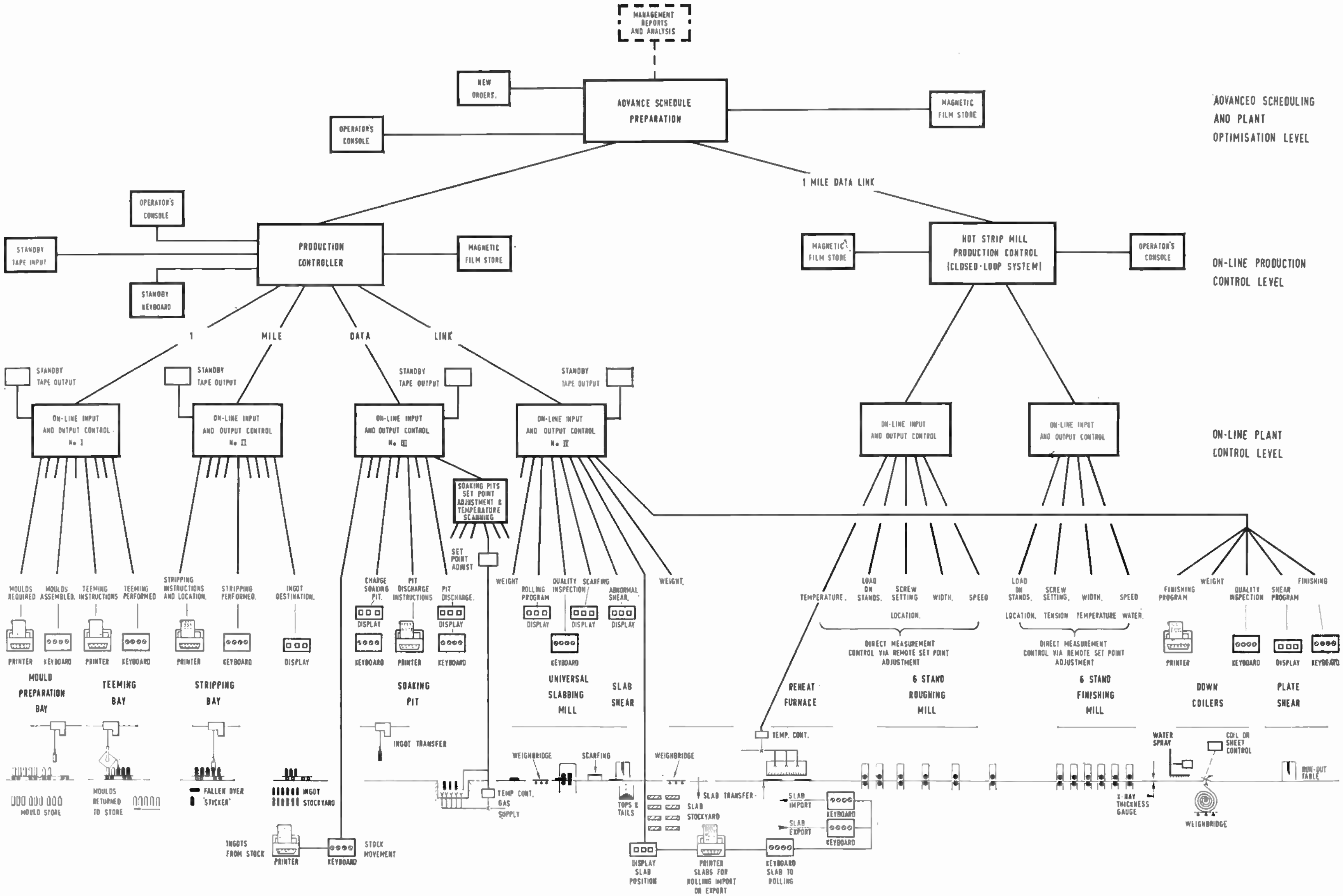


Fig. 3. Hierarchy system for steel strip mill.

weighed and finally inspected for quality or alternatively cut up into the required lengths.

The time scale from start to finish of the steel strip production process will vary between quite wide limits. Although the minimum period could be of the order of 36 hours the average is about 2½ days.

The various stages of the process are physically laid out to follow each other, and the distance from the teeming bay to the down coiler is about 1000 yards.

4.2. *Hierarchy Levels*

Three levels of the hierarchy can be readily seen in this system.

4.2.1. Upper level

The top machine prepares the advanced production schedules from the new information. In generating the necessary instructions it must take into account the items held in stock at the relevant points in the process. A schedule of operations is thus prepared for every stage in the process, and these are passed down to the next level so that the latter always stores sufficient instructions for the following eight hours. In this way a breakdown in the top machine is not catastrophic—as long as it does not exceed this period. This is highly unlikely, for the average time a computing system of this type is out of action due to a fault is under 3 hours. The magnetic film store also provides protection in the event of failure, as all information is held on the film and this provides a permanent storage medium.

Information about what has actually happened in the plant as a result of trying to operate the schedules is fed back to the upper machine. This may necessitate an alteration to the schedules not yet acted upon, and a modification can be prepared and sent down to the lower levels. For example, if the grade of steel produced was incorrect and the spare train used, the ingots may be kept at the ingot stock yard. Information now has to be generated to specify where the ingots are to be placed in the stockyard and the schedules for the succeeding stages modified to take into account the resulting gap. Also directions for dealing with the unused train of moulds must be produced, and instructions given to assemble a further spare train. In this way the slabbing mill, for instance, will be unaware that anything unusual has happened at an earlier stage of the process.

The upper level section is storing information concerning not only what appeared to be the ideal way of operating the plant but also what has actually happened. This information can be analysed and made available in a suitable form for management reports. This is precisely what management requires, for they can now be presented with essential up-to-date information and make more rapid decisions without getting submerged by a mass of mainly irrelevant documents and obsolete facts.

4.2.2. Middle and lower levels

There are two middle sections each with a number of lower level sections in this Hierarchy System. One part deals exclusively with the rolling mill and the other covers the remainder of the process.

The upper level sends to the rolling mill control section a list of the slabs which are to be rolled, and the final thickness for the resultant strip. From this information the rolling mill is operated as an independent closed-loop system. The rolling temperature and degree of reduction at each mill stand is calculated and checked to ensure that they fall within the specified stand loading limits. The required initial slab temperature is achieved by direct control of the set point of the reheat furnace temperature controller.

By adjusting the set point of a conventional controller, fail safe operation is achieved. In the event of a failure of the control system the temperature will remain where last set. The operator is also able to adjust the temperature directly in the conventional manner should he subsequently require a change. Similarly during the commissioning period, the operator will be able to see what temperature has been selected by the control system so that he may verify the operation. Limits can be applied to the adjustment of the set point so that if it does attempt to move outside them, the attention of the operator is drawn to the situation. As can be seen the set point adjuster forms the buffer store referred to in Section 3.4.2 as a requirement for the successful interconnection of two sections.

The remainder of the control system acts mainly as an information acquirer and instruction generation unit rather than a closed loop automatic control system like the strip mill. At each processing stage printers or visual displays indicate to the operator what he is expected to do and from a keyboard he indicates his actions to the system.

This part of the process is divided into four areas. Each area is controlled by a fixed program unit. Each fixed program unit has associated with it up to 32 different types of input message regarding actions in the plant. The messages are set on decade switches and can have a maximum length of 24 characters. Up to six types of input messages are grouped on a keyboard, the character switches being shared. A typical example of a message would be when a slab has been placed in the required position in the stockyard, the operator would key in the cast number, ingot number, and position placed in the stockyard. The cast number and ingot number uniquely describe an ingot and is thus used as its recognition number throughout every stage of the process. Messages can also be entered automatically, as for example, when a slab passes over a weigh-bridge.

Two examples will be taken to illustrate the use of the input and output facilities. At the mould preparation bay, a list of the moulds which are required to make up a given train is printed out two hours in advance. As the train is being assembled, the keyboard is used to inform the system which moulds have actually been placed on the train and their positions. This latter item is required as it does happen that during teeming a mould may not be filled.

In the stockyard a display will inform the operators where an item is to be placed just before it enters the yard. When the item has been put in its position the details are entered into the system via a keyboard. Provision is made so that if the area in the stockyard is occupied, the system can be informed of this apparent error. This may for instance be due to someone placing an item in the wrong position. Alternatively, when items are to be withdrawn from the yard, a list is produced on a printer giving the details of the item and where each is to be found. As each item is taken away its details are entered at a keyboard. Instructions may be given for an item to be taken from a certain place, and it is not actually there. In this case the keyboard will be used to indicate an erroneous instruction.

At each keyboard a lamp is provided which is used to indicate when the system does not agree with the information that the operator has entered. For instance, he may have made a mistake in setting his switches, or alternatively the computing system may be in error, most likely due to an operator fault at an earlier stage. When the operator sees that the query lamp is illuminated following his entry of information, he will check his switches and re-enter the item. Should the first entry have been correct the computing system will receive a duplicate message, and interpret this as an error in the stored information. If it cannot determine the source of error and make its own correction, an alarm will be indicated in the central control room.

As an operator can enter his information at any time a method must be used for organizing the messages. This is achieved by having a scanner in each fixed program unit which looks at each message initiator in turn. When it comes to a message that is ready, the scanner pauses to read the message and when this has been organized continues until a further message is found. In this way only one message has to be dealt with at a time, and the speed of scanning is selected so that no message is kept waiting for more than a few seconds before it is read. It is considered that an operator will be prepared to wait a maximum of 10 seconds for his message to be dealt with. If he should be kept waiting for a longer period he is very liable to assume that the equipment has failed.

Having read the message, and added a unique code to it to show the time and point of origin, the message is transferred to the middle level computing system. To maintain this information in the event of the link or the computer system failing, arrangements are made so that should this happen the messages are recorded on punched paper tape. When the fault condition is over, the tape is entered into the computer system which can now catch up. As with the upper level, this level also has a film store on which all data is permanently recorded.

Each of the fixed program units have a number of output facilities associated with them. Advanced information to cover a few hours' operation is printed out on a series of teleprinters placed in strategic positions. Actual instructions for the successive operations are given on visual decade displays or single lamp displays situated at the point where action is to be taken. In the event of a failure, the printed advance information copy can be used. Should the failure be prolonged some or all of the teleprinters can be disconnected from the automatic control system and interconnected with a further keyboard in the control room to form a conventional Telex type of communication network. In this way a skeleton control system can be maintained.

In the control room a stand-by console is provided which can take over the functions of any of the message input stations. Should any of these fail in the field, a telephone system is provided so that the stand-by console can be used to enter the relevant messages.

Other output information would be to adjust the set points of conventional controllers. For instance, the soaking pit can be set to maintain the required temperature which will vary according to the type of ingots charged in it. The temperature of each pit and the set point temperature are regularly scanned, and each measurement is used as an automatic input for one of the fixed program units. It is possible to consider this scanner as a lower level in the Hierarchy System than the fixed program units, and the conventional controllers then become an even lower level.

As has been previously stated each section (i.e. fixed program unit or computer system) has facilities provided to draw the attention of the operator to abnormal conditions and permit him to make modifications. The higher the level, the more comprehensive are these facilities.

From the descriptions given it will be readily appreciated that any section of the complete Hierarchy System could have been installed by itself, as each is self-contained and performs a useful function.

4.3. *Future Extensions*

It is of interest to consider briefly a couple of the possible extensions in the Hierarchy System that has been described.

With this plant, only a portion, which is in fact in the centre of the complete process, has been covered by the control system. The steel manufacturing portion which is the first stage, and the Cold Reduction and Finishing Mills which is the last, can each be provided with a similar control system. A further computing system can then be provided to link together those three systems thus giving a completely integrated Hierarchy System as shown in Fig. 4.

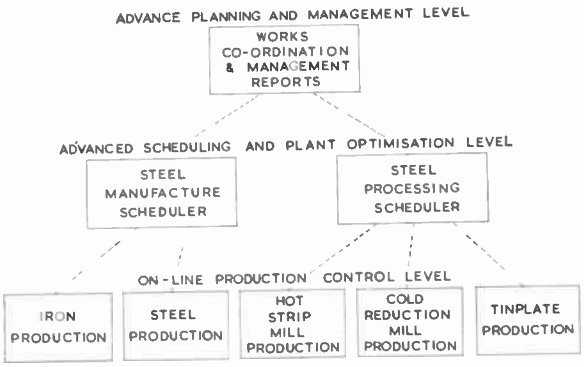


Fig. 4. Hierarchy system for integrated steel plant.

A complete record of the state of the plant is maintained within the computing systems. Thus it is possible to interrogate the system to inquire the position of a given customer's order. An extension of this facility will provide a prospective customer with a delivery date for a number of different types of steel from which he may choose the combination most suitable for his needs. It is quite practical to use direct interrogation and order entry from the customer's premises using the same style of approach as is already used for airline bookings. This may appear to be a long way from steel manufacture, but on reflection it will be seen that the principles of interrogation and reservation are precisely the same.

5. Conclusions

This paper has described the philosophy of approach and some practical methods for the integrated control of a complete plant by the use of the Hierarchy System. The sections of the system can be installed separately as each will perform a useful function on its own, and gradually additional sections can be added. The facility of gradually building up the control program is provided, and it is therefore unnecessary to have available a mathematical model of the plant. The method of using the sections of the Hierarchy System apply equally to small as well as large plants, the former using only the lower levels, whilst for the latter there may be numerous levels.

6. References

1. J. A. Reine, "Performance of a solid-state process computer", *Instruments & Control Systems*, **32**, p. 1820, December 1959.
2. R. D. Eisenhardt and T. J. Williams, "Closed-loop computer control at Luling", *Control Engineering*, p. 103, November 1960.
3. Leo Walter, "World's largest crude pipe-still controlled by computer", *Petroleum*, **24**, p. 426, November 1961.
4. R. G. Beadle and W. E. Millar, "Improving thickness control in hot strip mills", *Control Engineering*, p. 94, July 1961.
5. M. J. Feldmann and J. W. Purssell, "Monitoring Logging and Computing Performance for an Existing Coal and Oil Fired Station", American Power Conference Paper, 22nd March, 1961.
6. Arthur Freilich, "What's doing in computer control", *J. Instrum. Soc. Amer.*, **8**, p. 49, September 1961.
7. V. J. McMullan and P. Cox, "The reliability of an experimental transistorized data handling system", *J. Brit.I.R.E.*, **22**, No. 1, pp. 17–29, July 1961.
8. R. F. Shaw, "Preserving accuracy in digital data transmission", *Control Engineering*, p. 119, November 1961.
9. J. F. Roth, "Applications of computers to process control", *Chemical and Process Engineering*, October/November 1961.

Manuscript received by the Institution on 24th February 1962. Paper No. 745/C42.

POINTS FROM THE DISCUSSION

Mr. P. Huggins: I would like to ask Mr. Roth to enlarge on one or two thorny problems in connection with the acquisition of shop floor data. Let us take Fig. 3 as an example. A large proportion of the cost, to say nothing of the engineering difficulties, lies in the shop floor region covered in Fig. 3 from "on-line input and output control" and then extending to the transducers or other peripheral equipment. The engineering difficulties are (a) standard transducers are not always available, (b) peripheral equipment must be mechanically sound enough to meet a proper industrial specification†. The economic difficulty stems from the multiplicative element in the calculation. If each installed transducer cost £100 and there are *n* inputs, the costs soon mount up alarmingly. I would guess the *installed* cost of one of Mr. Roth's sub-nets (say on-line input and output control No. 1 and its associated shop floor equipment) to be in the order of

† D. Shaw, "A specification for electronic equipment for use in heavy industry", *J. Brit.I.R.E.*, **24**, pp. 133–9, August 1962.

£15,000. The design study associated with making it work could well double this figure. So the questions I would like to ask are:—

(1) Does the equipment of the sub-net exist in a form suitable for its environment (i.e. heavy steel mill)?

(2) One of the advantages claimed for hierarchial control is that the sub-nets are self-contained and the overall integration can be adopted piecemeal. How does Mr. Roth manage to find economic justification for the sub-nets, on their own merits, as distinct from the complete system?

Finally, I would like to ask about buffer storage at the lowest level. Mr. Roth vaguely mentioned "paper tape, cards, etc." Using tape/card punching equipment is all right if it can be adequately loaded and is the correct medium for computer input. Because of the sporadic nature of shop floor generated data, adequate loading means complicated and expensive multiplexing schemes. Considering this problem a few years ago led me to develop a slow-speed inexpensive core storage system.† However, on more mature consideration, I see economic and practical advantages to mechanical type buffer storage at shop floor level. This can be cheap, easily understood and serviced by shop floor personnel, and fast enough for most practical applications. By mechanical type storage, I mean such devices as the pin-wheel. Has Mr. Roth considered the use of slow speed mechanical buffer storage for this kind of work?

Mr. J. F. Roth (*in reply*): The equipment required to form the sub-units does exist in a modular form constructed so as to be suitable for industrial environments. By having a standard range of units which are all compatible, the cost of a given system is reduced to an absolute minimum. In addition the systems or plant engineer is not worried about detailed circuit design but need only be concerned with the function of the modules. Thus sections of the plant can be investigated independently, and the advantages of automatic control achieved piecemeal. Of course, if the detailed investigations of the separate parts of the plant has been preceded by an overall study, the resultant control system may well show an improvement. Thus justification for a sub-net can be made both on its own merits and in relation to the complete system.

The buffer store I referred to was for the purpose of recording information intended for another section of the control hierarchy which was temporarily out of action. When the section became operational again, the stored information could be rapidly entered, and the section thus able to "catch up". For such an application a punched paper type store is a most useful and economic device. However, for input of information from the shop floor, the requirements of a buffer store to ease the data acquisition problem are rather different. In this case only one message, which is normally quite short, is entered at a time, and arrangements can be made to ensure that one message is read before the next can be entered. For this type of temporary storage, multiway switches with suitable coding arrangements form a most convenient and cheap unit. These will, in addition, provide the facilities of being easily understood and serviced by shop floor personnel, and quite fast enough for most purposes. At the same time the coding of the information can be made to meet the requirements of the data acquisition system.

† W. H. Baker and P. Huggins, "Mechanizing shop floor form filling", *Trans. Soc. Instrum. Tech.*, **14**, No. 1, pp. 67–80, March 1962.

Mr. J. P. Dean: Since the system described may be applied to parts of the plant and the gradual build-up may not produce the most sophisticated system, is it possible that a point may be reached where any increase in the system produces no improvement in operation whereas a complete system designed around the entire plant might produce a better result?

Mr. Roth (*in reply*): To discuss the correctness of this supposition would require the detailed examination of a number of applications. If this were done—which cannot for obvious reasons be included here—it would be found that this situation does not exist in practice. The extension of the hierarchy control system to another section of the plant would only be proceeded with if it could be shown to be justifiable. The experience gained from the other sections previously installed would be used in arriving at this decision, and in fact in this way an uneconomic extension would be prevented. Had an overall control system been applied in the first place, the impracticability of bringing that section of the plant within the control system may not have been foreseen. Having finally arrived at a complete control system, the difference in performance between the initially complete system and the built-up system would be negligible.

Mr. D. Shaw (*Associate Member*): The effect of this hierarchy system will be to convert the operator into a supervisor who then has little experience of plant operation when breakdown occurs. How will you ensure that both management and operative are adequately trained to meet this situation?

Mr. Roth (*in reply*): This is a very real problem which must be successfully solved if control systems of the type under discussion are going to be universally accepted. I do know of one case where such a situation arose, although here the system was being used in an advisory capacity, producing performance coefficients. Previously the calculations, which are quite complex, had to be done by hand, and when the computer failed there was no one immediately available who could remember how to do them!

The solution to this problem can be seen by referring to such industries as atomic reactors where simulators are used for training operators. In a similar way a simulator can be constructed for the plant, and thus the operators can be trained to control it manually, and obtain practice in doing this.

Alternatively it may be possible to arrange for periods when the process is controlled manually so that the operators maintain their skill. For a large plant this will most probably not be practical in which case a simulator will have to be used.

U.D.C. 535.331: 681.33

The Application of Analogue Computing Techniques to Automatic Spectrochemical Analysis

By

F. SMITH, B.A.†

AND

K. P. RIPPON, M.A.†

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

Summary: The direct reading spectrometer has in recent years been widely adopted as an analytical tool in the manufacture of metallic alloys. So far the read-out of such instruments has normally been in arbitrary units and the concentrations of the constituent elements have had to be determined by referring the instrument output to graphs or tables. Various methods of converting measured intensities into actual concentration readings automatically have been devised, but have not gained wide acceptance in industrial circles, where reliability and simplicity are of great importance.

This paper describes the development of a method which employs diode function generators to cater for the non-linearity of the calibration curve of the spectrometer and is considered to have high inherent reliability. Capacitor storage devices are used in auxiliary circuits to enable corrections to be made to the output for third element interference effects.

1. Introduction

At the Brit.I.R.E. Industrial Electronics Convention held at Oxford in 1954, F. Holmes described the development of automatic spectrometers with electronic measuring circuits for routine production control.[1] Such "direct reading" spectrometers are now widely adopted as the standard means of analysis in the metal manufacturing industries. In particular, the steel industry has been widely equipped in this way during the last three years and a photograph of a typical steelworks installation is shown in Fig. 1.

The principal advantage of these instruments is the speed of the analysis. A typical steel sample can be analysed for the ten most important elements in less than two minutes and this is particularly important since the development of oxygen steel-making has reduced the process time in some cases to less than fifteen minutes.

1.1. *Fundamentals of Photoelectric Spectrochemical Analysis*

Analytical work has long been carried out by means of emission spectroscopy. In this technique, the optical spectrum of the material to be analysed is excited by passing an electrical discharge between electrodes formed of the material and of a standard material, the latter being chosen so that its spectrum does not interfere with the analysis.

The light emitted by this discharge is dispersed by means of a prism or grating spectrometer and the intensities of selected spectrum lines emitted by the elements for which the material is to be analysed are measured. In photo-electric instruments the required spectrum lines are isolated by means of exit slits arranged round the focal plane of the instrument, the light passing through each slit being arranged to fall on the cathode of a photomultiplier. Thus the magnitude of the photomultiplier collector current is a measure of the concentration of the required element.

A schematic diagram of the type of measuring circuit most commonly employed is shown in Fig. 2.

During the period that the discharge is running the collector of each photomultiplier is connected to a

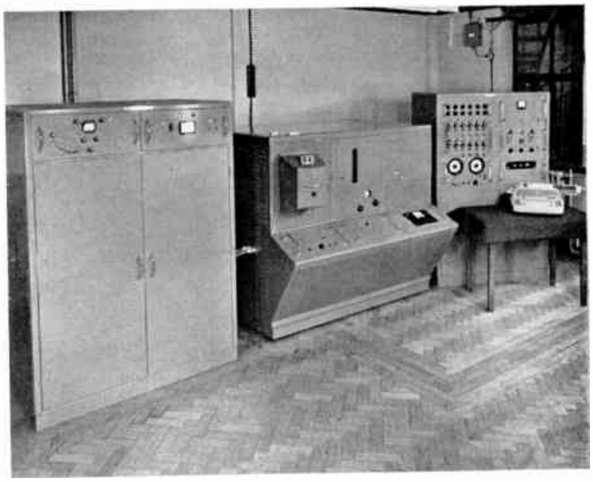


Fig. 1. The E600 direct reading spectrometer.

† Hilger & Watts Ltd., 98 St. Pancras Way, Camden Road, London, N.W.1.

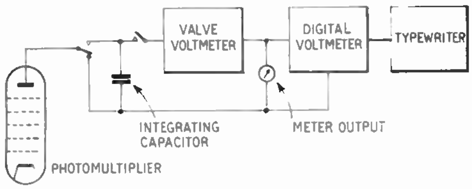


Fig. 2. Schematic diagram of direct reader measuring circuit.

capacitor which integrates the collector current. At the end of this period the capacitors are isolated and measurement of their potentials is made sequentially by means of a high impedance valve voltmeter, often followed by a digital voltmeter and electric typewriter or line printer.

The operator must now convert the voltage reading for each element into concentration, using a working curve of the type shown in Fig. 3. This conversion is time-consuming and a considerable saving can be made by converting the readings to concentration automatically, with a consequent reduction in the delay before the analysis is available to the furnace operator and the elimination of errors occurring in the interpretation of the graphs or calibration tables.

The interpretation process is further complicated by the fact that in some analyses, the reading obtained for one element is dependent upon the concentrations of other elements. This condition is shown in Fig. 4, which is a working curve for sulphur in steel and which is modified according to the amount of carbon in the sample.

2. Method of Conversion

Various devices for converting measured intensity ratios into actual concentrations have been developed. These fall into three categories.

(1) Those which make a quadratic approximation to the working curve and use either digital or analogue methods for calculation.

(2) Non-linear analogue-to-digital converters.[2]

(3) Analogue devices using adjustable function generators.

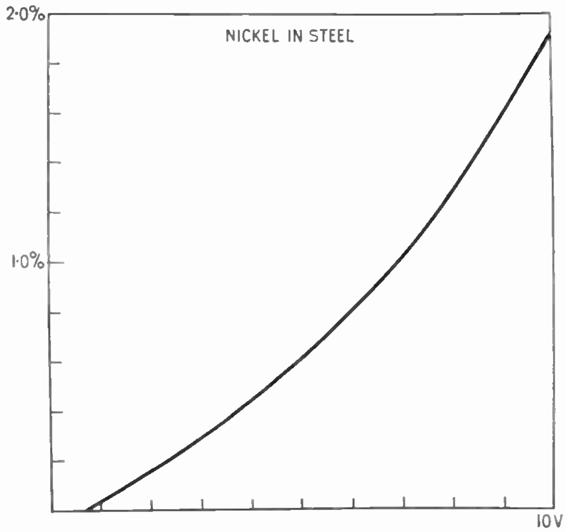


Fig. 3. Typical calibration curve and presentation of results.

Sample	Time	Fe	C	S	P	Mn	Si	Cu	Cr	Ni	Sn	As	Al
6001	230	570	378	249	628	323	219	741	187	828	483	255	071
6011	235	570	866	362	338	859	729	370	132	163	697	197	060
6013	234	570	344	156	457	265	478	272	104	132	338	592	660

Source conditions 20 μF 3 ohms.

The requirements to be met by a practical system are as follows:

(i) It should be automatic in operation so that concentration readings may be obtained directly from the spectrometer.

(ii) It must be extremely reliable since it will be employed continuously for process control.

(iii) Initial adjustment must be simple.

(iv) Drift in characteristics must be small to avoid the need for frequent re-calibration.

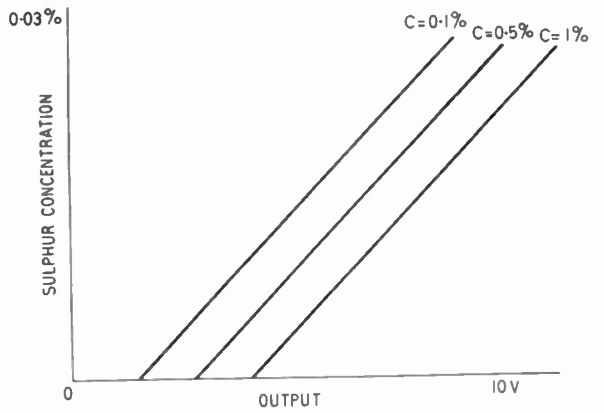


Fig. 4. Third element interference with instrument calibration.

(v) It should be compatible with existing installations. It would be desirable to retain the linear digital voltmeter which is already fitted to most Hilger direct reading spectrometers.

Practical assessments of method (1) using special purpose computers of both digital and analogue types have shown that not all of these requirements can be met. In particular, the best quadratic approximation to an arbitrary curve entails a certain arithmetical dexterity and sometimes is not sufficiently exact over an adequate range. This approach is also the most expensive one.

Method (2) is known to be practicable but all devices using this principle have been cumbersome. It is not compatible with the existing digital voltmeters.

However, method (3) does fulfil the requirements laid down and appears to be a relatively cheap solution.

The voltage output from the spectrometer valve voltmeter is passed into the function generator which produces a voltage directly proportional to concentration. This latter voltage may be measured on a meter or by the existing digital voltmeter. Hence the output is a direct concentration reading.

Since the function generators consist mainly of passive resistance elements, reliability is high, and drift in characteristics is low. A function generator can be easily and quickly set to approximate any arbitrary curve within its range, and the approximation does not depend on a knowledge of the mathematical equation of the curve, but can be made as accurate as required by a suitable choice of parameters.

2.1. *Method Used*

The principle of the device is shown in Fig. 5. The calibration curve to be used is divided into segments by selected ordinates, and the multi-segment approximation to the curve is then reproduced electrically by means of switched resistor networks. Generally, both the slope and the breakpoints (points at which the slope changes) of the segments are adjustable,[3] but in this application it was decided to have fixed breakpoints spaced at equal intervals.

This simplification was possible because the calibration curves to be dealt with show no sharp changes in radius of curvature, and indeed the maximum departure from linearity usually found is approximately 20%. It can be shown[4] that if e is the maximum tolerable error of approximation to the curve $y = F(x)$ in using the tangent to the curve at $x = a$ at a point $\delta x/2$ away from a, then

$$\delta x = \frac{16e}{F''(a)}$$

This gives the separation δx of the breakpoints, for a given error e.

In this case, ten segments are provided, which give an acceptable accuracy of approximation to all normally encountered curves.

Direct reading spectrometers are normally set up to measure twelve or more elements, some of which may occur in more than one concentration range. Each element and concentration range has a characteristic shape of working curve and there must be one function generator for each curve. In practice this may call for twenty or more generators, each being connected in circuit as required.

Many circuits of function generators using diodes as switching devices have been published, but for this application the bridge circuit used offered the advantage of ease of setting both positive and negative slopes to the segments with only one control per segment. Circuits of this type have been described in various forms[5] but this particular application allowed simplifications to be made, especially the use of fixed breakpoints, which helped in reducing both the bulk and cost of the apparatus.

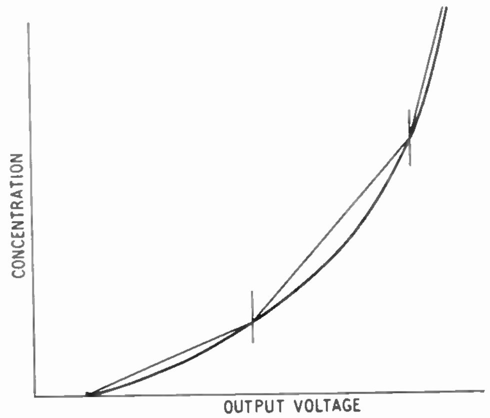


Fig. 5. Multi-segment approximation to calibration curve.

2.2. *Description of Circuit*

The basic unit of the circuit is shown in Fig. 6. Amplifier 1 is a high gain d.c. amplifier arranged with a suitable feedback ratio to give an overall gain of unity with sign reversal, and amplifier 2 is similar, but arranged to sum the voltages applied to the two input resistors R5 and R6.

The application of a voltage more negative than the diode biasing voltage $+V$ to the input of the bridge results in equal potentials appearing at points A and B (since R3 = R6) and the output of amplifier 2, which is the sum $A+(-B)$ remains at zero. If, however, the applied voltage is increased to a value greater than $+V$ so that the diode conducts, the potentials at A and B will now depend on the position of the wiper of the potentiometer RV1, and will not, in general, be equal. Hence an output appears at amplifier 2 which depends on the sign and magnitude of $(A-B)$, a quantity determined by the setting of the potentiometer.

The characteristic of this circuit is also shown in Fig. 6. Nine of these bridge circuits are connected in parallel at the summing junctions of the amplifiers, with the diodes returned to progressively increasing voltages (Fig. 7). The first segment is set by means of a simple potentiometer (RV10) since the calibration curve is drawn only in the first quadrant, and therefore negative slopes are not needed for this segment. A biasing device is provided which allows for the possibility of an intercept on either axis.

All the resistors and potentiometers for one function generator are mounted on a printed circuit card which is made as a plug-in unit. The cards are permanently plugged into a special chassis, each one being selected in turn by energizing a relay which connects the function generator to the two amplifiers and to the input voltage.

In order to make full use of the available output voltage swing of the d.c. amplifiers used, the output of the valve voltmeter, which has a range of 0–10 V, is amplified ten times before being fed to the input of the function generators, and this procedure also helps to reduce the effect of temperature drift in the switching diodes.

The nine fixed breakpoint voltages are derived from a chain of Zener diodes fed from a stabilized d.c. supply. All the function generators used are connected permanently to the breakpoint potentials which greatly reduces the amount of switching required to bring a function generator into circuit. The first segment can be given a slope in the range 0–3 and each subsequent segment has a range of ± 1, the intercept control allowing for an intercept of 10% on either axis.

A special unit is provided to hold the function generator and facilitate its adjustment. A ten-way switch is provided which feeds voltages corresponding to the ten breakpoint voltages into the function generator. In use, the operator selects each position of the switch in turn and adjusts the corresponding potentiometer in the function generator to give the concentration reading determined from the working

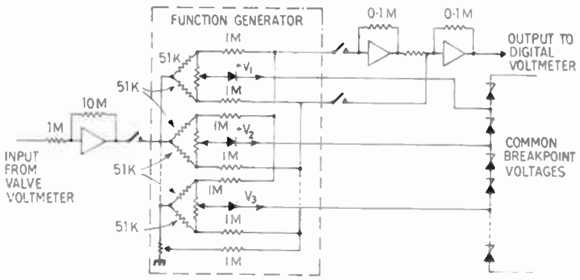


Fig. 7. Circuit to simulate a complete curve.

curve. The adjustments of the segments are independent providing they are made in the proper order, i.e. starting from the origin of the working curve.

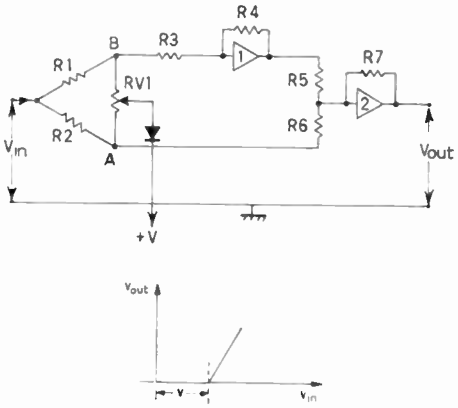


Fig. 6. Circuit of single segment and its characteristics.

2.3. *Analogue Store*

In order to make the correction of inter-element effects an automatic process, some form of store is required which will store the results of measurements on given elements, so that correction can be made to other elements measured later in the cycle. It sometimes happens that an element concentration must be corrected for the presence of two other elements simultaneously. However, not all elements are affected by the same interferences and by a suitable arrangement of the reading order, the amount of storage capacity can be kept small, since the instrument can be programmed to store measurements and erase as required, thereby making maximum use of the store.

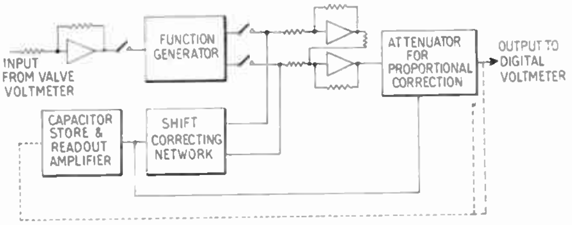


Fig. 8. Schematic diagram showing correcting circuits.

The storage device used is simply a polystyrene capacitor which is mounted in a desiccated airtight box to reduce leakage. The output voltage which is fed to the digital voltmeter is stored on this capacitor when necessary, and then "read out" later by means of a high impedance d.c. amplifier. In fact, four capacitors are provided, and switched by relay contacts of high insulation resistance. Two amplifiers are used so that two corrections can be made simultaneously (Fig. 8).

The accuracy of this store is governed by the leakage rate of the polystyrene capacitor, the input capacitance and grid current of the d.c. amplifier. Taking the time constant of the capacitor as 500 000 ohms-farads, the input capacitance of the amplifier to be 10 pF and its grid current as 10^{-10}A, a maximum error in the stored voltage of 0·1% will arise in the course of a measuring cycle. Since the correction never amounts to more than 20% of the result the error due to imperfections in the store is insignificant.

The voltages on the capacitor are fed via the d.c. amplifiers to resistance networks, which add a correcting voltage to the output of the element being measured.

Another form of correction is sometimes required which demands that the output voltage of a given measurement should be reduced by a percentage depending on the amount of the interfering element present. This is done by means of a five-step switched attenuator, the switches being transistors arranged so that the number operated depends on the magnitude of the signal from the store unit.

3. Conclusion

Spectrochemical analyses generally yield reproducibilities of 0·5%–5% relative to the concentrations of the elements being analysed (Table 1). This figure varies with the element concentration and on the type of alloy.

The computer described is capable of converting spectrum line intensities into concentrations with considerably greater accuracy than the overall accuracy of analysis.

The adjustment of the function generators is readily carried out to one scale part of a three decade digital voltmeter, i.e. 0·1% f.s.d. The error due to the multisegment approximation has been found not to exceed 0·25% for a typical low alloy steel analysis. The adjustment of the function generators did not vary by more than 0·2% during a test lasting 1000 hours, which again is insignificant in comparison with the basic accuracy of the system.

4. Acknowledgment

The authors wish to thank Dr. A. C. Menzies, Director of Research, Hilger and Watts Ltd., for permission to publish this paper.

5. References

1. F. Holmes, "Electronics in automatic direct reading spectrometers for the analysis of metals", *J. Brit.I.R.E.*, **15**, pp. 163–76, March 1955.
2. S. Minami, H. Yoshinaga, S. Fujita, "A digital circuit for measurement of spectral line intensities", *J. Opt. Soc. Amer.*, **51**, p. 674, June 1961.
3. W. J. Karplus and W. Soroka, "Analog Methods, Computation and Simulation" (McGraw-Hill, New York, 1959). W. Soroka, "Analog Methods in Computation and Simulation" (McGraw-Hill, New York, 1954).
4. C. C. Ritchie and R. W. Young, "The design of biased diode function generators", *Electronic Engineering*, **31**, pp. 347–51, June 1959.
5. G. A. Korn and T. M. Korn, "Electronic Analog Computers", p. 297 (McGraw-Hill, New York, 1956).

Manuscript first received by the Institution on 1st March 1962 and in final form on 15th March 1962 (Paper No. 746/C43).

Table 1. Typical Computed Results

	Time	Fe	C	S	P	Mn	Si	Cu	Cr	Ni	Sn	As	Al
4761	240	569	·155	·0178	·0572	·381	·072	·345	·142	·732	·0236	·0280	·0029
	237	569	·150	·0174	·0545	·381	·042	·340	·138	·730	·0230	·0264	·0049
4762	240	569	·403	·0203	·0459	·955	·247	·225	·739	·726	·0639	·0494	·0243
	244	569	·393	·0198	·0474	·965	·248	·227	·741	·736	·0642	·0494	·0240
4763	240	569	·199	·0182	·0312	·151	·373	·489	·260	·400	·0140	·0998	·0263
	240	569	·210	·0202	·0345	·150	·383	·503	·262	·409	·0153	·0997	·0263

Source conditions 20 μF 3 ohms.

APPLICANTS FOR ELECTION AND TRANSFER

The Membership Committee at its meeting on 26th July last recommended to the Council the election and transfer of 18 candidates to Corporate Membership of the Institution and the election and transfer of 33 candidates to Graduateship and Associateship. In accordance with Bye-Law 21, as adopted at the Special General Meeting held on 23rd May 1962, the Council has directed that the names of the following candidates shall be published under the grade of membership to which election or transfer is proposed by the Council. Any communications from Corporate Members concerning these proposed elections must be addressed by letter to the Secretary within twenty-eight days after the publication of these details.

CORPORATE MEMBERS

Transfer from Associate Member to Member

BARTLEY-DENNISS, Colonel George Arthur, R.A. *Swindon, Wiltshire.*
WALDRON, Richard Arthur, M.A.(Cantab). *Chelmsford, Essex.*

Direct Election to Associate Member

BAUST, Norman Charles. *Luton, Bedfordshire.*
LEAHY, David John, B.Sc.(Eng.). *Arlesey, Bedfordshire.*
PEAKE, Donald Thomas. *Gravesend, Kent.*
SCOTCHER, Roland John, B.Sc. *Weymouth, Dorset.*
SINCLAIR, Bryan Allen. *Harpenden, Hertfordshire.*

Transfer from Associate to Associate Member

BURGESS, Philip Hugh George. *Coventry, Warwickshire.*
GENT, Leslie Frederick. *Stevenage, Hertfordshire.*

Transfer from Graduate to Associate Member

BUBLOZ, Rene Maurice. B.Sc.(Eng.). *Brighton, Sussex.*
CUNNINGHAM, Squadron-Leader Robert Alexander, R.A.F. *High Wycombe, Buckinghamshire.*
GILHAM, Alan Edward. *Morden, Surrey.*
HOWARD, Alwyn George. *Stafford.*
McCARTNEY, Thomas. *Stockport, Cheshire.*
McGOW, Peter Charles. *Kuching, Sarawak.*
PULLINGER, Gordon Patrick. *Birmingham.*

Transfer from Student to Associate Member

HARRISON, Brian. *Northfleet, Kent.*
TIN-TUN, Maung, B.Sc. *Rangoon, Burma.*

NON-CORPORATE MEMBERS

Direct Election to Associate

FLETCHER, Albert William Arthur. *Leicester.*
HOBLYN, Squadron-Leader Frank John, R.A.F. *Dunfermline, Fife.*
LEE HON WING. *Hong Kong.*

Transfer from Student to Associate

SIMMONS, Roy. *London, N.9.*
SOUTHGATE, John William. *Chislehurst, Kent.*

Direct Election to Graduate

AGUTTER, Derek Stanley. *Sidcup, Kent.*
BAILEY, Robert Edwin John. *London, N.13.*
BURTON, Brian. *Romford, Essex.*
CANT, Graham Oliver. *Malvern Wells, Worcestershire.*
*CHARANJIT SINGH. *Calcutta, India.*
DAVIES, James John. *Birmingham.*
DOW, Barry John. *Northwood, Middlesex.*
GREEN, Robert Frederick Dennis. *Artarmon, Australia.*
HOLROYD, Flight-Lieutenant Frank Martin, R.A.F. *Southend-on-Sea, Essex.*
HOWARD, James Scott. *Chesham, Buckinghamshire.*
IRELAND, Peter Morris. *Northolt, Middlesex.*
*KINALLY, Dennis Raymond. *Guildford, Surrey.*
LEIGH, Anthony William. *Hertford.*
MORGAN, Eric John, B.Sc. *Wembley, Middlesex.*
NOTTINGHAM, Major William Joseph, B.A. *Ottawa, Canada.*
PAYNE, Roy David. *London, N.22.*
SHORT, Peter Ronald. *Hounslow, Middlesex.*
SMITH, Henry Robert James. *Marlow, Buckinghamshire.*
SWEET, Anthony William. *London, N.9.*
THOMAS, Victor Francis. *Wheathampstead, Hertfordshire.*
WOO-SAM, Jeremiah Calvin. *London, N.15.*

Transfer from Student to Graduate

BALASUBRAMANIAN, Mallagi Ramanatha, M.Sc., B.Sc. *Poona, India.*
JACKSON, Brian David. *Kisumu, Kenya.*
McFADZEAN, Gordon Robert. *Currie, Midlothian.*
MOHIDEEN, Mohamed Junaid. *Dehiwela, Ceylon.*
MORGAN, David Sydney. *Bristol.*
PHILLIPS, Edmund. *Cambridge.*
TONGUE, Brian Leonard. *Cheshunt, Hertfordshire.*

STUDENTSHIP REGISTRATIONS

The following students were registered on the 26th July.

AJAYI, Solomon Olawale. *Lagos.*
ALLSOP, David A. *Biggleswade, Bedfordshire.*
ANANTHAPADMANABHA, B.Sc. *Bangalore.*
ARORA, Dhalla Ram. *Panjab, India.*
AZIZ, Abdul. *Risalpur, West Pakistan.*
BALAKRISHNAN, A., B.Sc. *Ernakulam, S. India.*
BALAKRISHNAN, K., B.Sc. *Salim, India.*
BALARAMA. *Bangalore City.*
BHEEMA-RAO, Lakkur, B.Sc. *Bangalore.*
BOOT, Frederick William. *Crowborough, Sussex.*
BRAMHAVAR, S. M., B.Sc., M.Sc. *Bijapur, India.*
BRANCKER, Richard William S., B.A. *Montreal.*
CARTER, Dourne Harry. *Norfolk.*
CHANNABASAPPA, N. I., B.Sc. *Bangalore.*
CHELVAKUMARAN, V. R., B.Sc. *Bangalore.*
CHIDAMBARA-MURTHY, V., B.Sc. *Bangalore.*
CHONG, Fook Choy. *Coventry, Warwickshire.*
DE BONO, Victor Henry. *Malta G.C.*
DESHPANDE, Madhav D., B.Sc. *Bombay.*
EDRIDGE, John William. *Cambridge.*
FUSSELL, Gordon E. *Bracknell, Berkshire.*
GAUR, Bibhuti B., B.Sc. *Allahabad, India.*
GOPAL-RAO, T. H. *Bangalore.*
GUNANAND, T. R. *Bangalore.*
HARVEY, Basil William. *Bombay.*
HEASMAN, John B. I. *Lithgow, Australia.*
HEDGE, Anant Shivaram. *Siddapur, India.*
HUEN, James Cheung C. *Hong Kong.*
JAYARAM, S. B. *Bangalore.*
JOHNSON, John Vivian. *B.F.P.O. 40.*
JONES, Cyril Edwin. *Bletchley, Bucks.*
JOSHI, Arvind Govind. *Mysore State.*
KAMALESH, Mysore N., B.Sc. *Bangalore.*
KAPOOR, Suresh Kumar. *Bangalore.*
KARANDIKAR, Subash. *Mysore.*
KARPI, Veerabhadrappa D., M.Sc. *Hubli, India.*
KESHAVAREDDY, Hoodi M. *Bangalore.*
KRISHNA-MURTHY, Tarikee, B.Sc. *Chickmagalur, India.*
KULKARNI, Manohar Panditrao, B.Sc. *Bombay.*
LEONARD, Edward Malcolm. *Huddersfield.*
McSTAY, R. E. D. *Christchurch, New Zealand.*
*MALHOTRA, Madan Mohan. *New Delhi.*
*MASAND, Bhagwan Menghraj. *New Delhi.*
MEHTANI, Mahendra N., B.Sc. *New Delhi.*
*MOHAMMAD IBRAHIM, Siddiqi. *Baghdad.*
MOHINDER SINGH, Bains, B.Sc. *Southall.*
MORRISON, Alexander I. *Thurso, Caithness.*
MURTHY, Annavarapu V. R., B.Sc. *Nagpur.*
NAGARAJA-RAU, S. *Bangalore.*
NAGENDRAN, L. R., B.Sc. *Tumkur, India.*
NARASIMHA-SWAMY, B., B.Sc. *Bangalore.*
NARAYANA-RAJ, Y. R., B.Sc. *Bangalore.*
OKEKE, Christian N. *London, N.7.*
ORIAN, Jean Serge M. *London, N.4.*
PADMANABHAN, T. M., B.Sc. *Bangalore.*
PAGE, Kenneth E. *Salisbury, S. Australia.*
PEARSON, Godfrey. *Bangalore.*
PRABHAKARA-RAO, Kunala. *Bangalore.*
PRATT, David M. *Bingley, Yorkshire.*
RAJDERKAR, M. K., B.Sc. *Nasik City, India.*
RAMACHANDRA, B. V. *Bangalore.*
RAMACHANDRA, K., B.Sc. *Bangalore.*
RAMANNA, Hanagud S., B.Sc. *Bangalore.*
RAMA-CHANDRA-BHATTA, H. N., B.Sc. *Koppa Kadur, India.*
RAMAKRISHNA, J. S. *Bangalore.*
RAMAKRISHNA SARMA, Kalluri. *Bangalore.*
RAMAKUMAR, R., B.Sc. *Chittoor, India.*
SANU, Saluel O. *Agege, Nigeria.*
SATISH KUMAR TYAGI, M.Sc. *Buland Shahr, India.*
SCOTT, Marthinus J. *Johannesburg, S. Africa.*
SHARMA, Babu R. *Dehra Dun, India.*
SHARMA, Suraj B., B.Sc. *Bhopal, India.*
SHEKAR, Chikmagalur N. C., B.E. *Bangalore.*
SIVANAPPAN, M., B.Sc. *Ernakulam, India.*
SREEDHARA-MOORTHY, N. S., B.Sc. *Bangalore.*
SRIDHARA, A. R., B.Sc. *Bangalore.*
SRINIVASA-RAO, B., M.Sc. *Srikakulam, India.*
STEPHENS, Garnik Gerald. *London, W.5.*
SUBBA-RAO, Chowdaiah L. S., B.Sc. *Bangalore.*
SUBBA-RAO, China M., B.Sc. *Andhra, S. India.*
SUBBA-RAO, J. V., B.Sc. *Bombay.*
SUBRAMANIAN, Capt. H., B.Sc. *Mhow, India.*
SUBRAMANYAM, S., B.Sc. *Chengleput, India.*
TAN EWE CHENG. *Penang, Malaya.*
THAKAR SINGH DATTA. *Ludhiana, India.*
THOMAS, V. I. *New Delhi.*
TILBURY, David C. *London, N.W.4.*
UWABOR, Godfrey A., B.Sc. *Yaba, Nigeria.*
VARMA, D. S. *Karlsruhe/Baden, West Germany.*
VENKATADRI, Veerapanaei. *Krishna DT., India.*
VENKATARAMANAIAH, V., B.Sc. *Kolar, India.*
VENUGOPALAN-NAIR. *Trichur, India.*
VOHRA, Rajeshwar L., B.Sc. *Agra, India.*

*Reinstatements.

U.D.C. 621.37/8.004: 62

A Specification for Electronic Equipment for use in Heavy Industry

By

D. SHAW, B.Sc.

(*Associate Member*) †

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

Summary: In the past inadequate attention has been paid to the mechanical design of electronic equipment for heavy industrial environments, resulting in poor life. A specification has therefore been produced which sets out the conditions the equipment will have to endure. The specification discusses general design considerations, advocating unit construction and other points which have been found of value, and insisting on adequate handbooks and circuit diagrams. Mechanical design requirements are similar to Government specifications. Typical environmental characteristics are listed and a series of proposed tests for the equipment is given. Mains supply voltage specifications are also covered.

1. Introduction

The history of electronic equipment development, were it ever to be written, would show that the vast majority of equipments were intended for use either in laboratory conditions or in connection with domestic entertainment. It is generally accepted that laboratory equipment is maintained by highly trained personnel who are continuously in attendance and that, even then, appreciable delays due to breakdown or malfunction may occur. On the other hand, although domestic equipment does not have the advantage of this type of maintenance, its hours of operation are usually restricted, often severely so, and simple feedback systems can be employed in most cases to mask the effects of variations in equipment performance. The demand for reliability in such systems has therefore been met by the use of design techniques adequate to ensure short-term reliability in ambient conditions of temperature, vibration, etc., which are quite moderate. So much is this the case that it is by no means uncommon to find equipments which cannot be used in winter without preheating and others which fail regularly on hot summer days. Failure due to other variations in ambient conditions has been accepted quite cheerfully also, one writer reporting that his radio failed after about half an hour's use in his bathroom and being quite surprised it had lasted so long.

At the same time there has been a general recognition that some attention to improved reliability in difficult ambient conditions was required and efforts were made in this direction. The advent of the computer accelerated this requirement since large numbers of components were grouped together, there now being so many that, if the old standards of reliability were not improved on, then the equipment had little chance of ever operating at all. The first, and most obvious line of attack, was to provide the equipment with its own environment which could be made as mild and beneficial as possible. This led to the production of equipments which required a greater expenditure on their housing, air conditioning, vibration isolation and mains stabilization than their first cost and it came to be realized that whilst reliability could always be gained in this way, it might well be more economic to "design in" reliability than to attempt to add it on afterwards.

Unfortunately, since the requirement for this extra reliability occurred in a comparatively small number of equipments, there was little experience in designing for reliability and hardly any attention was given to it in those courses dealing with electronic design principles. In general, the lecturers giving these courses were laboratory-employed and designed equipment for laboratories, so that it was not surprising that they overlooked reliability since to them it was not a pressing problem. Hence at first it was not possible to design for reliability since it was not known what factors influenced it and even now the data on which some reliability techniques are based are meagre.

As a starting point it was therefore argued, as an extension from mechanical engineering reliability design, that the reliability of an equipment should be dependent on the individual reliabilities of its components and on the amount by which they are under- or over-run. The first premise is obviously true although the dependence is by no means always a

† The United Steel Companies Limited, Research and Development Department, Swinden Laboratories, Rotherham, Yorks.

straightforward one but the second premise can be false and, although considerable work has now been done, more work is needed to determine under what conditions components are reliable. Such data also serve to pinpoint causes of unreliability which may then be rectified where possible. There are, therefore, now available three methods of increasing the reliability of an equipment in a difficult environment: (1) the use of reliable components, (2) the use of circuits designed to cope reliably with the environment, and (3) the modification of the environment to a less demanding one.

One of the more difficult environments which electronic equipment has to meet is that of heavy industry. Here it may be exposed to heat or cold, dampness or dryness, vibration, shock, dust, fume, corrosion, running water or other fluids. These may occur singly or in any combination and the rate of change of these quantities can be large particularly where batch processes occur. For some time there has been a growing requirement for electronic equipment for use in heavy industry, but its use has often been impractical because of its poor record of reliability. There have been many occasions when equipment stated to be reliable in heavy industrial conditions has been found to fail and investigation has often shown that there has been a lack of appreciation of the effects of these conditions on the equipment. However, the time has come when British heavy industry must have electronic equipment to compete effectively with the rest of the world and means must be found to make it reliable. In order, therefore, to guide the author's organization when purchasing equipment and to guide the electronics industry when supplying this equipment, a specification has been produced which summarizes present knowledge of environments and of the modes of design and construction which have succeeded to some extent in the past. In general it will be necessary to use all three methods of reliability improvement mentioned previously, namely the use of reliable components, reliable circuits and the modification of environments by appropriate housings and even then it may be that there will be positions on plant where electronic equipment cannot reliably be used. However, it is our hope that by the use of this specification and its refinement through experience we shall arrive at a time when the saying that "an electronic equipment may replace two men but it often requires three to maintain it" is no longer remotely true.

2. Reliability

As with many other words which have become scientific terms, reliability has two meanings. It is used in its general sense to express the ability of a system to survive in its environment and to carry on doing its job. It is in this sense in which it is used in this Specification. No attempt has been made to define the quantitative reliability index values which are inferred by the second, more narrow but more precise, meaning of the word. This has been done in order not to restrict the equipment supplier. However, some figures have had to be stated for our own protection and a compromise has therefore been made by pointing out that a fault-free life of 1000 hours must be regarded as a minimum and that 10 000 hours would be a much more desirable life to aim for. It is realized that "fault-free" is not a well-defined expression of reliability since failure is often a random effect and there is always some probability of failure. Arguing by analogy with other equipment it would be expected that an acceptable definition of "fault-free" would be "a not greater than 10% probability of failure in the time allowed", but it is likely that the exact definition would depend on the circumstances and we would expect to have to arrive at a reasonable compromise with the manufacturer in this respect.

3. Maintenance Techniques

Once an equipment has failed, its subsequent reliability is critically dependent upon the ease with which maintenance can be carried out. If an equipment is not easy to maintain then it will either be out of commission for a considerable time when it has failed, or alternatively it must be hurriedly patched up and returned to use, when it will inevitably have a shortened life in future. In the extreme, if it is very difficult to maintain, then there is a great temptation to be too busy doing something else so that the equipment is eventually discarded as useless.

To try to prevent this, adequate test facilities are suggested to allow both performance checks and fault diagnosis to be carried out. These checks should follow the logical sequence of the equipment so that it is easy to understand the relationship between the tests and the equipment. Unit construction methods are considered useful together with repair by replacement of units wherever possible in order to minimize repair time. Other points have been made about construction methods which are usually well-known but often ignored. As an example, it is well-known that it is often not possible to protect equipment cables adequately and that cable breakage may well occur but it is rare to find protection built into equipment against the damage that this might cause.

4. Handbooks and Circuit Diagrams

There is no doubt that it is impossible to build any equipment without some drawings and circuit diagrams existing somewhere and without someone being aware of how the equipment should operate. Nevertheless, equipment suppliers all too often state that this information does not exist and that handbooks are impossible to produce. Since, however, these are necessary in order to maintain the equipment, this implies that the circuits in an equipment have to be traced out by the user and a circuit diagram drawn up and a handbook produced by him. This is hardly an economic procedure and it is obvious that a handbook and correct circuit diagrams are an absolute right of the user of equipment.

5. Heavy Industrial Environments

In the early stages of the production of this Specification it became obvious that the severity of the environments in the Companies' works was not known in every case and a programme to investigate this was carried out. The resultant records showed that impressions of the severity of our works conditions were substantially correct, and that the conditions had been under- rather than over-estimated. The environments were therefore classified into severe, moderate and protected grades, and typical works areas where these environments are met have been listed. This list is to be regarded as a guide only, since it is obviously not possible to cover every site even on one individual works without considerable elaboration. The environmental conditions considered are shock and vibration, temperature, humidity, dust and fume, corrosive fume, condensation, rainfall and water spray.

6. Equipment Tests

In order to ensure that equipment supplied to a specification of this kind is in fact adequate for the expected duty, certain tests are involved. The tests do not cover any protection used to ensure safe delivery since this is assumed to be the responsibility of the manufacturer and, where environmental protection is fitted, then it is expected to be operative during the tests. Most of these tests will be seen to be similar to those specified in certain parts of K.114: Climatic and Durability Testing of Service Communication Equipment, since it was felt that many manufacturers would already have equipment to carry out the K.114 tests and that expenditure would be minimized as a result.

Most suppliers contacted seem to expect the vibration tests specified but are surprised by the drop test requirement. This has been included in the Specification to cover the large impact shocks which occur in heavy industry. These shocks are due to legitimate impacts occurring, for instance, at rolling mills and forging hammers, and to many other causes which may well be imagined. Most buildings readily propagate these shocks and the cumulative effects of the accelerations thus produced can appreciably shorten the life of an equipment.

The temperature tests are straightforward and the temperatures specified are realistic ones. It will be observed that the severe environment rules out many commonly used components and also rules out germanium transistors unless some form of cooling is used. However, cooling is difficult to apply since it must be considerably more reliable than the equipment it protects if it is not to cause failure rather than prevent it. In general it would appear better to design the equipment to operate at the specified temperature and to offset extra expense thus incurred against the saving on the cooling plant which is not required.

Certain areas of heavy industrial plants, notably ore crushing and sintering plants and foundries, are extremely dusty. This dust consists of many materials but may well be abrasive, semiconducting, magnetic, or metallic, and any combination of these properties may occur. The Specification therefore recommends a mixture of dusts to cover the range and simply proposes an upper limit to the particle size. The test has been titled "Dust and Fume" to emphasize that sub-micron particles should be present in appreciable quantities, although no detailed size analysis is specified. The aim of the test is to find out whether dust can build up in the equipment in dangerous amounts and, where equipments are ventilated, whether the air filters are adequate for the conditions of use.

Corrosive fume is met in many heavy industrial conditions near pickling plants, plating baths and furnaces for example. The most usual corrosive agent is sulphur dioxide but provision has been made to specify other materials where necessary.

The tests for condensation, rainfall and water spray are obvious. It should be realized that even where equipment does not have to withstand rainfall and water spray at the time of its installation, nevertheless it is an advantage for it to be designed to be drip-proof since roofs are notoriously unreliable in industry.

7. Equipment for Particular Environments

Examples have been given of the types of equipment which have given satisfactory service in particular environments. It will be seen that only in the protected environment is it expected that faults will be repaired *in situ*. It is felt that considerable damage can be caused to equipment in more severe environments whilst repairs are being carried out and therefore, in order to minimize the time for which equipment is opened up, it is recommended that repair be carried out by replacement of units which are themselves repaired in a protected environment.

8. Electrical Mains Supply

This aspect of the environment has been dealt with separately. The user should try to make available a mains supply which is held within reasonable tolerances but the manufacturer cannot expect to ask that mains variation is eliminated and the Specification therefore asks that permissible tolerances should be stated for both long and short term variations.

9. Mode of Usage of the Specification

It is proposed to send out a copy of the Specification with each invitation to tender and with each order where there has been no invitation to tender. In order to complete the Specification, the person originating this must decide for himself the relevant environmental conditions and specify them. The Specification will thus ensure that the major environmental conditions are specified and it is left to the discretion of the specifier whether additional requirements must be included.

The supplier may then either propose an equipment which meets the Specification or may propose a modification to the Specification which he thinks justifiable. Such modifications must obviously be agreed between the two parties and, when a final Specification has been agreed, this will form a part of the contract.

10. Suppliers' Reactions to the Specification

The Specification has been discussed with about a dozen suppliers, two of whom were invited to make specific detailed comments. The general reaction has been one of approval although it has been pointed out that the cost of equipment may well increase if it is designed to this Specification. It is felt, however, that the increased cost will be less than the cost of the maintenance required on, and lost production caused by cheaper and less reliable equipment.

Some suppliers doubt whether they can build equipment to meet this Specification and have suggested that it may be that it is not practical to do so. We doubt whether this will prove to be the case but, even if it is so, we shall have made a step forward in establishing this melancholy fact and will be able to plan our maintenance requirements on a more realistic basis.

Finally, some suppliers have been disturbed by the thought that the tests could damage the equipment they have produced. We think that this shows a wrong attitude to the problem. The equipment looked for must pass these tests with flying colours, rather than pass them only to require extensive repairs before it can be used. The Specification is a minimum requirement, not a maximum, and if there is any danger of an equipment failing these tests then its design is not adequate.

11. Conclusions

A Specification has been produced which should be of benefit both to heavy industry and to the suppliers of electronic equipment in improving the reliability of such equipment. Fully to implement the ideas behind this Specification may take some time due to the lack of formal instruction in design principles for improved reliability.

12. Acknowledgments

This Specification has been prepared by the Automatic Control Applications Committee of the United Steel Companies, Ltd., of which the author is a member. The author is indebted to Dr. F. H. Saniter, Director of the Research and Development Department for permission to publish this paper.

13. Appendix:
General Specification of Electronic Equipment for use by The United Steel Companies Limited

Introduction

This specification is designed to ensure that electronic equipment used in the United Steel Companies Limited shall be capable of withstanding for an adequate period the conditions generally encountered in the Company. Part 1 of the Specification covers general design features which are required for all electronic equipment and Part 2 covers the physical conditions which are experienced in various places throughout the Company. The design of the equipment shall be such as to ensure an adequate life and to satisfy all Parts of the specification. The specification should be regarded and used as a guide both by the Company and by its suppliers. It should be realized that there may well be difficulties in its application which shall be settled by agreement between the parties involved.

When using this specification, the environmental conditions of Part 2 (A) shall be specified in "Severe", "Moderate" or "Protected" grade as appropriate, the tests being specified by Section 2 (B), except in (v) Corrosive Fume, which will require defining. If any test is to be omitted, this shall be stated and in the case of complex equipments where it may be that some units are in different locations from the remainder, the specification shall be repeated for each location.

1. General Design Requirements

Life of Equipment

In general, electronic equipment in the Company is working in conjunction with plant of high standing costs/hour. Hence reliability is the essential requirement, fault-free lives of 1000 hours being regarded as a minimum and 10 000 hours a desirable aim. In order to reduce possible down-time still further, the time required for fault diagnosis and correction shall be reduced to a minimum. To achieve this the following requirements also shall be fulfilled:

Unit Construction

Wherever possible equipment shall be designed as an assembly of units, each unit being designed to perform one complete electrical function or group of functions. These units shall be quickly and easily replaceable by one man without the use of special tools. Where this cannot be achieved consideration should be given to duplication of units and the provision of switching.

Test Facilities

Adequate test facilities both for fault diagnosis and for performance checking shall be incorporated. Where possible the test procedure shall follow the logical sequence of the equipment and it is a prime requirement that failure of the test facilities shall not cause a failure or inaccuracy of the main equipment.

Replacement Units

(i) When a fault has been diagnosed, quick repair shall be possible by replacement of units where spare units are available. Replacement of units shall not require interference with permanent wiring or other mechanical structures.

(ii) It shall be possible to repair and test replacement units where appropriate without interrupting the operation of the main unit.

(iii) In order to minimize repair time, relays shall be plug-in types suitably protected to ensure adequate life. Where the operational state of the relay cannot be determined visually, provision shall be made for its indication.

(iv) Plugs and sockets shall be coded so that incorrect connection is not possible. They shall be equipped with adequate means of retention and have attached to them a cover to prevent ingress of dust and moisture when disconnected.

Fuses

All special internal supplies shall be separately fused by cartridge-type fuses. All power supplies passing through external cables shall be separately fused.

Handbooks and Circuit Diagrams

(i) A summary of the proposed equipment, its purpose, major units, brief technical description, etc., together with provisional circuit diagrams and unit interconnecting cables shall be given as soon as possible after the order has been placed.

(ii) In any case, before delivery of the equipment, the dimensions of all units shall be supplied together with specifications of all interconnecting cables and appropriately up-to-date circuit diagrams.

(iii) On delivery of the equipment an adequate handbook covering maintenance and operation of the equipment shall be supplied together with final correct circuit diagrams.

Fail to Safety

Where possible it shall be arranged that equipment shall fail to safety but this facility shall not allow any relaxation of the requirement not to fail. The condition of the equipment which is safe shall be specified.

Unauthorized Removal of Equipment

Components shall be mounted so that their removal is not possible by any unauthorized person. This may be achieved by providing locks on the equipment and arranging for the maintenance and assembly of such components from the interior of the equipment or by any other agreed procedure.

2. Mechanical Design Requirements

Electronic equipment in the Company may have to be operated in all types of environment from the most severe conditions of temperature, humidity, vibration, dust and fumes to fully protected "office" conditions. In general there is considerable similarity with ship-borne and vehicle-mounted conditions such as specifications K.114/A, K.114/B and K.114/E cover. The Services' general specifications of good practice such as DEF. 5000, ED.1000/R2, R.C.G.4 and C.P.1005 should be followed where possible.

No account is taken in this specification of packaging requirements to withstand transit from manufacturer to the site.

(A) Environmental Conditions and Locations

(*i*) *Mechanical Shock and Vibration*

SEVERE

Directly adjacent to any steel-making or processing plant, e.g. rolling mills.
Cranes in melting shops, forges, etc.
Locomotives.

MODERATE

Within the same building as steel-making or processing plant but at some distance from it.
Iron-making plant.
Cranes in machine shops, etc.
Vehicles other than locomotives.
Instrument panels.
Coke ovens.

PROTECTED

Control rooms.
Offices.

(*ii*) *Temperature—In addition to normal British climatic variation*

SEVERE

Directly adjacent to any iron- or steel-making or processing plant, e.g. blast furnaces, open hearth furnaces, rolling mills. Directly adjacent to high temperature metal or slag, e.g. ladles, ingots, cooling beds, etc.

MODERATE

Within the same building as above but at some distance from the source of heat.
Some control rooms.

PROTECTED

Air-conditioned control rooms.
Offices.

(*iii*) *Humidity—Normal British climatic specification except where otherwise specified*

(*iv*) *Dust and Fume—May be magnetic and/or conducting*

SEVERE
Sinter plants.
Iron foundries.
Melting shops.

MODERATE
Iron- and steel-making and processing plant.

PROTECTED
Air-conditioned control rooms.

(*v*) *Corrosive Fume*

SEVERE
Pickling plants.
Coke ovens.

MODERATE
Vicinity of furnaces.

PROTECTED
Cannot definitely be ruled out in any location but unlikely in offices. Atmospheric contamination such as is normal in heavy industrial districts.

(*vi*) *Condensation, Rainfall, Water Spray—Normal British climate if externally situated*

SEVERE
Water-cooled plant, e.g. rolling mills.

MODERATE
Sides of metal clad or unwalled buildings.
Basements.

PROTECTED
Air-conditioned control rooms.
Offices.

(B) Equipment Tests

Where equipment is supplied with environmental protection, the following tests shall be applied with the protection operative. It will be seen that these tests are similar to those specified in K.114 and where test equipment is specified in K.114, similar equipment shall be used for these tests.

(*i*) *Mechanical Shock and Vibration*

The shock test shall consist of a free drop of 1 in. on to a ½ in. thick steel plate which in turn is mounted on a concrete block. The test shall be repeated for the following number of times:

SEVERE	MODERATE	PROTECTED
50	15	6

The vibration test shall be carried out with the equipment functional, the vibration being applied mainly along the direction of the principal mounting. Each frequency range specified shall be swept through uniformly at such a rate that it requires not less than 10 minutes.

Frequency range c/s	Vibration amplitude in.	Applicable to
1–10	± 0·100	SEVERE
10–30	± 0·020	SEVERE
30–150	± 0·001	SEVERE
1–10	± 0·060	MODERATE
10–30	± 0·020	MODERATE
30–150	± 0·001	MODERATE
1–10	± 0·020	PROTECTED
10–30	± 0·005	PROTECTED
30–150	± 0·001	PROTECTED

(*ii*) *Temperature*

The equipment shall be placed in a heated enclosure and be functionally tested for 12 hours at the following temperature:

SEVERE	MODERATE	PROTECTED
75° C	50° C	40° C

(*iii*) *Humidity*

No separate humidity test is proposed but see (*vi*).

(*iv*) *Dust and Fume*

During the temperature test the equipment shall be exposed to the following dust conditions:

Dust constitution	Equal parts of silica, magnetite and iron powder by volume.
Dust size	250 mesh and below.

Duration of test shall be:

SEVERE	MODERATE	PROTECTED
30 minutes	10 minutes	Nil

It shall be examined for dust accumulation as well as being functionally tested.

(*v*) *Corrosive Fume*

The equipment shall be stored at the temperature required by the temperature test for four weeks in an atmosphere of 100% humidity charged with SO_2 (sulphur dioxide) or other specified fume in the following concentrations by volume:

SEVERE	MODERATE	PROTECTED
1%	0·2%	0·05%

The equipment shall then be examined for deterioration or corrosion of parts, finishes, materials and components and functionally tested.

(*vi*) *Condensation, Rainfall, Water Spray*

SEVERE
The equipment shall be sprayed with water at the temperature required by the temperature test simultaneously on all external surfaces.

MODERATE

The equipment shall be placed in an enclosure and heated to the temperature required by the temperature test. The humidity shall be raised to 100% and the test shall commence when this is attained. After 30 minutes in these conditions the equipment shall be allowed to cool, saturation being exceeded.

PROTECTED

As "moderate" but the humidity shall be raised to 60% only.

After the appropriate test has proceeded for 30 minutes, the equipment shall be switched on and functionally tested for 30 minutes.

3. Mechanical Construction for Particular Locations

Class (S): All conditions severe

The mechanical design shall be of a very high standard and all components hermetically sealed in protective boxes. At the design stage consideration shall be given to the duplication of components and the provision of remote switching to reduce the possibility of failure in operation. In many cases it will not be possible to obtain access to carry out maintenance except when the plant is completely shut down; reliability is therefore very important.

Class (M): All conditions moderate

This is the situation which calls for the maximum use of replaceable plug-in units for maintenance. Test procedures shall allow for the rapid identification of faulty units but need do no more.

Class (P): Fully protected conditions

This equipment will usually be in the form of centralized racks and the main requirement is for rapid identification and repair of faults. This calls for considerable care in the design of test facilities and in the accessibility of the equipment for maintenance.

Appendix A: Electrical Mains Supply

It is realized that it is not reasonable to expect equipment manufacturers to protect their equipment against the severe mains fluctuations which can occur on works where large loads employing long cables are in use. There appear to be two normal specifications for mains supplies for electronic equipment in use in the industry which are:

Normal tolerance	Nominal voltage	± 6%
	Nominal frequency	± 2%
Wide tolerance	Nominal voltage	± 10%
	Nominal frequency	± 10%

It shall be the responsibility of the equipment specifier to ensure that appropriate steps are taken if these tolerances are exceeded and if "surge" conditions exist.

It is reasonable, however, to require the equipment manufacturer to state the supply voltage tolerance acceptable for his equipment and to indicate the levels of transient and harmonic voltages which it will tolerate.

Appendix B: Specifications Quoted

K.114. Issue 2—Climatic and Durability Testing of Service Communication Equipment. Obtainable from: The Secretary, Radio Components Standardisation Committee, Ministry of Aviation, Castlewood House, 77–91, New Oxford Street, London, W.C.1.

DEF.5000. General Requirements for Service Telecommunication Equipment. Obtainable from H.M. Stationery Office.

ED.1000/R2. General Requirements for Admiralty Surface Weapons Establishment Development Work. Obtainable from: Captain Superintendent, A.S.W.E., Portsdown, Cosham, Portsmouth, Hants.

R.C.G.4. Guide to Approved Components, Valves and Insulating Materials. Obtainable from: Radio Components Standardisation Committee, as above.

C.P.1005. The Use of Electronic Valves. Obtainable from: British Standards Institution, 2 Park Street, London, W.1.

Manuscript received by the Institution on 1st March 1962 Paper No. 747).

POINTS FROM THE DISCUSSION

Mr. P. Huggins: The proper solution to an industrial electronics problem is to use equipment that is just good enough for the job at the cheapest price. If the equipment is not good enough it costs too much to maintain; if it is too good it must have been too expensive! I would like to thank Mr. Shaw for providing the industrial electronics industry with a blueprint which can be used (with suitable modification) in any factory having a special environment.

Mr. D. Shaw (*in reply*): In general I would agree with Mr. Huggins. However, there are occasions when the best equipment is not the most expensive and this must be considered. We have heard in this symposium how reliability has been increased by removing components, thereby cheapening the equipment. It can well be that the cost of developing an unreliable system is greater than that of developing a reliable system simply because the reliable system requires less "debugging" time once it has been made.

The "brilliant" engineer who can squeeze the last ounce out of a circuit is in fact the biggest enemy of reliability since, if the last ounce has been squeezed out, the circuit has no reserve left for deterioration. The truly brilliant designer is the man who can design an equipment which does its job comfortably with a reasonable margin in hand on all its parameters, environment included, using the minimum number of components.

ATOMIC ENERGY DEVELOPMENT

Colonel G. W. Raby, C.B.E.,† a Vice-President of the Institution, addressed the Parliamentary and Scientific Committee on 24th July 1962 on some of the present problems of developing a new reactor system. The following extracts from his address emphasize the cost of research and subsequent development in a field where electronics plays a major part in control and instrumentation.

"Any reactor system which is developed is a very expensive venture, whether it is the existing Magnox system or the Advanced Gas-cooled Reactor (A.G.R.) or an entirely new system, such as the Steam Generating Heavy Water reactor. When such an expensive venture is started the only basis on which it can be made to pay its way is that it must have a good run for its money after the initial development has been completed. Otherwise there will be a waste of finance and effort on an immense scale, morale will be lowered amongst the professional staff engaged on the work, and in the long run reactor development may be retarded.

"The present Magnox system can be further exploited before being superseded by a more economic system, and if this further exploitation does take place, it will enable at least some of the initial expenditure incurred on this system by both Government and Industry to be recouped. The Consortia of industrial companies engaged on building power reactors have invested at least £12M and probably more in the Magnox system on research and development alone—and this figure does not include the development costs of fuel handling equipment, or the more conventional items such as heat exchangers, etc., most of which have entailed prolonged experimental and metallurgical investigations.

"What is true of the Magnox reactor will be equally true of the A.G.R. or of any new reactor system developed, and although we are somewhat more cautious in these matters to-day, nevertheless, it is inevitable that further sums of money will have to be invested by industry when new types of reactors are ordered. And this investment can only be made if sufficient numbers of the new reactors are ordered thereafter to bring a return. A primary requirement of a new reactor system for power generation purposes is therefore that it must show sufficient economic promise to justify potential users anticipating an extended life for it. And as the largest potential users are the Generating Boards, it is to them that we must look for help and guidance in such an assessment.

"Assuming the Generating Boards agree that a system has extensive economic possibilities, then the actual process of development falls naturally into two separate stages, first the development of the basic design up to prototype stage, and second the extrapolation and commercial exploitation of the basic design thereafter.

"The first stage covers the initial reactor development and design right through to construction and commissioning of a prototype of sufficient power to provide reliable basic data. If the prototype is of, say, 100 electrical megawatts then the probable cost of design and construction will be between £35M–£40M and will require not less than 2500–3000 man-years of professional effort. The size of this effort can be gauged by the fact that it represents an appreciable slice of the country's total capacity in this particular field.

"Another point to be borne in mind is that a prototype reactor must be completed in the shortest possible time if the Generating Boards are to obtain actual performance data on which to base an extensive forward programme. For example, if the Generating Boards are to have actual operating data within, say, six years of the commencement of a new development, it means that the prototype must be completed at least one and a half years before that. Every endeavour must be made to meet such a short time-scale by planning and by co-ordination of all available resources. It is of particular importance that the facilities and resources of the Consortia are fully used in the design, construction and commissioning of prototype reactors so that the knowledge and experience gained in this early stage is available for the commercial exploitation of the reactor later on. During the whole of this first stage the primary responsibility for co-ordination and control presumably rests with the U.K. Atomic Energy Authority.

"Stage II is probably the most critical in the life of a reactor development, for it is then that a decision must first be made as to whether the basic performance data obtained from the prototype will enable the Generating Boards and the Consortia to extrapolate the design to larger outputs, and whether the system as a whole confirms the earlier promise of economic advantages.

"If the A.E.A. and the Generating Boards agree that the reactor system has possibilities for further exploitation and will meet user requirements, then presumably the primary responsibility for the necessary further design, development and exploitation in the home market will thereafter rest with the Generating Boards.

"The development of a new reactor system is thus an expensive venture, and it must be developed with a sense of urgency. Above all, it must be developed economically by making full use of all resources on a true partnership basis between the Generating Boards, the A.E.A. and the Consortia with clear divisions of responsibility to ensure the successful outcome of these developments. Such an arrangement would at least enable us to go forward in the reactor development field on a united front, and this in itself would give a much-needed boost to the sale of British reactors abroad."

The following speakers also addressed the Committee:

Sir William Cook, C.B., United Kingdom Atomic Energy Authority.

Mr. S. A. Ghalib, B.Sc.(Eng.), M.I.E.E., M.I.Mech.E., Deputy General Manager, The Nuclear Power Group.

The discussion was opened by Sir Christopher Hinton, K.B.E., F.R.S., Chairman, Central Electricity Generating Board.

† Managing Director, Atomic Power Constructions Ltd.

U.D.C. 621.391.88 : 621.394/5 : 681.39

Error Correction in Data Transmission Systems

Presented at the Symposium on "Data Transmission" in London on 3rd January 1962.

By

E. R. AYLOTT, B.Sc.(Eng.)†

AND

E. S. SIMMONDS†

Summary: The use of the existing telephone and telegraph networks as data transmission media is discussed with special reference to the correction of errors. These networks as they stand are far from ideal for data transmission because the high-level noise pulses and short interruptions which are relatively unimportant for normal usage cause errors in data messages. The data terminal equipment must contain some means of error correction, and various ways in which this can be done are explained. Typical error detection and correction equipments for both the telephone and telex networks are shown. Finally methods for the evaluation of the relative efficiencies of the various error correction and detection codes are described.

1. Introduction

The present-day trend, especially in large organizations, is to make greater use of electronic computers, or data processing equipment, to carry out much of the routine office work normally associated with business management. Examples of this that readily spring to mind include centralized stock control, centralization of banking, accounting and payroll, and airline seat reservations. In many cases the computer and the information source and destination or both, may be many miles apart; possibly in different countries or even in different continents. It is necessary, therefore, to be able to transmit information, in a suitable form, to and from the computer centre. For many applications the more normal methods of sending information, for example the postal service or courier, are not fast enough. The need has thus arisen for a new form of system for the fast and accurate transmission of information, or data, between machines.

The language of a computer is usually in the form of characters comprising a number of elements; these elements may be in either one of two states, and are referred to as binary digits. In this respect the machine language is similar to the language of the teleprinter and telegraph network. Because of the widespread availability of the existing telegraph and telephone networks, consideration has thus been directed to the use of these for the transmission of data.

The circuits to be used for data transmission may be either leased circuits on a permanent point-to-point basis, or circuits within the public switched telephone or telegraph networks, dialled up when required. The choice of circuit will be dictated, to a large extent, by the volume of data to be transmitted, by time limitations imposed, and by the relative line rentals. It may be expected that with the expansion of subscriber trunk dialling the use of dialled connections over the telephone network will become increasingly attractive for data transmission at speeds in the order of 1200 bauds, and possibly higher, whilst the telex network already provides circuits which may be used for data transmission for speeds up to about 60 bauds.

2. Limitations of Existing Networks

The existing telephone network was designed for the transmission of speech where the main requirements are those of bandwidth and relative background noise level. Noise clicks and very short interruptions, as may occur from time to time on a telephone connection, are relatively unimportant, because, firstly the ear can tolerate them, and secondly, because there is a large percentage of redundancy present in speech which enables the listener to apply automatic error correction to a certain extent; and in any case he is able to ask for an immediate repeat of uncertain information.

The main requirement for most data transmission links is that of error free transmission of data in the form of binary digits. A bit error rate better than 1 in 10^7 is generally considered to be necessary. The noise and interruptions that occur in a practical circuit can cause errors in a data message, but, because the size of the existing networks makes it inconceivable that they be modified to suit the special requirements of data transmission, the data transmission terminal equipment must incorporate means to eliminate errors as far as possible.

The main sources of high level noise pulses on a telephone connection are: induced currents in the cable caused by dialling impulses in adjacent cable

† Standard Telephones and Cables Ltd., Transmission Division, North Woolwich, London, E.16.

pairs, and current variations caused by vibration of the dry contacts that are present in the switches and distribution frames within the telephone exchanges. On a switched connection the selector contacts are the major source of noise, as their contact resistance is likely to fluctuate due to vibration when adjacent selectors or relays operate.

Short interruptions can be caused during maintenance operations on the distribution frames by the

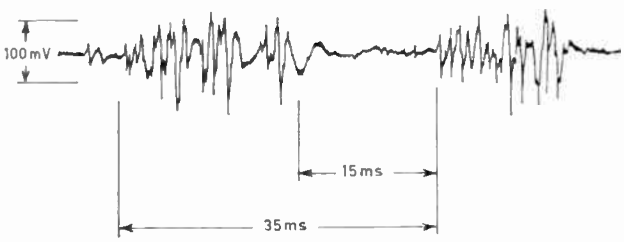


Fig. 1. Typical noise burst on a telephone circuit.

disturbance of the dry contacts such as are present in the fuse and protector mountings. These disturbances can affect some rented circuits even more than switched circuits because over many rented circuits only low-level voice-frequency signals are transmitted, the potential of which is insufficient to wet the connections. Other causes of short breaks in transmission are changeovers of carrier and power supplies on wide-band multi-channel carrier circuits. As previously stated, these noise pulses and interruptions are generally insignificant in a telephone conversation, but because the duration of a data signal element is only in the order of one millisecond, they may extend over several signal elements of a data message, causing many, or all, of these to be in error.

The oscillogram of a typical noise burst occurring on a telephone circuit involving two-motion selectors is shown in Fig. 1. It is seen that such disturbances may have amplitudes exceeding 100 mV, and last for several hundred milliseconds.

Figure 2 shows the cumulative frequency distribution of such noise pulses versus the response level of the detector. These measurements were made on calls established by dialling on a telephone, and then replacing the telephone with suitable terminations. The number of noise spikes exceeding a given response level were then counted. The left-hand ordinate shows the frequency with which noise spikes occur, and the right-hand ordinate shows the faulty bit rate probability assuming a transmission speed of 1000 bauds.

The effect of the relative background noise level, fluctuations of signal level and the like on the error rate are dependent to a large extent on the design of the modulation and demodulation equipment. It is generally accepted that frequency- and phase-modulated systems are to be preferred from this point of view, but it should be emphasized that whatever modulation system is employed, it cannot cater for the high-level noise pulses and interruptions described above.

3. Error Detection and Correction of Data over Telephone Circuits

It can be seen that, generally, for the satisfactory transmission of data, where human checking is impossible, and where there is insufficient redundancy in the information to permit error detection or correction within the processing equipment, some means of error correction is necessary.

In certain applications of data transmission systems, for example telemetering, the information is continually being brought up to date by later information. In these cases there is no value in attempting to correct errors if retransmission is involved.

The types of circuit available for data transmission may be classified as simplex, duplex, or half duplex. In the simplex circuit transmission can take place in one direction only. The duplex system allows for transmission in both directions simultaneously, although not necessarily at the same transmission speed in bauds. The half duplex circuit permits transmission in either direction, but not in both directions at the same time.

On most two-wire leased or switched circuits duplex operation can be readily obtained by operating in the two directions over separate parts of the available frequency band. However, on circuits where echo suppressors are present, steps must be taken to disable the echo suppressors to enable duplex working to take place. This can be done with suitably designed or adapted echo suppressors, which react in the presence

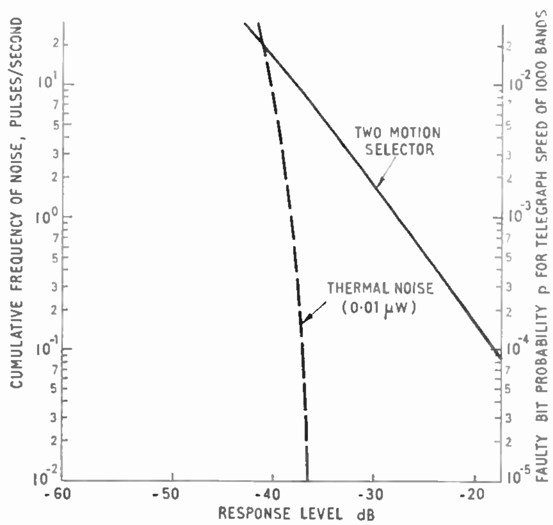


Fig. 2. Cumulative frequency distribution of noise pulses.

of continuous tone to switch themselves into an inoperative condition. If no such steps are taken only half duplex working is possible and the lost time due to the "turn round" time of the suppressors becomes considerable.

Error correction methods where correction is by retransmission can be used on the duplex and half duplex circuits only. Simplex circuits are of greatest value when error correction is not required; otherwise, when simplex circuits are used information for the correction of errors must be present in the transmitted data.

3.1. *E.D.C. Methods*

Five schemes for error correction are shown diagrammatically in Fig. 3. The first three make use of redundancy in the form of check, or parity, bits derived from the information signals.

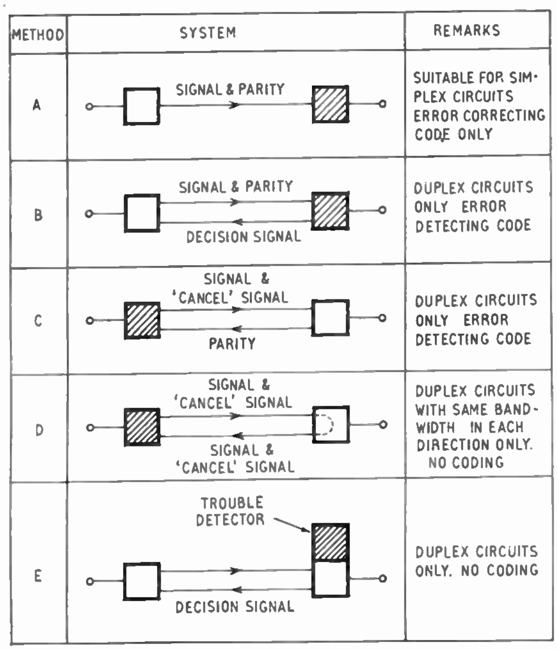


Fig. 3. Error correction schemes.

The first method (A) uses a forward-correcting code in which a number of parity checks are derived from the digits of each character, and these are transmitted with the information bits in predetermined positions. At the receiving terminal the parity bits are used to check the information, and, if necessary, to correct errors. The protection gained depends upon the amount of redundancy that is added. In practice with such codes the percentage of redundancy necessary to provide a substantial reduction in the error rate of data signals transmitted over telephone circuits is so high as to reduce seriously the effective transmission speed, even under error free conditions. These codes have been fully described elsewhere.†

For a given error rate, much less redundancy is required when error detection only is employed. Correction, in these cases, is by retransmission of the orginal information.

In the arrangement of scheme B redundancy is transmitted together with the information, and error detection takes place in the receiving equipment. Usually the information is split up into blocks each containing a fixed number of bits to which redundancy, i.e. parity bits, is added. A block may be as small as one character or, on the other hand, it may contain several hundred bits. After each block has been received a signal is sent back to the transmitting equipment, via the return path, indicating "correct" or "repeat".

Scheme C is an alternative arrangement in which redundancy is not transmitted with the information. Parity bits are derived in the receiving equipment. and are sent back to the transmitting station, where they are compared with similarly derived parity bits. Should an error be indicated the faulty block is retransmitted together with a prefix or cancel signal to indicate a repeat.

This method usually requires a wider bandwidth in the return direction than that in the previous scheme B. When working over a full duplex circuit the relative bandwidths of the forward and backward channels will influence the size of block and percentage redundancy that may be used.

An extension of method C is shown in scheme D. In this arrangement the whole of the transmitted message is repeated back to the transmitting equipment, where a direct comparison with the original signal is made. An error will cause a cancel signal. referring to a predetermined number of characters. to be transmitted followed by a repeat of the original signal.

The fifth scheme E does not make use of any redundancy, and the signal itself is not compared or examined for errors. An analogue detector is used at the receiving terminal to indicate the presence of interference which is likely to cause an error. The detector could be applied to a parallel channel conveying no information, but responding to similar noise and interruptions, or it may take the form of a device indicating excessive telegraph distortion of the received signals, etc. Whatever the form of the detector, whenever the appropriate output occurs, a signal requesting a repeat is sent back to the transmitting equipment, regardless of whether or not the information received is correct.

† R. W. Hamming, "Error detecting and error correcting codes", *Bell Syst. Tech. J.*, **26**, No. 2, pp. 147–60, April 1950.

A well-known example of scheme B is the Van Duuren ARQ system used in radiotelegraphy.† The five-unit telegraph signals are converted into a seven-unit code with a fixed ratio of marks and spaces. The receiving equipment checks each character for this correct ratio and when an error is noted it requests a repeat from the transmitting equipment.

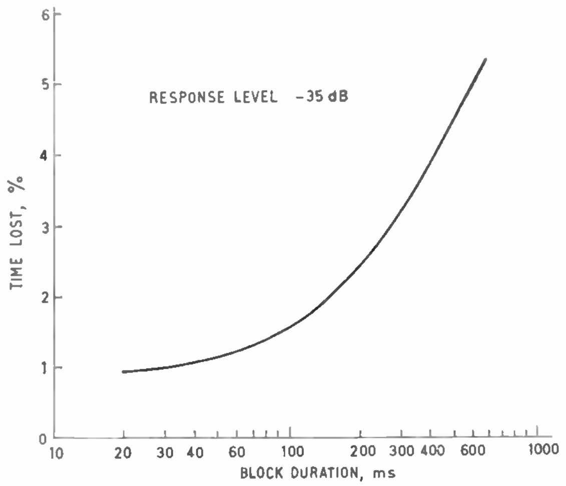


Fig. 4. Time lost as a function of the block duration.

Of these five schemes, schemes B, C, and D offer the greatest protection against errors; scheme D, however, can be operated only when the channels in each direction are of similar bandwidth. When two-wire circuits are used, the forward channel is generally made a wider bandwidth than the return channel, so that the maximum transmission speed may be realized, and because of this limitation the choice lies between schemes B and C. Both these schemes operate on a block by block basis of transmission with correction by retransmission, and stemming from this, the ratio of the number of blocks containing errors to the total number of blocks transmitted is obviously of interest.

3.2. *Optimum Block Duration*

As the length of a block is increased, the probability of an error occurring within it is increased. The time for retransmission is also increased; and so the effective transmission rate is decreased. If the block length is too short the redundancy forms a greater percentage of the total message so that again the effective transmission rate is decreased. Some compromise is thus necessary. Figure 4 shows the percentage time lost due to the retransmission of blocks containing errors, versus the block length, with noise as in Fig. 2 and a detector response level of −35 dB. From this curve it can be seen that the block length may be increased up to several hundred milliseconds before the increase in time lost becomes substantial.

3.3. *Modes of Operation of E.D.C. System*

There are two possible modes of operation involving block-by-block transmission. In the first mode, transmission is stopped after each information block until a signal is received back from the receiving equipment, indicating acceptance or rejection. The time lost between blocks is equal to the total loop propagation time plus the duration of the decision signal plus the delay time within the apparatus. The block duration should be made relatively large, so that this lost time is small compared with the total transmission time.

Because of the lowering of the effective transmission speed due to the lost time between blocks, even under error-free conditions, this arrangement is usually reserved for applications where half duplex operation only is possible.

A preferred arrangement when full duplex circuits are available is to transmit successive blocks of information without any break. During the transmission of the second block, the decision signal referring to the first block is received at the transmitting terminal. If this is the signal "correct", block three is transmitted after block two; but if the decision signal requests repetition, block 1 is retransmitted after block 2, followed again by block 2, and so on. In this system the minimum block duration is equal to the total loop propagation time plus the computation time plus the duration of the decision signal. If the return channel is of a narrower bandwidth than the forward channel, it must be remembered that the propagation time in the return direction may be greater than that in the forward direction.

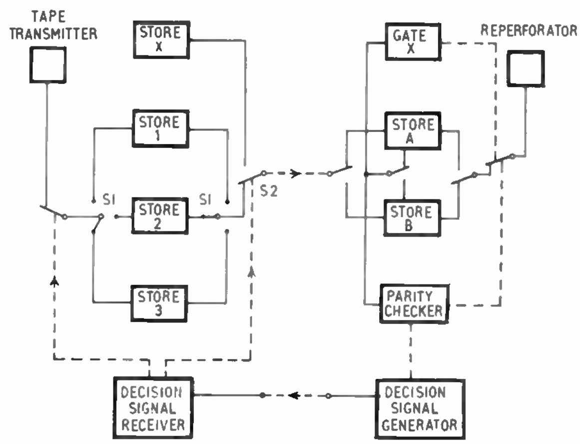


Fig. 5. Block layout of an error detection and correction equipment for use on the telephone network.

† H. C. A. Van Duuren, "Teleprinting over radio links", *Tijdschr. Ned. Radiogenoot*, 16, pp. 53–67, March 1951.

In both of the modes of operation just described the operation has followed that of scheme B in Fig. 3, in which the error detection is carried out in the receiving equipment. However, the same modes of operation are applicable when the redundancy is derived at the receive end and transmitted back, as in scheme C, or when an analogue detector is used, as in scheme D.

3.4. *Typical E.D.C. Equipment*

An error detection and correction system (e.d.c.), based on the method B shown in Fig. 3, is shown diagrammatically in Fig. 5. The input to the equipment is derived from a tape reader, and the output operates a tape reperforator.

The transmitting equipment contains three temporary stores which are connected in turn to the output of the tape reader and to the input of the data modem equipment. When the data representing one block have been read into a store, parity check bits are added before a read-out takes place.

In the particular arrangement shown, the input is read into store 3 whilst the output from store 2 is being transmitted. During this time a signal is received back from the distant end equipment indicating whether or not block 1 is acceptable. If block 1 is correct the data from store 2 are followed by those of store 3, and new data are fed into store 1.

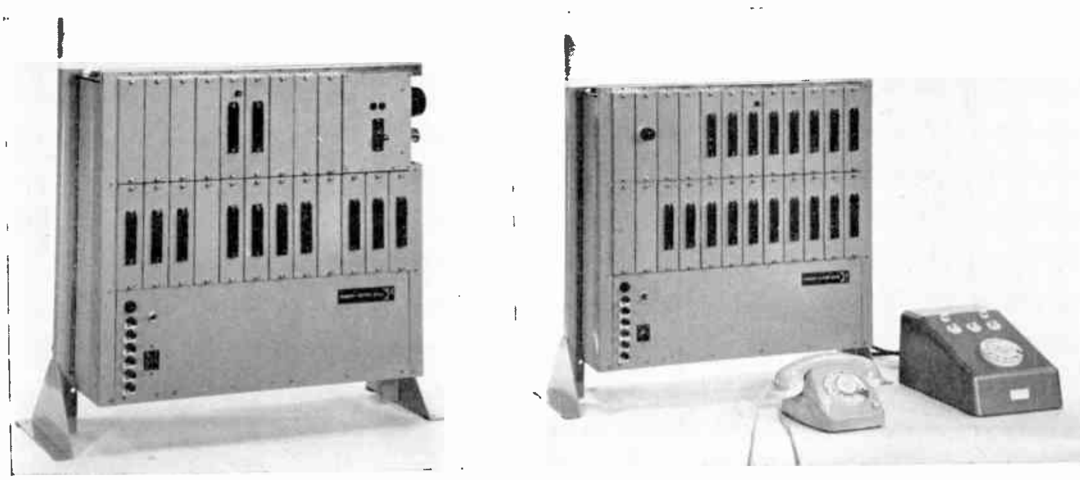
If the signal received back is other than "correct", the next information to be sent is that from a permanent store X, followed by a repeat of blocks 1 and 2, during which time the tape-reader is stopped. Provided that no further errors occur, block 2 is then followed by block 3, and so on. Store X contains a signal which indicates to the receiving equipment that a repeat is to take place, and which also maintains synchronism.

The forward signal indicating that the next blocks are repeats is necessary because the repeat may have been initiated by an error in the supervisory channel, and not by a signal from the receive equipment.

The receive equipment contains two stores which alternately read in and read out. Whilst data are being read into one store, e.g. store A, from the line equipment, they are checked for parity in the parity check equipment, and also for coincidence with the repeat signal in gate X. If this checking indicates that the data are satisfactory, these are then read out into the reperforator, whilst the next block of data is received into the other store, B. If the data contain an error, a signal is sent back to the transmitter requesting a repeat, and the output from the store is blocked. The output from the store is also blocked in the event of the "repeat" signal being received.

To ensure that the data read out to the reperforator are in the correct block sequence, the blocks are serially numbered from 1 to 3. There is a means within the receiver for identifying the block number and comparing it with the next number expected, which, in turn, is advanced each time a correct block is received. This is important, especially when a repeat is initiated by an error in the supervisory channel, because in this case some of the repeated data will already have been read out into the reperforator.

Photographs of the transmitting and receiving equipment are shown in Fig. 6. In the experimental models of the equipment described a block length of 63 bits was used which at 800 bauds gives a block duration of about 79 milliseconds.



TRANSMITTER RECEIVER

Fig. 6. Experimental data transmission system with automatic error correction by block retransmissions.

4. E.D.C. for Data Transmission on Telegraph Circuits

In a similar manner to its application to telephone circuits, e.d.c. may be applied to telegraph lines, whether these be private wire or switched as in the telex networks. When used in this latter mode, the e.d.c. equipment must be designed so that normal telex operations may be carried out, and the e.d.c. equipment only brought into the circuit as required after the line connection has been established. Thus the most flexible arrangement at the subscriber's terminal would be for the e.d.c. apparatus to be a separate unit and to interconnect with a standard signalling unit by means of plugs and sockets.

4.1. *Data Transmission on Telex Circuits*

In operation, one telex subscriber should be able to dial another and converse in the normal telex manner. If data are then required to be sent accurately, the e.d.c. equipment could be switched in at the sending end, either by manual or automatic means. This action must cause instructions to be sent in the normal telegraph code (C.C.I.T.T. Alphabet No. 2) to the called subscriber's apparatus such that it will automatically "switch to data", since the position at that time may be unattended. Also, to avoid interference with the data elements, this signal may be used to disable such automatic devices as will respond to telex sequence signals, otherwise it would be necessary to restrict the data alphabet. Whilst the use of this "switch to data" signal will preclude the use of the devices mentioned it does not preclude the return to printing should the data users concerned agree to provide the necessary signal. When transmission of data is over, the breakdown of the connection can be attained by a clearing signal sent during the data condition.

The power on the e.d.c. equipment would not need to be on the whole time, but could be switched on by the signalling unit associated with the telex position, in a similar manner to the supply of power to the teleprinter motor when the position is in use.

4.2. *E.D.C. Methods*

The type of e.d.c. that may be employed depends on what type of circuit is available, i.e. half or full duplex. The former must inherently give a slower overall signalling rate than the latter, since transmission must stop in order that the decision signal may be received in the backward direction. This delay inevitably causes a loss of time equal to at least twice the line delay plus the switching time. It has been assumed that the half duplex circuit must use an error detecting code since, although the speed of working is the same in both directions, it would take too long to repeat all the information back to the sender. In the full duplex case, however, this is not so, it being possible to transmit both ways simultaneously with only twice the line delay to take account of. As the transmitted information has to be stored for retransmission purposes, a simple extension of the store will cater for the maximum line delay to be encountered.

In considering the capital outlay required to give error detection and correction, the e.d.c. arrangement on a full duplex circuit with information feedback would appear to be attractive. The e.d.c. equipment could be self-contained in a unit which would mount on a desk, and only need be associated with the transmitting terminal for the transmission of data in one direction. Since, with information feedback, only the information bits are passed to a store and to line, the decision whether to repeat or not can be achieved by straight comparison, and therefore no encoding is necessary.

4.3. *Overall E.D.C. Arrangement*

A block diagram of a typical overall low speed e.d.c. arrangement with information feedback is shown in Fig. 7. Either station A or station B can instigate the transmission of data which, in the con-

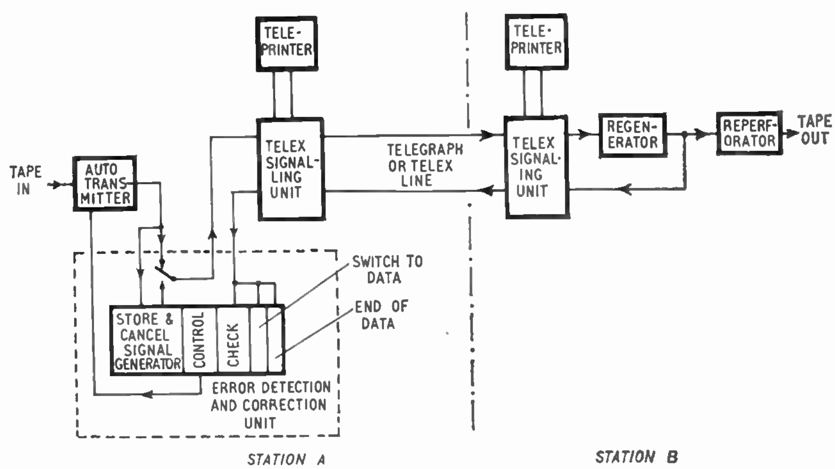


Fig. 7. Typical arrangement of e.d.c. on a full duplex telegraph circuit.

figuration shown, can only be sent from A to B. To transmit in either direction station B would also have to be equipped with an e.d.c. unit. It would then be necessary to fit a send/receive switch into the e.d.c. unit in order that the e.d.c. unit at the receiving end could be turned into a regenerator, as shown in station B. This send/receive switch could then be operated either by manual or automatic means in a similar manner to that of the "switch to data" and "end of data" arrangements.

5. Evaluation of the Relative Efficiency of Error Correcting and Error Detecting Codes

Many different error detection codes have been proposed from time to time and a means of comparison of their relative efficiencies is required before a decision can be made as to the best code for use in a data transmission system. Any method of comparison must take into account the "fine structure" of the errors within a block which are encountered on a particular network and with a particular mode of operation.

One obvious method of comparison is by the direct measurement of detected and undetected errors obtained by various equipments, designed to operate in different modes, and with different codes operating over practical circuits. To obtain a reliable comparison, measurements must be made for many hours on each of many telephone connections. This is a laborious and time-consuming exercise, and as such must be considered unacceptable.

Other methods make use of a statistical analysis of the error or fault occurrence obtained with a particular mode of operation when tested over a number of telephone connections. Once this statistical analysis has been obtained from tests over typical circuits, no further on-line tests are necessary.

The error analysis may be either a complete record of the position of all faults, or a simplified record showing the totals of a number of predetermined parameters. Using the simplified record a useful comparison is obtained simply, but it does not allow for the "fine structure" of the error pattern.

The simplified fault record is arranged to show, for a given mode of transmission, the proportion of element errors which occur as 1, 2 or 3, etc., within a block.

From the properties of the code under examination, a number of error reduction factors are calculated. Each factor indicates, for a given number of errors occurring within a block, the proportion of errors that would remain undetected, assuming that all possible patterns of such errors were equally likely to occur.

From these two sets of figures the fraction of errors that would remain undetected for all values of errors per block may be obtained. Because of the assumption made in the calculation of the reduction factors, this method of comparison does not give a complete evaluation; it is nevertheless useful as a guide and has the virtue of being relatively easy to calculate. It does not take account of the fine structure of the distribution of errors, and tests have shown that for some codes the undetected error rate is under-estimated, and for others it is over-estimated.

To overcome the shortcomings of the simplified evaluation, comparisons may be carried out using a more comprehensive recording of errors. This recording of errors should be fully representative of the pattern of errors occurring with a particular mode of operation, and should be based on the results of many tests. The information may be presented as an error sequence recording which, for convenience, may be compressed by the omission of some of the long error-free periods. Any compression must be carried out with care, so that the gaps between errors are not made so small as to affect the results; and, in addition, the degree of compression must be known. Error codes may be applied directly to these error recordings for evaluation.

An alternative method is to present the error information in a statistical form as the distribution of the separation between errors. Using this information, a computer may be programmed to generate a series of errors conforming to the correct separation characteristic, and to apply to this any particular error-detecting code that one may wish to test. This permits evaluation tests of the code to be carried out at the computer speed for long periods. With this arrangement a whole family of codes may be examined at the same time.

Because the fault record used in this method is fully representative of actual conditions, the reduction factors calculated by it take account of the "fine structure" of the error distribution, and this gives a complete evaluation.

These methods, which enable the relative efficiencies of various error detecting codes to be made quickly on an "off-line" basis, make it possible to decide upon the best code to use before designing an e.d.c. equipment.

From such evaluation it is apparent that the most efficient codes derived so far are those using a regular or scrambled interleaved parity arrangement.

6. Conclusions

When it is required accurately to transmit binary coded data from one point to another, the telephone and telegraph networks, though far from ideal, provide suitable and very convenient transmission media. However, in order to overcome the shortcomings of these networks, as far as data transmission is concerned, some form of error detection and correction is essential.

Except in special cases forward error correcting codes cannot provide sufficient accuracy, and a system of correction by retransmission is required. In the case of the telephone network, transmission is most efficiently carried out on a block-by-block basis, with parity bits derived for each block. The parity bits may be transmitted together with the information, or they may be derived at the receiving terminal and transmitted back to the transmitting equipment; the method chosen depends, to a large extent, upon the type of telephone circuit and modem equipment available.

In the case of networks where full duplex circuits are available, error detection can be carried out simply in the transmitting equipment by comparison of the signal repeated back from the receiver with the original signal.

The efficiencies of various error-detecting codes may be evaluated using a computer once sufficient statistical information regarding the circuits and mode of operation has been obtained.

7. Acknowledgment

The authors wish to thank Standard Telephones and Cables Limited for permission to publish the information contained in this paper.

Manuscript first received by the Institution on 8th December 1961 and in final form on 28th April 1962. *(Paper No. 748/C44).*

The Symposium on "Data Transmission"

A Symposium of seven papers on Data Transmission, sponsored jointly by the Programme and Papers Committee and the Computer Group Committee, was held at the London School of Hygiene and Tropical Medicine on 3rd January, 1962. Some 150 engineers were present to hear and discuss the papers which were arranged in two sessions—Line Considerations and Terminal Equipment. The Chair was taken by Dr. G. L. Hamburger (Member), a member and past chairman of the Programme and Papers Committee, who opened the meeting with some remarks on the growing importance of data transmission as a means of fully utilizing computing facilities. The papers presented and their dates of publication in the *Journal* are as follows:

"Telephone circuit evaluation for data transmission"—K. L. Smith, J. Bowen and L. A. Joyce. (*September*)

"Considerations in the choice of the optimum data transmission systems for use over telephone circuits"—A. P. Clark. (*May*)

"A comparison of the merits of phase and frequency modulation for medium-speed serial binary digital data transmission over telephone lines"—F. G. Jenks and D. C. Hannon. (*July*)

"Data collection and distribution"—D. J. Dace. (*May*)

"Data collection systems, their application and design"—J. A. Pearce. (*October*)

"Error correction in data transmission systems"—E. R. Aylott and E. S. Simmonds. (*August*)

"Some recent developments in data transmission"—K. L. Smith. (*October*)

Discussions of the papers took place after each session and many points of interest were raised. Some of the principal contributions to the discussion are given below:

DISCUSSION

Mr. M. B. Williams: Although data transmission over private point-to-point circuits should be subject to the minimum of technical restrictions, if data transmission over the public telephone network is to develop rapidly and freely it is desirable that a few essential parameters should be standardized. The British Post Office has joined with other telephone administrations in studying the problem of data transmission over national and international public telephone connections, and a recent meeting of an international study group (International Telephone and Telegraph Consultative Committee Special Study Group A) had agreed on certain recommended characteristics for serial f.m. and p.m. systems, namely, modulation rates, carrier frequencies and deviation. The preferred modulation rates for transmission over telephone connections were 600 bauds (which should be capable of unlimited application) and 1200 bauds (on favourable connections). Other modulation rates proposed for further study were 1800, 2000 or 2400 and 3000 bauds. It is hoped that future designs of data processing equipment would be guided towards these nominal rates.

A further useful standardization should result from the concept of an "interface" at the boundary between data transmission equipment and data processing equipment. At the interface are defined the interconnecting wires which carry binary data and control signals.

There is a considerable possibility of confusion in the words "synchronous" and "asynchronous" being used in a particular sense by some of the authors to denote that the modulation rate bore a simple numerical relationship with the transmitted carrier frequency of p.m. systems. I suggest that it would be better to refer to the coherence between the element rate and the carrier frequency.

Mr. W. F. S. Chittleburgh: I would like to draw attention to the recent meeting of the C.C.I.T.T. Special Study Group A referred to by Mr. Williams at which considerable information was contributed on test results of data transmission over practical circuits. From a study of these results, the meeting concluded that the test results did not enable a definite recommendation to be made for a particular type of modulation. The opinion was also expressed that the various performances would appear to depend as much on the individual design of equipment as on the modulation technique employed. This seems somewhat surprising, considering the outstanding theoretical superiority that has been claimed for p.m.

I endorse the statement that it is not so much the superiority of one system measured in decibels which is so interesting as the actual error rates of the two systems. However, I would go further and say that a user is interested in the relative error rates, when measured over practical circuits, and I still submit that such evidence, collected from diverse sources, shows no marked superiority for either method of modulation.

Mr. W. D. Worthy: We are indebted to Messrs. Smith, Clark and Jenks for their very convincing demonstrations of the potential superiority of phase modulation. It is now clear that this gives the maximum possible change from $+1$ to -1 in the autocorrelation of the signal—auto-correlation in the sense that the signal is multiplied by itself with a delay equal to the minimum allowable interval between elements of data and integrated over the same period.

However, in principle the determination of the autocorrelation involves the process of multiplication which it is difficult to perform electronically. I would like to ask Mr. Smith to show how this is accomplished. I would also like to ask exactly why the data for transmission is filtered before being fed into his "linear phase modulator". Phase modulation of the carrier by simple switching with a square wave would seem to be a more straightforward process, followed of course, by a low-pass filter before transmission.

Mr. A. C. Croisdale: The Post Office will provide, on hire, the f.m. modulator/demodulator described by Mr. Williams for use on telephone circuits. It is hoped that this modulator will be available in a year to 18 months. This will involve the interface, described by Mr. Smith, between the P.O.-supplied modulator and the data processing equipment. The details of this interface have in fact been accepted by the U.K. at the October 1961 C.C.I.T.T. meeting on Data Transmission. The U.K. delegates included the Electronic Engineering Association which also represented the computer industry. Standardization of these details will be of value in many ways, particularly as a means of gaining comparative experience of data transmission systems as soon as possible.

It is possible that, after further study the C.C.I.T.T. will recommend a further standard modulator based on phase modulation. It seems likely to offer a higher signalling speed in the same bandwidth compared with f.m. The comparative tests of phase and frequency modulation quoted by Mr. Jenks and Mr. Smith were made in the laboratory. However, most of the mutilations which cause errors in data transmission on lines are caused by brief disconnections and these would cause errors in both p.m. and f.m. Similarly, a percentage of the noise bursts which cause other errors are due to high levels of interference which would cause errors on both p.m. and f.m. Tests in a laboratory are limited in value and the real test is performance on circuits. I personally consider that error control, by means of repetition after detection of the errors, will be necessary to provide an acceptable standard of transmission of data over communication circuits.

One of Mr. Smith's illustrations shows the interface connections. At Geneva a working party studied the standardization of the interface and one result was to improve the terminology, for example the American term "interlock" was replaced by "data set on line" which should help in understanding the function.

Concerning data transmission over telex circuits, since the Post Office provides all the equipment at the telex subscriber's office, it intends to provide, on hire, the equipment to provide loop check error detection as described by Mr. Simmonds. The arrangement requires a small change in wiring at the subscriber's office to provide for full duplex operation (for which there would be a small charge). Since the U.K. telex system provides for full duplex operation it favours the use of loop check which provides error detection without reduction in circuit efficiency.

Mr. R. Buttery: Mr. Smith has described an admirable complete data collection system, but I think he is wrong in inferring that the only sensible use of data transmission is as part of such a complete system. There are many concerns who already produce punched tape information and who wish to process it at some central office, requiring some means of transference of the tape, preferably at a higher speed than is obtainable using telex or the normal post system: such a requirement, in fact, as has been described by Mr. Dace. To meet such a demand we have produced what we hope is a reasonably priced equipment to provide a simply-operated data link for use on the public switched telephone network. The transmission rate of 5-hole tape is 67 characters per second and the output is in the form of reperforated tape or a magnetic recording with start and stop signals for teleprinter operation. The latter can be replayed at 6·7 characters per second for print-out on a standard teleprinter.

The cost of such a system involves the cost of buying or hiring the terminal equipments, plus maintenance and depreciation charges and cost of telephone lines, against the number of characters to be transmitted per working day. We have produced figures for typical installations showing when the use of such a link becomes an economic proposition.

Mr. Dace referred to a trial of our equipment with his organization, which involved transmission between Liverpool and Croydon. We ran into difficulties here, mainly because of the presence of echo-suppressors on some of the lines used. Difficulties with echo-suppressors have been mentioned briefly to-day, but I think these might be explained more fully. All the data transmission systems use some form of "feedback" information from receiver to transmitter. In our case supervisory tones are sent back using a lower bandwidth of approximately 300 c/s to 450 c/s. When an echo-suppressor is present the information signal in the band 900–2000 c/s switches the device in its favour and the supervisory channel is blocked. Thus the data transmission equipment cannot operate. The G.P.O. are no doubt doing all they can to obviate this, but it will obviously take time.

Mr. R. W. Brown: All the papers presented to-day have dealt with data transmission over lines. There are, however, many applications where lines cannot be used and where recourse has to be made to radio, regardless of the increased error rate.

It was pointed out by Mr. Aylott that an ARQ system has the virtue that in the absence of error no redundancy is required, and in this respect it approaches more closely than most error correcting codes the condition of having what I believe Mr. Smith called the "inverse characteristics" of the transmission channel. However, when only a simplex channel is available, real time transmission is required, or the necessary re-transmission creates too many practical problems, sufficient redundancy for error correction must be added to the data. This redundancy should match as nearly as possible the error statistics of the channel. The most successful method of achieving this would be to employ "dynamic matching", i.e. to monitor continuously the error statistics of the channel and to adjust the error correcting code accordingly. This, however, would be an extremely costly and complex business and the most practical alternative is to collect error statistics over as long a period as possible, and to predict the future error statistics from these.

We have been conducting some experiments to examine the feasibility of using a short-wave radio link for the transmission of seismic information over a distance of 300 miles. The experiments have consisted of a determination of error rate and error distribution as functions of transmitter power, frequency, transmission rate and time of day. Our approach has been similar to that described by Mr. Smith in his first paper. We have developed an error rate and distribution analyser which consists mainly of synchronizing pattern and test pattern generators at the sending end, and a synchronizing pattern recognizer, test pattern generator and comparator at the receiving end. When an error occurs in transmission the comparator generates an error pulse which is fed to a counter. The received data is also recorded on magnetic tape which provides the input to a data reduction system, the output from which gives the distribution of correct bits between errors, the frequency of occurrence of 1, 2, 3 etc., errors per character and also provides information on the distribution of character errors. One difference between our system and that described by Mr. Smith is that in place of his pseudo-random pattern we use the 1024 10-bit binary numbers generated in sequence.

From the results of these tests it will be possible to determine within certain probability limits the grade of service obtainable with a given transmitter power and error correcting code. In order to maintain a given grade of service throughout a substantial period of each day it will probably be necessary not only to vary the transmission speed, but to change the error correcting code to combat changing types of error distribution. This could best be achieved by employing a group of mathematically related codes in which a change of code could be effected by changing the order of occurrence of a set of basic logic operations common to all codes of the group. The cyclic codes mentioned by Mr. Smith might well have possibilities in this field.

THE AUTHORS' REPLIES

Mr. K. L. Smith: I agree entirely with Mr. Williams' comments on the need for standardization of certain essential parameters in modulating systems and the considerable amount of agreement on these characteristics at C.C.I.T.T. is due in no small measure to the leadership shown by the G.P.O. This will do much to ease the problems of international data transmission in future. It is rather unfortunate, though, that the clear advantages of p.m. over f.m. were not developed sufficiently to allow a clear decision to be made in favour of p.m.

I sympathize with Mr. Williams' preference for "modulation coherence" to describe the relationship between the modem carrier frequency and data clock frequency instead of the more common "synchronous" although, from my own experience, I know of no confusion which has arisen. However, Mr. Williams has the support of C.C.I.T.T. who define modulation coherence as "modulation in which the succession of significant instants is simply related to the characteristics of the current transmitted to line". There is possibly some confusion in Mr. Williams' mind, though, on what is meant by a synchronous p.m. system, since in addition to the simple numerical relationship between the data lock frequency and the carrier frequency, there is also a distinct phase relationship, since in all practical systems exhibiting modulation coherence the modulation instant occurs when the carrier is passing through zero.

The availability of the G.P.O. modem with full compatibility with the C.C.I.T.T. interface will greatly facilitate the development of data transmission applications in the U.K. As Mr. Croisdale points out, the original term "interlock" was replaced by "data sent on line" at the last C.C.I.T.T. meeting, but whether this is an improvement on the original term or not is, I suggest, a matter of opinion, albeit a majority one in favour of the latter term in view of its adoption. Personally I believe that more attention should be given to brevity and convenience in use in choosing these terms, rather than their precise meaning. After all, language is the conveyance of thought and provided everybody with a need to understand does understand what is meant by a term, then that term should be adopted. My reference to the interface itself was intended more as an illustration of the form it takes rather than as a precise definition of a particular interface. The complete interface adopted at C.C.I.T.T. is, in fact, far more complex than that shown, involving 23 links from which suitable selections can be made by individual communications authorities.

Mr. Croisdale's expectation that C.C.I.T.T. might possibly recommend a standard p.m. modem implies that they have already recommended a standard f.m. modem. According to reports I have available, C.C.I.T.T. have in fact recommended neither a standard f.m. nor p.m. system, but have rather restricted themselves to the recommendation of two alternative sets of carrier frequencies for use in each system.

I agree with Mr. Croisdale that many errors are caused by circuit interruptions and bursts of impulse noise, but suggest that the actual effect is more complex than a catastrophic loss of data regardless of the type of modulation employed. During the period of an actual complete circuit interruption clearly no data will get through, but at the beginning and end of such interruptions marginal conditions exist and each system has its own characteristic threshold for its ability to reduce the error rate under such conditions. The characteristic threshold for p.m. is higher than for f.m., i.e. it has better rejection capabilities. Furthermore many interruptions are not complete but appear as loss variations introducing marginal conditions where the better rejection capabilities of p.m. do count. It is true that the curves shown in the paper refer to laboratory tests, but these are the only meaningful data which can be presented since they represent a repeatable performance under controlled and specifiable conditions. Similar curves are available showing the relative performance of the two systems over actual circuits, but these can only be regarded as indicating a general order of performance at one instant of time on a particular circuit. If repeated on the same circuit at a later period they might differ from the earlier results by as much as an order of magnitude. However, from such measurements made on a wide range of circuits in many countries in Europe it is clear that the improved performance of p.m. over f.m. is even greater than that shown by the laboratory tests and has a typical value of 10 dB. On one circuit only in the large number tested did the f.m. system show an advantage over the p.m. system and this occurred only when the line loss was artificially increased by 10 dB—under normal conditions the asynchronous p.m. system had the better performance.

I should like to answer Mr. Worthy's questions by reference to the asynchronous phase modem. The demodulation process is effected by multiplying the received signal by a locally produced carrier signal in a normal balanced modulator, designed especially to achieve a high level of balance so as to reduce the jitter effect caused by interaction with higher level intermodulation products. The local carrier is derived from the received signal through a full-wave rectifier which, in effect, removes the modulation.

The reason for pre-modulation filtering of the data signal is to remove the higher harmonics which would otherwise produce intermodulation products near the carrier frequency. These cause the characteristic jitter of the demodulated signal which occurs in badly designed asynchronous systems. A typical intermodulation product is $(3f_c-3f_s)$, e.g. if we have a data speed of 2420 bauds (f_s = 1210 c/s) and a carrier f_c = 1800 c/s, then $(3f_c-3f_s)$ = 1770 c/s, which is close to the carrier frequency and could cause a 30 c/s jitter in the data output from the demodulator.

Thus elimination of these dangerous harmonics in the data signal is essential. Also great care must be exercised in the design of the modulator to achieve the best possible balance to avoid the re-introduction of these harmonics.

I think that Mr. Buttery has misunderstood my remarks on the need to regard data transmission systems as part of a larger system designed to integrate the total data processing requirements of a business organization. The main advantage of the integrated data processing approach is the ability of management to determine the true interaction between the various operating units of their complex since all relevant data are stored within the one system. My point was that the true value of data transmission systems and their economic value only emerge by taking such a total system approach. I trust Mr. Buttery will concede that the data collection operation is but one type of data transmission system used to speed the entry of data into the data processing system and ensure that its calculations will be based on most recent, and hence meaningful, data. Other, more advanced data transmission systems are necessary to interconnect data processing systems. My remark was intended primarily to answer those opponents of data transmission systems who point out that it is cheaper to send data through the parcels service. It is only by adopting a total systems approach that the function of the data transmission system and its speedy transfer of data is appreciated in the correct perspective.

Mr. A. P. Clark: In drawing his conclusions from the results of tests with different data transmission systems over telephone circuits, Mr. Chittleburgh appears to have overlooked one rather important fact, namely the extremely wide spread in the noise characteristics of different telephone lines. Thus when two different data transmission systems work over two different telephone lines, the effects of the noise characteristics of the two lines on the resultant error rates will in general completely swamp the effects of the tolerances of the two data transmission systems to the noise. Consequently what one really measures in such a comparative test is not the relative performance of the two data transmission systems but the relative noise levels of the two telephone lines. It is largely for this reason that most of the information on test results which has been contributed to the C.C.I.T.T. Special Study Group A is of such an inconclusive nature.

Mr. Croisdale's statement that most of the mutilations which cause errors in data transmission on lines are caused by brief disconnections, deserves some further comment. If taken literally this statement is almost certainly true, in so far that the greatest single hazard over any telephone circuit on the public network is a temporary disconnection lasting anything from a fraction of a second to several minutes and bearing in mind that a disconnection for one second may produce hundreds of errors. However the conclusion which has been implied appears to ignore the fact that the receiver in any data transmission system using f.m. or p.m. signals, as for instance the proposed Post Office f.m. system, should normally have a circuit for detecting the loss of the received signal whenever this exceeds some ten milliseconds, thus enabling the data processing equipment at the receiving terminal to detect the errors produced by any such disconnection. Strictly speaking then, the errors resulting from any disconnection exceeding ten milliseconds should be neglected in describing test results, since this condition is treated by the equipment not as noise but as a fault condition and cannot in any case cause false operation in a correctly designed system. Furthermore, in the published results of various data transmission tests over the public network in western Europe in which tests a disconnection had to exceed some 300 milliseconds before it was regarded as a fault condition and the resultant errors ignored, as recommended by the C.C.I.T.T. Special

Study Group A, the results show in general a significant reduction in error rate as the signal level is increased. This necessarily implies that the majority of the errors are caused by additive noise, such as for instance impulsive noise, and not by multiplicative noise which here includes brief disconnections lasting less than 300 milliseconds. The tolerance to additive noise (for which white noise is most often used in the laboratory) is therefore a most important measure of the suitability of a data transmission system for use over telephone circuits.

Messrs F. G. Jenks and D. C. Hannon: The advantages of standardizing parameters for data transmission are of course considerable but, once standards are established by so august a body as the C.C.I.T.T., they are not easily changed. It is therefore extremely important that the technical aspects should be fully understood before standards are agreed.

In the present instance there seems to be some foundation for the view that agreement has been reached with undue haste. It seems, for example, that those responsible for the agreement have been thinking primarily in terms of frequency modulation and the digit rates proposed could probably be raised if phase modulation were used. It would also seem to be more logical for the proposed rates to form a geometric rather than an arithmetic series. Again, in connection with the interface, the decision that digit timing extraction shall not be a function of the receiver is, from a technical viewpoint, indefensible. The sacrifice in performance is apparently to be condoned in order to reduce the number of grounds on which the customer can complain about equipment performance.

Mr. Williams is, of course, right in drawing attention to the ambiguous use of the word "synchronous" and we would support the use of the alternative term "coherent" which he suggests, though this also has been frequently used in another sense.

It is true, as Mr. Croisdale observes, that an appreciable proportion of errors on practical circuits will be due to brief disconnections (a further reason why digit timing should be extracted in the receiver whose designer, being aware of the problem, will arrange to bridge the shorter gaps and so keep the number of lost digits to the minimum) and it is true that our laboratory measurements make no allowance for this factor. On the other hand, we think that both Mr. Croisdale and Mr. Chittleburgh will agree that it is not possible to make truly comparative tests over actual circuits unless these are of formidable duration, due to the extreme variability in line performance.

Our tests do allow for the errors due to high levels of interference since our impulsive noise test recording contained samples of this kind. For example at 1500 bauds, on average one digit per thousand suffered interference from a noise peak exceeding $-22{\cdot}5$ dBm. We do not agree that there is no difference between the modulation systems in error performance at high noise levels unless the noise is so strong that severe limiting occurs. It is for this reason that a receiver which uses a linear detector and which does not limit until the noise is well in excess of normal signal level, is desirable. We accept the need for error control by repetition after detection of errors, which Mr. Croisdale advocates, and feel that our tests show conclusively that there will be need for a substantially greater number of repeated messages if f.m. rather than p.m. is used. Moreover under difficult conditions and particularly at higher data rates, it will be found that a line may still give a useable output with phase modulation when it can no longer be used with frequency modulation.

We would agree with Mr. Worthy that some form of analogue multiplication is necessary if a true correlation process is to be achieved. While it is difficult to design a multiplying circuit which would be suitable for a practical receiver it is possible to approach it closely in laboratory equipment using, for example, the square-law characteristic to which a diode approximates. A practical alternative is to use a circuit of the switching modulator type which in effect multiplies the incoming signal by a square wave, rather than a sine wave, of the appropriate frequency and phase. We have compared the performance of p.m. receivers using true and switching multipliers and have found a worsening of about 1 dB in signal to noise ratio for equivalent error performance when the switching multiplier is used.

We were interested to learn of Mr. Buttery's difficulties with echo-suppressors. This is a problem which only the G.P.O. can solve and we hope that guidance from that quarter will soon be forthcoming. We should also welcome written confirmation from the G.P.O. that the practice of allowing rapid corrections of frequency translation has been abandoned in view of its relevance to methods of p.m. detection.

Mr. D. J. Dace: The only comment I would make on Mr. Buttery's remarks is that although our own trials with Ericsson equipment were unsuccessful, there is no reason to believe that they would have been anything but successful between any other two locations. In fact we would have welcomed the opportunity of further testing the system from different locations.

Messrs. E. R. Aylott and E. S. Simmonds: We thank Messrs. Williams and Croisdale for their statements regarding the plans of the Post Office in respect of data transmission and are glad to hear of the policy of standardization. We endorse Mr. Croisdale's remarks regarding the usual form of interference encountered on telephone circuits, which will cause errors in a data message whatever form of modulation is used, and because of this, a comprehensive error control system is necessary. Mr. Croisdale also gives, as his personal view, that error correction should be by retransmission. This system is covered in our paper by Schemes B, C, and D, and we give our wholehearted support of this view.

In ARQ systems as used in radiotelegraphy and referred to by Mr. Brown, in the absence of errors, no signals are required in the reverse direction. Redundancy is always present in the transmitted message. The advantage of this arrangement is that the reverse direction channel is available for traffic in the reverse direction, except when errors are detected.

U.D.C. 621.387.46

Semiconductor Nuclear Radiation Detectors

By
G. DEARNALEY, Ph.D.†

Summary: Semiconductor detectors of the junction type have recently revolutionized work with low energy charged particles in nuclear physics. Solid conduction counters have also been improved through the use of modern semiconducting materials but are as yet at an early stage of development. The mode of operation, methods of preparation and characteristics of these various detectors are described, together with the requirements for pulse amplification. Finally some of the many applications of semiconductor junction detectors in nuclear physics are discussed in outline. 70 references.

1. Introduction

The past three years has been a period of rapid and important development in the field of nuclear radiation detectors which has been stimulated by the growing availability of semiconductor materials of high purity and controlled characteristics. Already for charged nuclear particles such as alphas, protons, etc., solid detectors of silicon and germanium have been shown[1-7] to have excellent energy resolution and linearity of response, added to a fast rise-time, very compact size and a relatively low sensitivity to unwanted gamma-rays and neutrons. They also possess the advantages of low cost, ease of fabrication, simple power supply requirements, and adaptability to a growing collection of special applications using simple modifications to the basic design. For very many experiments the old scintillator and gas-filled particle counters have been superseded by semiconductor counters and, with more recent work on bulk conductivity detectors of larger sensitive volume, it seems that semiconductors in one form or another will ultimately replace other detection media for gamma-rays and high-energy particles also. It is, after all, a simpler mechanism to convert the energy of radiation into electron-hole pairs directly rather than through an intermediate stage of optical radiation as in the scintillation process.

2. Bulk Conduction Counters

Chronologically, the development of the semiconductor junction detector preceded improvements in bulk detectors, but we shall deal with the latter first as they are based upon a simpler physical situation.

The detector of nuclear radiation is almost always based on the ionization or excitation of the atoms of the detection medium by the passage of charged particles. Uncharged radiations, such as neutrons and gamma-rays, are detected by the transfer of their energy to charged particles which themselves cause further ionization. The subsequent history of the ions and electrons produced depends on the state of the medium and the presence of applied electric fields. In semiconductor counters the situation is in many ways analogous to the gaseous ion chamber, for in both cases an electric field is established across a medium of low conductivity. The traversal of this medium by a nuclear particle liberates charges which are separated by the field and collected at the boundary electrodes, producing in the detector capacitance a voltage pulse which can be amplified and recorded. The circuit of such an arrangement is shown in Fig. 1.

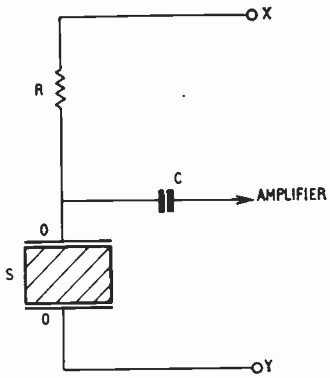


Fig. 1. Bulk conduction counter circuit. S is the semiconductor crystal, with ohmic contacts O,O. Since the counter is symmetrical the high tension collecting potential may be applied at either X or Y.

The earlier type of solid ionization chamber known as the crystal counter was developed by Van Heerden[8] and Chynoweth[9] and consisted of rather small single crystals of diamond, zinc sulphide, cadmium sulphide and other materials of large band gap. A high electric field applied by surface contacts results in collection of a certain amount of charge liberated in ionization. Minority carrier lifetimes are low however; for example the lifetime is about 10^{-8} seconds in diamond, so that appreciable recombination and trapping occurs before the carriers can reach the boundaries.

† Nuclear Physics Division, Atomic Energy Research Establishment, Harwell, Didcot, Berkshire.

There are a number of conditions which must be satisfied if a detector is to be of practical use. The first is that a sufficiently large number of ionization events occur, and this is measured by the mean energy absorbed for each electron liberated. Here solid detectors have a considerable advantage, for whereas in a gas counter some 20–30 eV are absorbed per electron-ion pair produced, in silicon and germanium the energy required per electron-hole pair is only about 3 eV. The output signal, other things being equal, is therefore seven to ten times greater in the latter case.

Secondly, the number of carriers produced by ionization by the particle detected must exceed the random fluctuations in number of free carriers normally present in the field region, so that the signal should be detected above noise. This condition calls for a high resistivity medium, easily attained in a gas, but only with difficulty in the most suitable solids.

Finally, the carriers must be swept out of the field region before there is too serious a probability of recombination or trapping, and preferably this time should be short so that the counter may be operated in a high flux of particles. In gas counters the mobilities of the ions are about 10^3 times less than that of the electrons so that special measures must be taken to construct fast chambers which operate by electron collection. In semiconductors the electron and hole mobilities differ only by a factor of about three so that this arrangement is unnecessary. However, trapping and recombination are a much more serious problem in the solid medium, and here the necessity for very high purity and crystalline perfection makes great demands on present semiconductor technology. Generally the holes are trapped more readily than the electrons, and if μ_h and τ_h are the hole mobility and lifetime, respectively, then the velocity of carriers in a uniform field E is $\mu_h E$ so that for efficient collection the product $\mu_h \tau_h E$, the hole trapping length, should be large compared with the width of the sensitive volume of the detector. In Table 1 are listed some of the principal semiconductors and the values of μ_h and τ_h which obtain for the purest materials at present available. It can be seen that only silicon and germanium are suitable if serious difficulties from trapping are to be avoided.

Besides loss of carriers there are other effects of slow carrier collection. Polarization results from the charge which accumulates in the field region and which produces a field in opposition to that applied for collection. Unequal collection due to preferential trapping of one carrier type results in a net charge which can cause injection from the boundaries of carriers of the opposite sign in order to approach equilibrium. During this time an additional "secondary" current flows and the integration of this current yields a total collected charge which is not proportional to the particle energy. In cadmium sulphide this secondary current flow is large[10] and is of long duration so that the charge collected may be many times that produced by the initial ionization. This is an advantage in a measurement of a mean level of radiation, but the non-linear response renders it unsuitable for the detection of individual particles.

Materials such as diamond, cadmium sulphide and silver halides are usually non-uniform and their properties are not easily controllable or reproducible, so that further development of the subject awaited the advances in semiconductor technology which yielded pure materials with long carrier lifetimes and high carrier mobilities.

Silicon at room temperature possesses an intrinsic resistivity of about 3×10^5 ohm cm, which has rarely been approached owing to impurity conduction. This is much too high a conductivity, as any useful volume would contain more free carriers than would be injected by an ionizing particle. Two methods of

Table 1

Material	Temperature	Hole mobility, μ_h	Hole lifetime, τ_h	Product $\mu_h \tau_h$
		cm² volt⁻¹ s⁻¹	seconds	cm² volt⁻¹
Diamond	300°K	1200	10^{-8}	10^{-5}
Silicon	300°K	500	2×10^{-3}	1
Gold-doped Silicon	140°K	10^4	10^{-7}	10^{-3}
Germanium	300°K	1800	10^{-3}	1·8
Germanium	78°K	$1{\cdot}5\times10^4$	10^{-3}	15
Tellurium	300°K	560	10^{-8}	5×10^{-6}
Indium Arsenide	300°K	3×10^4	6×10^{-8}	2×10^{-3}
Gallium Arsenide	300°K	10^3	7×10^{-7}	7×10^{-4}
Cadmium Sulphide	300°K	50	10^{-8}	5×10^{-7}

utilizing silicon for detectors have been applied. The first is to dope it with gold when, because of the multiplicity and positions of the levels introduced in the conduction band, compensation is achieved for both acceptor and donor impurities. Thus Van Putten and Vander Velde[11] obtained a resistivity of over 10^9 ohm cm at $-130°$ C, and at this low temperature the mobility of carriers is very high, about 6×10^3 cm^2 volt^{-1} s^{-1} so that collection can be achieved throughout a volume several millimetres thick without too much recombination. The gold-doping reduces the carrier lifetime drastically to 10^{-7} seconds so that the method is not applicable to large volume detectors for which the transit time would be greater. The hole trapping length is about 1 cm.

A second method has been to use high-purity silicon without doping at a lower temperature of 77° K. Gibbons and Northrop[12] prepared detectors by ohmic contacts of diffused boron on opposite faces of *p*-type silicon blocks with initially 5000 ohm cm resistivity (similar units were prepared by surface layers of phosphorus on *n*-type silicon). The thermal cycling process necessary for diffusion caused an increase in resistivity and a decrease in carrier lifetime.

Such detectors with a collecting field of 500–1000 volts/cm have shown an energy resolution of 10% for 1·3 MeV Co^{60} gamma-rays, and 6·6% for 30 MeV protons. Problems of trapping and the effect of surface states are under investigation.[13]

Perhaps the most promising material for bulk or junctionless semiconductor counters for use at room temperature is gallium arsenide. Normally the purest material possesses a resistivity of about 10 ohm cm, but by a process of oxygen doping[14] a stoichiometric compensation occurs and the resistivity rises abruptly to the region of 10^8 ohm cm. The carrier mobility is reduced by a factor of approximately two and is therefore about 2×10^3 cm^2 volt^{-1} s^{-1} for electrons, and one tenth of that value for holes. Efficient collection was obtained over several millimetres in a field of several hundred volts/cm. The carrier lifetime is probably very low, and smaller than the observed pulse rise-time of the order of 1 μs, which was attributed to temporary trapping without appreciable recombination. An energy resolution of a few per cent was achieved for 5 MeV alpha particles in such a detector.

In all these detectors the type of contacts made is of considerable importance, and difficulties are often experienced through injection of carriers into the field region. Some form of blocking contact is needed in which the electrons are prevented from entering the bulk semiconductor,[15, 16] and it seems that further research on this aspect of the problem is required.

The value of an efficient bulk conductivity detector of dimensions 1 to 10 cm cannot be over-estimated. High resolution gamma-ray spectroscopy would then be possible, as well as applications to high energy charged particle detection. The scintillation detector systems at present employed yield roughly one photoelectron for each 1 keV energy loss of the incident radiation. This yield might be increased 100 times in a semiconductor detection medium, with corresponding increase in resolution, while problems of optical coupling to a photomultiplier tube are eliminated. There is likely to be a much lower sensitivity to magnetic fields, which often calls for heavy iron shielding of photomultipliers.

In nuclear physics we are faced with the situation in which gamma-ray emission processes can be studied with far less precision than cases of particle emission, and an improvement in the present methods would open a wide field of new research.

3. Semiconductor Junction Detectors

The problem of noise due to the presence of carriers in the semiconductor can be solved in a limited volume by making use of the properties of an *n-p* junction. In the *n*-type, or donor-rich region, the majority carriers are electrons donated by impurity atoms with electron

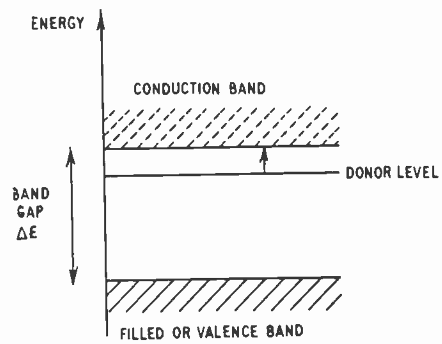


Fig. 2. Electron levels in an *n*-type semiconductor.

energy levels near the conduction band (Fig. 2), and such holes as are produced by random processes soon recombine. If a contiguous *p*-type region, rich in electron-acceptors, is formed electrons will tend to leave the *n*-region and be captured until a small potential difference is established which prevents the intermingling of carriers; this potential difference is generally equal to about half the band gap, ΔE. A shallow volume at the interface is thus depleted of carriers leaving an excess of ionized impurity atoms, positive in the *n*-region and negative in the other. An applied voltage of the same sign will enhance this effect and extends the "depletion layer" to a thickness which may be as great as 1 mm under the most favourable conditions. Carriers due to impurity atoms are thus absent from this layer and if the intrinsic conduction by thermally-excited carriers is low there will be little noise to act as background to the detected signal. In materials such as germanium with rather a small band gap cooling is necessary to reduce intrinsic conduction. The field in the depletion layer is non-uniform, decreasing to zero at a certain depth from the junction, so that the collection field is a function of the position in the detector at which carriers are produced. The first measurement of pulses in a semi-conductor junction due to carriers injected by nuclear particles was by K. G. McKay[17] in 1949. In order that the junction should be easily accessible to the ingoing particles one of the regions, *n* or *p*, is made very thin compared with the range of the particles detected.

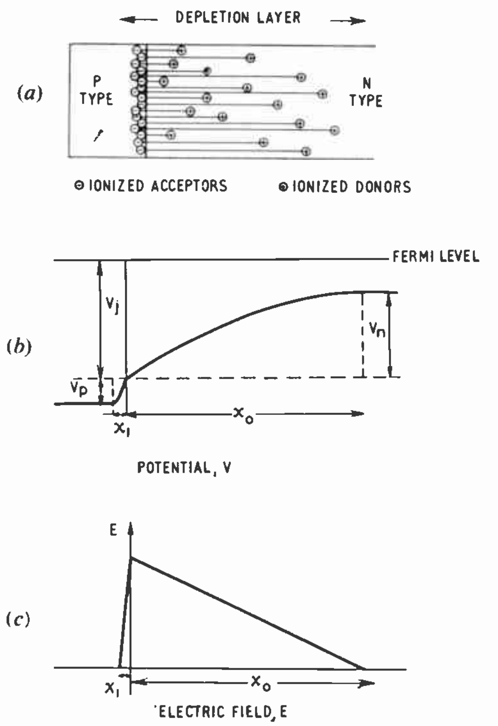


Fig. 3. Potential and field distribution in the depletion layer of a *p-n* junction.

Whereas in the general case a depletion layer is formed on each side of the junction, extending from the boundary into both *n* and *p* regions, it is usual to ensure a much greater impurity concentration in the surface layer so that the depletion layer formed in it is very thin. This is simply because for overall neutrality the number of ionized centres in each region must be the same.

It is very easy to derive the essential properties of the depletion layer on the basis of a simple theory. A more complete treatment has been given by Walter *et al.*[18] Consider the *p-n* system illustrated in Fig. 3(*a*), in which the potential V of an electron is represented in the graph 3(*b*) as a function of the distance x from the interface. We assume for simplicity that all the donor centres are ionized up to the point x_0 at which the field $\mathrm{d}V/\mathrm{d}x$ becomes zero, as it must be in the normal semiconductor. Similarly all acceptors are ionized up to x_1, in the *p*-region. We neglect the presence of acceptor impurities in the donor-rich *n*-region, and vice versa. Finally we assume that the number of thermally-excited carriers is negligible at the temperature chosen. If N_d is the donor density throughout the *n*-region, N_a the acceptor density in the *p*-region with an abrupt discontinuity between, and with $N_a \gg N_d$, we have by Poisson's relation in *n*-region:

$$\mathrm{d}^2V/\mathrm{d}x^2 = 4\pi N_d e/k \text{ (practical c.g.s. units)}$$

where e is the electronic charge and k the dielectric constant, which remains invariant for the low impurity concentrations considered here. Similarly, in the *p*-region:

$$\mathrm{d}^2V/\mathrm{d}x^2 = 4\pi N_a e/k$$

Integration yields $\mathrm{d}V/\mathrm{d}x = 4\pi N_d e\,(x-x_0)/k$

and $V = V_j + 2\pi N_d e\,(x^2 - 2xx_0)/k$ in the *n*-region

so that $x_0^2 = V_n k/2\pi N_d e$ and $x_1^2 = V_p k/2\pi N_a e$

Since for neutrality $N_a x_1 = N_d x_0$,

$$V_n/V_p = N_a/N_d \gg 1$$

and we therefore neglect V_p compared with V_n in what follows.

When a potential difference V_0 is applied to the junction so that the *p*-type layer is negative the height of the potential barrier is increased by V_0 and

$$x_0^2 = k(V_0 + V_n)/2\pi N_d e.$$

Generally $V_0 \gg V_n$ since the latter is only about 0·5 volts, and then $x_0^2 \simeq V_0 k/2\pi N_d e$.

Since the impurity conductivity $1/\rho$ is proportional to $\mu_e N_d$ and $\mu_n N_a$ in the n and *p*-type volumes respectively where μ_e, μ_h are electron and hole mobilities we have:

$$x_0 \propto \sqrt{\rho V_0 \mu} \text{ and approximately}$$

$$x_0 \simeq 0{\cdot}5\sqrt{\rho V_0} \text{ microns in } n\text{-type silicon}$$

$$\simeq 0{\cdot}3\sqrt{\rho V_0} \text{ microns in } p\text{-type silicon}$$

where the resistivity ρ is measured in ohm cm and V_0 in volts.

It is to be noted that the field $\mathrm{d}V/\mathrm{d}x$ at a given point is proportional to $\sqrt{V_0}$, a factor which is important when we consider recombination effects.

The capacitance of the junction arises from the existence of an insulating layer of width x_0 between conducting regions in material of dielectric constant k, so that the diode capacitance per unit area C_d is given by

$$C_d = k/4\pi x_0 \simeq 1{\cdot}8\times 10^4/\sqrt{\rho V_0}\ \text{pF/cm}^2 \quad \text{in } n\text{-type silicon}$$

and $$C_d \simeq 3\times 10^4/\sqrt{\rho V_0}\ \text{pF/cm}^2 \quad \text{in } p\text{-type silicon}$$

It has been realized for some time that the problem of obtaining a deep field region for the linear detection of long-range particles might be approached by a different type of structure to the *p-n* junction. If a volume of near-intrinsic silicon is formed between regions doped respectively *p*-type and *n*-type and a reverse bias is applied to the resulting *p-i-n* structure the residual carriers will be swept from the intrinsic region which thus becomes an insulator. Under these conditions the whole intrinsic volume is sensitive to charged particles which enter it, and the situation is most analogous to a parallel-plate ion chamber. A number of attempts have been made[19] to construct a useful detector in this way, but the most important development has been the introduction of a method of

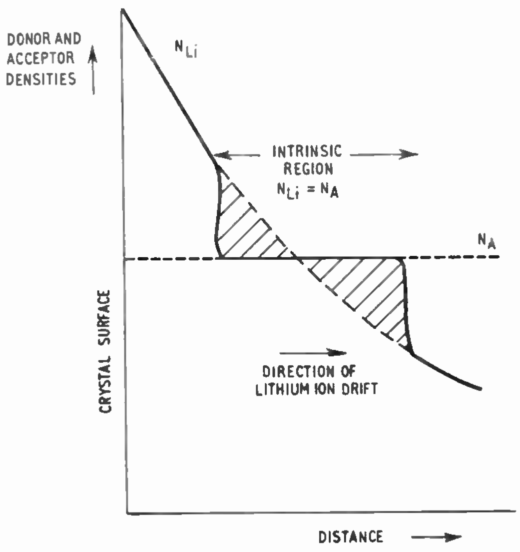


Fig. 4. Ion density distribution during the lithium ion-drift process. N_{Li} is lithium ion density, N_A the initial density of acceptor ions.

compensation of silicon by ion drift, due to E. M. Pell.[20] Lithium ions are small and have a high mobility in silicon, in which they act as donors. If a high concentration of lithium is diffused into the surface of *p*-type silicon a normal *p-n* junction results. If now a reverse bias is applied, preferably at an elevated temperature of about 150° C, the positively charged lithium ions drift through the junction into the *p*-type region. After a while the impurity distribution changes as shown in Fig. 4. The compensation which takes place is automatically stable, for if the Li concentration falls a little at one place the field will become higher on the side nearer the Li-rich surface so that more ions diffuse there; if the lithium concentration rises somewhere, the field will become higher on the opposite side, tending to remove the excess Li. Intrinsic regions as wide as 4 mm have been produced in this way. Of course the introduction of additional impurity centres lowers the carrier lifetime and increases recombination effects, but even so a resolution of 84 keV for 1·32 MeV protons has been achieved.[21]

4. Preparation of Junction Detectors

4.1. *Diffused Junctions*

A number of techniques have been described[22, 23] for the diffusion of *n*-type doping agents into *p*-type silicon. Principally phosphorus is used, either in the gaseous phase or by a "paint-on" technique as P_2O_5 in ethyl cellusolve at temperatures of 700–900° C. The surface of the crystal is initially cleaned and smoothed by chemically etching in a solvent such as CP4, or by mechanical polishing with ¼ micron diamond powder. The diffusion depth required is about 0·1–1 micron, and it is important to avoid a non-uniform coverage which causes a high sheet resistance over certain areas, leading to noisy performance. After diffusion the *n*-type layer is stripped from all but the required sensitive surface by masking with Apiezon W wax followed by etching and careful washing. Contact to the back face can then be made by soldering, by amalgamating to a metal plate with indium-mercury amalgam, or by a conducting silver paste. Contact to the front electrode is achieved by thermal bonding of gold, by spark welding of a copper wire, or by silver paste. Care should be taken to avoid as far as possible any sudden temperature changes during the whole process. Encapsulation of the device has been carried out by masking the junction edge with Apiezon W wax, or with arsenic glass, or by soldering a can to the front surface to give a hermetic seal around the whole unit. The encapsulation seems invariably to result in poorer performance, but without it the diodes are noticeably affected by ambient conditions, in particular high humidity and the presence of ozone. Encapsulation also reduces to some extent the available sensitive surface.

The majority of diffused counters have been of *p*-type base material, of which somewhat greater quantities are available with high resistivity. Buck[24] has shown however that thin surface *n*-type layers are affected by the common ambients of oxygen and water vapour so that more stable performance might be expected from counters made from *n*-type silicon with a surface *p*-type doping. In France the Compagnie LTT have prepared *n*-type counters with a shallow diffusion of gallium. It is, however, difficult to carry out the diffusion process with high-resistivity *n*-type silicon owing to a tendency for it to change into *p*-type, presumably as a result of the rapid diffusion of unwanted impurities.

A few typical diffused junction detectors are shown in Fig. 5.

4.2. *Surface Barrier Detectors*

In these a thin *p*-type surface layer on *n*-type silicon or germanium is allowed to form spontaneously by the oxidation of a chemically-etched wafer. This oxidation takes place over a period of a day or two,[25] after which contact may be made to the requisite sensitive area by evaporation of gold *in vacuo*. Other metals such as platinum or nickel have been used successfully, but gold is most commonly employed owing to its good conductivity and chemical stability. Contact to the gold film, which is about 20–50 $\mu g/cm^2$ in weight, can be made by a fine gold wire or strip bonded with conducting silver paste (for example, Johnson, Matthey FSP 43) which may also be used for the back contact to the silicon. Alternatively the crystal may be encapsulated in epoxy resin to form a surface flush with the crystal and gold is evaporated over the whole face so that contact can be made some point away from the crystal. However, organic contaminants cause deterioration of surface *p*-type layers and the possibility arises of impurities from the resin penetrating the silicon and altering its characteristics. Borkowski[26] has described a method by which iodine is incorporated in the resin to form what has been called "*p*-epoxy" in the presence of which the *p*-type surface oxide layer on the silicon is more stable. At A.E.R.E. it has been preferred to avoid as far as possible any material in contact with the silicon, apart from the two electrodes. The base contact of silver

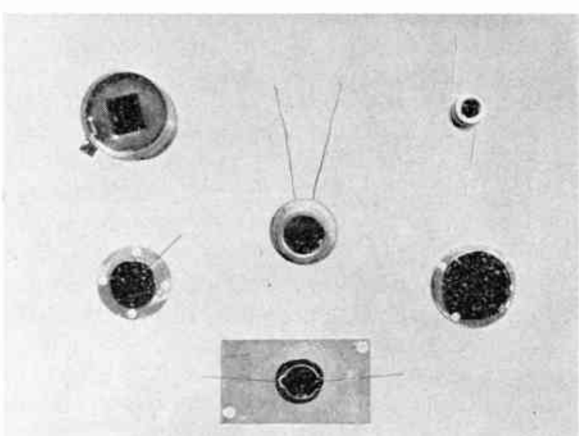


Fig. 5. Some typical semiconductor counters. Upper left—a commercial surface-barrier counter; centre and upper right—commercial diffused-junction counters, encapsulated; lower left and right—two laboratory-prepared surface-barrier counters; lower centre—a thin surface-barrier counter 0·002 in. thick for energy loss-rate measurements.

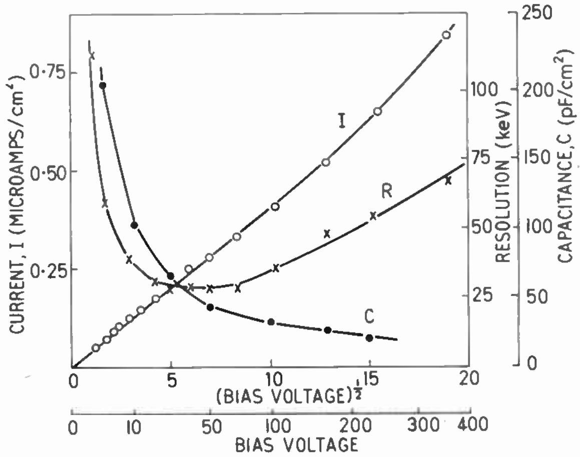


Fig. 6. Capacitance, leakage current and resolution as a function of applied bias in a typical surface-barrier detector, of area 20 mm² and of 3000 ohm cm silicon.

paste possesses fairly low injection properties, or in other words is a suitably poor rectifier. Work by P. T. Andrews[27] has shown that an evaporated film of aluminium on the base is even better and overcomes the problem of carrier injection, which is severe in very thin detectors (see final section) or whenever the depletion layer extends close to the base contact.

It is also preferable to apply bias to the front gold electrode and take the resulting positive pulses from it to the amplifier, the base contact being earthed. This is because whenever a multiple detector is prepared on a single crystal of silicon the individual counters are defined by the separate gold areas at the front, at which the separate depletion layers are formed.

A number of surface barrier detectors are shown in Fig. 6.

4.3. *Lithium-drifted Detectors*

Since the paper by E. M. Pell[20] on the compensation of silicon by lithium-ion drift in 1960 several groups[28, 29] have applied this method to the preparation of particle detectors. *p*-type silicon of 100–300 ohm cm is the usual starting material, in discs of 2–4 mm in thickness. Lithium is applied to one surface either by a "paint-on" technique using a lithium-in-oil suspension (obtainable from the Lithium Corporation of America) or by evaporation *in vacuo*. If the crystal is heated in an inert gas or vacuum at 450° C for 1–2 minutes diffusion to a depth of several hundred microns takes place. Contacts to the lithium-doped surface and to the base are made by nickel plating[30] followed by masking with Apiezon wax while the unwanted nickel is dissolved in acid. Alternatively the contacts may be painted on with conducting silver paste (e.g. Johnson, Matthey FSP 43). The disc is then placed in a silicone oil bath which can be heated and efficiently stirred while a reverse bias voltage of several hundred volts is applied. The temperature of the bath should be maintained at 135° C and the bias voltage must be regulated over many hours since as drifting proceeds the current increases and overheating can easily occur. The depth of the intrinsic layer may be monitored by capacitance measurements. Mayer[28] finds it advisable to store the detectors with a permanent bias voltage applied to maintain the balance in compensation.

5. Characteristics of Junction Detectors

These will be described by commencing with the electrical d.c. characteristics then treating the physical processes involved in detection in the order in which they occur in a single detection operation. Finally we shall consider time-duration effects such as ageing and radiation damage.

The capacitance and depth of depletion layer in a detector depend upon the resistivity of the silicon and the reverse-bias voltage, as was shown above. Approximately, the capacitance is inversely proportional to the square root of the product of resistivity and bias. A convenient nomograph for obtaining a value for the capacitance has been given by Blankenship.[3]

The reverse current of a junction detector arises from a number of sources. One is surface leakage, which will vary under different ambient conditions, and is sometimes a function of time owing to migration of surface ionic contaminants under the strong field. Buck[24] has studied the effect of ambient vapours on silicon detectors, and has shown that in general the surface breakdown voltage and reverse current below breakdown are both increased by ambients which induce a surface inversion layer. Surface breakdown is caused in *p*-type detectors by oxygen, ozone, and chlorine and in *n*-type detectors by water-vapour, ammonia and amines. An explanation of these effects has been given by Statz.[31]

The second contribution to leakage current is due to "space charge generated current" produced at recombination centres throughout the depletion layer. On the theory of Sah, Noyce and Shockley[32] these are also centres for the generation of carriers. The density of recombination centres governs the minority carrier lifetime, τ, and clearly the greater the density of centres the shorter is the lifetime. Space charge generated current is therefore lowest in material of high carrier lifetime, and its magnitude is proportional to the volume of the depletion layer if we assume a uniform distribution of recombination-generation centres. This current is thus proportional to the square root of the bias voltage.

A third factor is diffusion current, which is due to the presence of minority carriers in the field-free volume. These may diffuse to the edge of the field region when they are swept across and contribute a current which is independent of the volume of depletion layer, and therefore is dominant at low bias.

Finally there may be injection of carriers at the base contact and if the lifetime is sufficient for these to diffuse across to the depletion layer they will augment the diffusion current. This effect will depend on bias since this governs the width of the base region. Further studies of this phenomenon on the lines begun by Andrews[27] should be carried out.

Because of this complex behaviour it is not easy to distinguish the various contributions to the leakage current. Often surface leakage can predominate, but in most cases space-charge generation is more important. Low-lifetime material generally makes poor detectors, and in the case of diffused junction counters it is important to learn more about the processes which cause reduction in carrier lifetime as a result of the heating cycle.

The detector also behaves as a source of electrical noise, and is almost always coupled to the input of an amplifier, the noise spectrum of which is itself modified by the detector capacitance. A study of these factors has been made by F. S. Goulding[33] and by E. Fairstein.[34] The detector noise consists of shot noise due to fluctuations in the bulk leakage current, and two types of flicker noise characterized by a $1/f$ frequency dependence. One of these is surface noise due to fluctuations in the density of recombination centres at the diode surface, and the other is leakage noise arising in a conducting channel which by-passes the junction at its perimeter.

When charged particles enter the detector they are brought to a stop in about 10^{-10} seconds causing ionization of the silicon atoms and producing electron-hole pairs in the lattice. Measurements of the average energy absorption per electron-hole pair created yield results about 3·3 eV in silicon and 2·1 eV in germanium. No significant variation of these values with particle type has been observed in studies with electrons, protons, alphas, heavy ions, and fission fragments. However, in the conditions of extremely high ionization density which result with fission fragments some effects of recombination become important. Baldinger and Czaja[35] have shown that in obtaining a true value for the energy expended in the production of an electron-hole pair it is necessary to eliminate the effects of recombination by extrapolation of a plot of $1/Q$ against $1/E$, where Q is the charge collected and E the electric field at a given point in the junction. Here E is proportional, as shown above, to $(V_0+V_i)^{\frac{1}{2}}$ in which V_0 is the applied bias and V_i the internal bias of about 0·6 volts. A linear extrapolation to infinite collecting field yields the saturation charge, which in conjunction with the detector capacitance and incident particle energy gives values for ε, the energy per pair in silicon or germanium, which are lower than the values previously quoted in the literature. A theoretical derivation of the energy per pair has been given by Shockley[36] who suggested that when such a pair is created the residual kinetic energy is shared equally between the two carriers. Measurements on avalanche multiplication[37] and photoproduction efficiency[38] in silicon indicate a threshold for pair production of 3·4 eV so that with an energy gap of 1·1 eV, 1·15 eV remains as kinetic energy for each particle, greatly exceeding their thermal energy. These so-called "hot electrons" were previously considered most likely to create further pairs, but Shockley showed that a more likely process is phonon emission, principally at the Raman frequency, corresponding to 0·06 eV. In fact the processes do not become equally likely unless the electron energy is 1·2 eV above the threshold. As a result of these ideas the threshold value above leads to a figure of 3·5 eV for the mean energy for pair-production in silicon, in quite good agreement with experiment. The work of Baldinger and Czaja[35] showed that the ratio of mean energies per pair in silicon and germanium was equal to the ratio of the energy gaps. Czaja[39] has shown that this is a plausible result of Shockley's theory, and should extend to other group IV semiconductors, such as diamond and α-tin.

Once the electron-hole pairs have been created they begin to separate under the applied field. In the dense plasma of about 10^{18} carriers per cm^3 formed by a fission fragment separation is slow, since the mutual electrostatic forces between carriers are appreciable compared with the collecting field which cannot penetrate the plasma. Such "ambipolar" effects have been observed by Miller[40] who has considered the relative importance of recombination by excitons, direct radiative recombination, and at recombination centres. Miller shows that since the true recombination lifetime, τ, is a function of the carrier density[41] and may be 10^4 times smaller than τ_0, the lifetime for low injected carrier concentrations, the observed losses of about 10% in carriers produced by fission fragments can be satisfactorily explained. The non-linearity observed between pulse height and particle for light and heavy fragments[42, 43] is also accounted for on this basis.

Apart from these effects at very high carrier densities the number of carriers produced for a given incident particle energy is remarkably independent of the particle type. Moreover, the linearity of pulse height with particle energy is excellent, showing that ε, the energy required for each electron-hole pair produced, is constant. This is an improvement over the behaviour of gaseous ionization chambers in which E varies slightly with particle type[44] and with particle energy[45] for alphas and heavier ions. The results in gases are consistent[46] with the assumption that the particles cease to ionize when they have been slowed to the speed of the outermost electrons of the gaseous atoms, i.e. a threshold for ionization of about 48 keV for alphas and 84 keV for Li^7 ions in argon. In the case of semiconductors, calculations by Schweinler[47] yield a threshold of only 0·34 keV for protons in silicon. To test further the absence of this effect in semiconductor counters detailed measurements with low energy Li^7 ions are planned at A.E.R.E.

Linearity of response ceases, of course, when the particles penetrate beyond the depletion layer, since some of the electron-hole pairs are produced in a region with no collecting field and therefore recombine or are trapped. A convenient nomograph for obtaining the range of linear response for protons and alphas in silicon has been given by Blankenship.[48] Carriers produced near the field region of the depletion layer may diffuse towards it and be collected, particularly if the carrier lifetime is very long as in surface barrier detectors. The proportion of this collected charge which contributes to the pulse depends upon the amplifier time-constants, and hence these will determine the shape of the response curve beyond the linear region.

Once separated the carriers move under the applied electric field, inducing charges at the electrodes. The mobility of electrons in silicon at room temperature is about 1300 cm^2/volt sec while that of holes is approximately one-third that value. For a given barrier depth and bias the charge collection time can be calculated, and typical values lie between 10^{-9} and 10^{-8} seconds. In germanium at liquid nitrogen temperatures the mobility is of the order $1{\cdot}5 \times 10^4$ cm^2 $volt^{-1}$ s^{-1}, leading to a calculated pulse rise time of about 5×10^{-10} seconds. However, two important factors have here been overlooked. The detector possesses capacitance, due to the depletion layer, in series with a resistance arising from the volume of undepleted silicon and any base contact impedance. In the pulse amplification, charge must flow through this circuit into the amplifier input capacitance, and the pulse rise-time is accordingly limited. The existence of this effect was brought out in a discussion with J. W. Mayer, who has since verified it experimentally.[49] There were some discrepancies, however, at high detector bias, for which the rise-time was 20×10^{-9} seconds compared with a calculated value of 12×10^{-9} seconds.

There is another factor which limits the rise-time, namely a limiting velocity of carriers in silicon of about 10^7 cm/s owing to the increasing probability of inducing optical transitions in the silicon atoms. The mobility is reduced for velocities above 10^6 cm/s and is therefore not, as Mayer assumed, "well-behaved", while for the case he considered, a 9500 ohm/cm p-type counter with 100 V bias, the hole velocity exceeds 10^6 cm/s for 70% of the transit across the depletion layer. It seems likely that the magnitude of this effect could easily account for the apparent discrepancy which he observed.

Apart from these fundamental limitations, some lower-quality silicon contains so many trapping centres that the carrier collection efficiency is appreciably below 100%. Some of the trapping is temporary, and if the carriers are released within the time-constant of the pulse amplifier, they contribute a slow-collection component to the pulse. This is undoubtedly what was observed in work by Amsel.[50] Since the density of trapping centres probably varies throughout the crystal a poor resolution results. The solution is to obtain silicon as free from this defect as possible. The resolution is naturally improved by the use of amplifier time constants sufficiently long to enable the collection of a high proportion of temporarily trapped carriers.

The requirements for optimum rise-time in coincidence detection experiments are therefore complex. Detector capacitance must be low, and the series resistance minimized by choosing a semiconductor wafer thickness no greater than the depletion layer depth required to stop the particles studied. The evaporated aluminium base contact technique of Andrews[27] is very important here for minimizing the injection current and also the base contact impedance. Surface barrier counters are superior to diffused counters owing to the lower sheet resistance in the electrodes. Finally long-range particles present a serious problem, since the depletion layer depth to stop, for instance, 5 MeV protons is 0·2 mm. With even sufficient bias to produce the limiting velocity of 10^7 cm/s over the whole transit, i.e. over 1000 volts, the rise-time is 2×10^{-9} seconds, and the limiting value rises to 10^{-8} seconds for 12 MeV protons.

The spread in size of the output pulses, which determines the energy resolution of a detector, is due to electrical noise in the detector and the amplifier to which it is coupled. At low bias voltages, amplifier noise is usually the dominant factor, but at high bias the detector noise predominates. The importance of the various types of noise has been discussed by Goulding and Hansen[33] and Fairstein.[34] Another factor involved is the uniformity of collection efficiency over the detector, which depends upon the distribution of trapping centres. In severe cases this effect can give rise to "multiple peaking" in the spectrum of particles from a monoenergetic source. This is avoided only by the use of uniformly pure silicon, and, in the case of diffused junction counters, a diffusion procedure which results in a uniform carrier lifetime throughout the crystal. With detectors of 50 mm^2 sensitive area the best resolution which has been achieved for 5 MeV alphas is in the region of 13 keV and an average performance would be 30–50 keV. With larger area detectors the resolution is poorer, because of the increased noise. The resolution for low energy β-particles seems to be rather better than has been obtained with alphas. Since the factors which govern the resolution are for the most part independent of the ionization in the crystal the value of the energy resolution, in keV, is almost independent of the particle energy.

Turning now to some longer-term characteristics of the detectors, their useful life under normal room conditions of atmosphere and temperature, fortunately, appears to be very long. p-type diffused junction counters are generally encapsulated, owing to their susceptibility to moisture and ozone, and provided the encapsulation is efficient they appear to retain their characteristics over at least two years. n-type surface-barrier counters do not require encapsulation, and also have remained unchanged after two years exposure to air. Some groups have reported deterioration of surface-barrier detectors under vacuum, but we have not experienced this at A.E.R.E. One possible cause of such an effect might be insufficient time allowed for oxidation before the coating by gold, so that the surface inversion layer might be incomplete. It is perhaps possible that some loss of surface oxygen occurs *in vacuo*, or alternatively a film of pump-oil vapour could accumulate to give a positive charge at the surface. Statz[31] has shown that this process could narrow the space-charge region and cause surface breakdown. Deterioration has been observed with germanium detectors, probably as a result of this mechanism, which is enhanced because the detectors are operated at 77° K and so tend to trap vapours at their surface.

Radiation damage is however well established as a cause of deterioration in semiconductor detectors. Lattice defects are produced and in silicon this results in an increase in resistivity due to the trapping of carriers in levels formed near the middle of the band gap. The presence of these trapping centres lowers the carrier lifetime and increases the reverse current. The increased number of recombination centres causes poorer resolution in the detector. Measurements by Klingensmith[51] on a number of silicon surface barrier detectors exposed to fission neutrons showed that deterioration was not appreciable up to about 5×10^{11} fast neutrons/cm^2. Above this dose the resolution worsened and multiple peaking appeared in the spectrum obtained for monoenergetic alphas. After 10^{13} fast neutrons/cm^2 the pulse height was greatly reduced, presumably owing to serious recombination. These measurements were made for a very low bias of 6 volts in 3000 ohm cm silicon, so that the collecting field was low. It is to be expected that recombination effects would be reduced by operating with higher fields, obtained by choosing a resistivity of silicon which will give the required depletion depth, and no more, at a high bias. At A.E.R.E. several detectors were exposed to a high flux of 5·5 MeV alphas from a strong Am^{241} source. After about 10^8 alphas/cm^2 some increase in reverse current could be observed, while after 2×10^9 alphas/cm^2 the resolution began to deteriorate.[7] Some multiple peaking appeared in some of the counters, strikingly similar to the effects observed by Klingensmith[51] for neutron damage. This can only be explained if the recombination centres are produced non-uniformly throughout the silicon. Radiation damage is thought to be due principally to the displacement of atoms from lattice sites, and the ease of damage will depend critically on the energy required for this displacement, which is about 30 eV in silicon. Similar factors affect the ease of chemical solution during the etching process, and it is well known that etching proceeds at a different rate in the vicinity of dislocations, producing "etch-pits". It is therefore proposed that radiation damage occurs preferentially near the dislocations and that this is the cause of the appearance of multiple peaks. This explanation would account for the differences in damage rate observed between one counter and another. There was some evidence that the peaks tended to merge together on allowing the counters to stand at room temperature without further damage, as though diffusion of recombination centres was occurring from local regions of high density. After about 10^{11} alphas/cm^2 the resolution had deteriorated by a factor between 5 and 10 in different detectors. These measurements were again for a low bias of only 2 volts in 1000 ohm cm silicon, while at 20 volts bias giving a field three times greater over the shallow region of damage, the resolution deteriorated only by a factor of 2 to 3 and multiple peaking was never apparent. Further studies of these effects are planned.

Neutrons can also induce charged-particle emission processes by interaction with the nuclei in the detector, and when this occurs within the depletion layer a background is observed due to these particles. At A.E.R.E. Dr. A. T. G. Ferguson and the author[52] have studied the reactions induced in a silicon counter by neutrons in the energy range 5 to 8 MeV. W. Deuchars[53] has made similar measurements at a neutron energy of 14 MeV. The most abundant isotope in natural silicon is Si^{28} (92%) so that the most important background contributions arise from $Si^{28}(n, \alpha)$, the threshold for which is 2·66 MeV, and $Si^{28}(n,p)$ with a threshold of 3·86 MeV. The cross-section[54] for the latter reaction rises from 0·02 barns at 5 MeV to 0·4 barns at 8 MeV, falling to 0·22 barns at 14 MeV. Since the whole depletion layer is effectively a target for these processes the yield due to these reactions may be appreciable in a strong neutron flux, but it can be minimized by operating with a depletion layer no greater than is necessary for the charged particles under study. Obviously, this problem is equally important for diffused junction and surface-barrier detectors, and could be serious with very wide depletion layers in lithium-drifted counters. The (n,p) reactions in the germanium isotopes have much smaller cross-sections[54] owing to the higher potential barrier for charged-particle emission, and alpha emission will be even more strongly inhibited. Also, the depletion layer thickness required to stop protons of a given energy contains three times fewer nuclei than in the case of silicon. The result is, as we have found qualitatively, germanium counters produce a lower background from a flux of fast neutrons compared with a silicon counter of corresponding depletion layer thickness.

All the detectors are sensitive to light, which on absorption liberates carriers in the depletion layer. Photoconductivity is especially marked in surface barrier detectors, owing to the close proximity of the junction to the surface. Tuzzolino *et al.*[55] have studied this effect in some detail and find a photosensitivity of the order of 0·1 μA/microwatt around 4000 Å, uniform over the sensitive area. The sensitivity is such that they were able to use a gold-silicon surface-barrier diode with a scintillation crystal in place of a photomultiplier and achieve comparable resolution for 70 MeV protons.

6. Comparison of Diffused Junction and Surface Barrier Counters

A discussion of the relative merits of the two varieties of junction counter has been attempted in a previous publication[7] but a few further comments may be added.

Both methods give counters of essentially comparable performance, and in the U.S.A. both types are now commercially available. There is no evidence that in general one type of counter deteriorates more readily than the other. Comparative measurements by Buck and Gibson (unpublished) have shown that the sensitivity of the characteristics to a change in ambient vapour conditions for a surface barrier diode is less than that of a bare diffused diode, but equivalent to a diffused diode encapsulated to give protection to the edge of the junction. Experience at Chalk River and elsewhere has indicated that diffused junctions are altered by a relatively low bombardment by heavy ions, while surface barrier detectors are not. The reasons for this are not at present understood.

There are major differences in counter preparation which affect their applications. The diffusion technique is more suitable, at the present stage, for the manufacture of a large number of similar units, and manufacturers are more familiar with this method of diode formation. However, for laboratory preparation the surface barrier is obviously simpler and also more adaptable for special purposes. Here in fact, in our experience, lies the great advantage of the surface barrier detector. Counters can be made rapidly as and when required to meet specified characteristics, which may vary as we have seen over a wide range. By modifying the shape of the mask used during evaporation of the gold single or multiple counters can easily be produced for particular requirements. Some of these variants will be described at a later stage. Such devices could almost certainly be produced by diffusion followed by an etch process after photo-resist masking, but their manufacture would be expensive for small numbers of units. It seems preferable to have available a technique for the construction of such counters in the laboratory, and experience in this is almost certain to suggest other applications of the method.

Both the diffusion and surface barrier processes can be augmented by the lithium-drift method for producing compensated silicon, which allows depletion layer depths of as much as 3–4 mm to be obtained.

7. Comparison of Silicon and Germanium Detectors

The earliest junction particle detectors[17, 1] were of germanium, and although at a time when germanium has been completely superseded by silicon for detector preparation it may appear pointless to compare them, there are a few circumstances in which germanium has important advantages which should not be overlooked.

The great disadvantage in the use of germanium is the necessity for cooling to the region of liquid air temperatures in order to reduce intrinsic conduction. This brings attendant difficulties owing to vapour condensation on the diode surface which alters[31] the characteristics unless there is adequate vapour trapping provided elsewhere. Such effects may account for the generally lower breakdown voltage of germanium detectors.

The zone-refining of germanium is easier than that of silicon, and lower uncompensated donor densities can be achieved. Further, because of the greater stopping power of germanium compared with silicon the same barrier depth will stop particles of about twice the energy. Thus it has been possible to achieve[7] linearity of pulse height with proton energy to over 6 MeV for only 40 V bias, while to give the same performance in n-type silicon would require a resistivity of 10 000 ohm cm.

Germanium detectors are capable of giving faster rise-times than silicon for two reasons. The mobility of carriers is greater in germanium than in silicon, at any given temperature, and at 78° K reaches $1{\cdot}5 \times 10^4$ $cm^2\ volt^{-1}\ s^{-1}$. It is therefore easier to reach the limiting carrier velocity of about 10^7 cm/s with a field of about 2000 volts/cm. Secondly, since the width of depletion layer required to stop a particle of given energy is less in germanium, the carriers are not required to travel as far during collection.

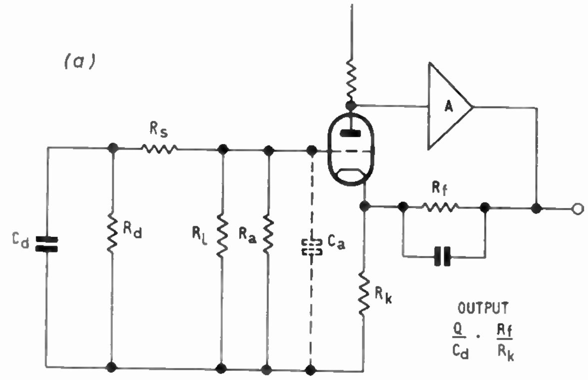
Owing to the higher atomic number of germanium, 32 as compared with 14 for silicon, there is a greater barrier against charged particle emission, reflected in the lower (n,p) cross-sections as mentioned above. Thus the background from neutron-induced reactions within the depletion layer is lower in a germanium detector than in an equivalent one of silicon. There are possible situations, for instance in low-yield polarization experiments, which would render this a considerable advantage.

8. Pulse Amplifiers and Data Handling

The charge liberated by a nuclear particle in a silicon or germanium detector is approximately ten times that which would be produced in a gaseous ionization chamber. If full advantage is to be taken of the signal-to-noise ratio generally obtained a particularly low-noise amplifier is required. Over recent years scintillation detectors or gas-filled counters with avalanche multiplication have been more commonly employed than ion chambers, with the result that there has not been a very wide demand for low-noise amplifiers. The potentialities of the new detectors are leading to renewed developments in this field.

There are two types of preamplifier circuit which may be used, termed the "voltage sensitive" and "charge sensitive" configurations. The former is characterized by a low input capacitance C_a and very high input impedance R_a to provide a fast rise time from a detector with capacitance C_d in series with some internal resistance R_s. There is also a diode leakage resistance R_d. Referring to Fig. 7 (a) the pulse rise time is governed by the time taken to transfer charge from the detector to the input tube grid through the series circuit of C_d, C_a, and R_s. C_a is generally less than C_d and then the rise is R_sC_a, which for $R_s \sim 100$ ohms is of the order 10^{-9} seconds. This is of the same order as the charge collection times within the detector, but it must be stressed that for fast rise-time applications R_s must be kept as small as possible by minimizing the thickness of undepleted silicon, especially if this should have high resistivity.

In the charge sensitive or integrating amplifier[56] the input capacitance is rendered very large by capacitive feedback to the input grid. If the gain is A the input capacitance is $(A+1)C_f \gg C_d$ and the output signal is approximately Q/C_f, independent of the detector capacitance. As this capacitance is a function



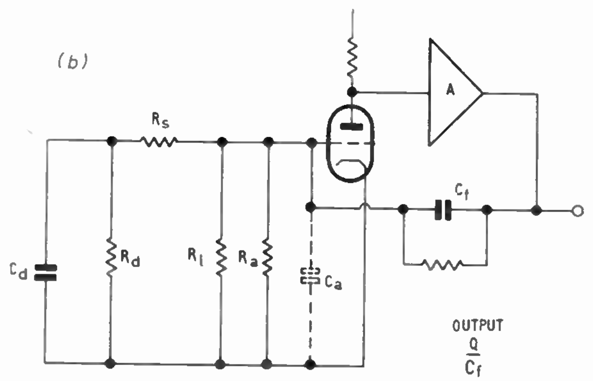


Fig. 7. Input circuit parameters for (*a*) a voltage-sensitive and (*b*) a charge-sensitive amplifier configuration.

of the detector bias it proves very convenient in practice to be able to obtain with the charge sensitive amplifier an output almost independent of bias. The pulse rise-time is now R_sC_d since the input capacitance is so high, and for a large detector at low bias this could easily approach 1 μs. This feature must be considered before the charge-sensitive amplifier is used for fast pulse requirements.

The performance as regards amplifier noise is essentially the same for these two configurations, and arises principally from fluctuations in the input tube anode current ("shot noise") and in the grid current (grid current noise). Assuming equal differentiation and integration time constants, τ, a situation close to the optimum for signal-to-noise ratio, it can be shown[57] that shot noise is proportional to $1/\tau$ and grid current noise is proportional to τ. Thus an optimum setting of τ can be chosen, dependent upon input tube characteristics and detector capacitance, so as to minimize the sum of these noise contributions (see Fig. 8). This value is generally between 1 and 10 μs for typical tubes and input capacitance around 50 pF. Another factor to be considered with detectors containing excessive trapping centres is the slow-collection component due to temporarily-trapped carriers. The resolution of such counters is improved by adopting larger time-constants which allow a greater overall collection efficiency.

Throughout this treatment we have assumed the input tube to be a triode and thus free from partition noise which is the cause of the inferior performance of multi-electrode tubes. In order to achieve sufficient gain from the input stage a cascode arrangement of two triodes is preferred, and charge-sensitive cascode circuits have been described by Blankenship[58] and Chase.[59]

It is often necessary to operate the detector at a relatively high bias, in order to obtain a sufficiently deep depletion layer and under these conditions the detector noise will probably exceed the input tube grid current noise. Then an optimum time constant must be chosen to compromise between shot noise and detector noise, and this optimum will move to shorter time constants as the detector noise increases. This is seen from Fig. 8. A theoretical treatment of this has been given by Goulding,[33] but in general the situation is complicated by the presence of a number of types of detector noise arising from bulk and surface leakage. Goulding has attempted to eliminate surface noise effects by a guard-ring structure.

Owing to the higher base current of transistor circuits the noise level is an order of magnitude greater than for vacuum tube amplifiers. However, the small size and low power consumption of the transistor amplifier may be of prime importance, for instance in space research. For detector capacitance greater than about 1000 pF it is no longer possible to optimize the conditions as above, and the transistor amplifier gives the better signal-to-noise ratio.

In order to make use of the fast rise-time of the detectors an amplifier of high band width is required, but the signal-to-noise ratio and stability of gain are much inferior to the slow amplifiers described here.

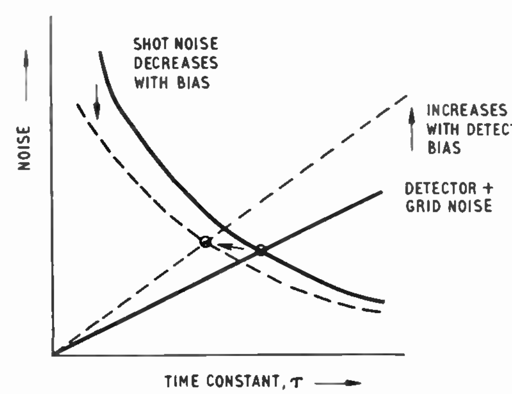


Fig. 8. Variation of noise contributions with amplifier time constants. It can be seen that the optimum time constant setting decreases with increasing detector bias.

It has been shown elsewhere[7] that it is possible to use a combination of fast and slow amplification so that fast coincidence measurements can be made simultaneously with high resolution pulse height analysis. With an A.E.R.E. type 2002A 200 Mc/s amplifier coupled in parallel with a type 1430 2 Mc/s amplifier to the same detector, output pulses rising in less than 6×10^{-9} seconds were obtained from the fast output and a resolution of 50 keV achieved in the slow output spectrum.

Many new problems arise in the handling of the data which can be obtained with semiconductor detectors. For example, an energy resolution of 20 keV or better requires at least 500 to 1000 channels for the pulse height analysis of a spectrum which extends over 5 MeV. If a bias is set for pulse height discrimination a stability of 0·1% may be required over long periods and for fluctuating pulse rates. Further difficulties arise in the study of weak particle groups in the presence of a high intensity of other groups since pulse pile-up causes loss of resolution, while it is impossible to clip the pulses very short without upsetting the optimum signal-to-noise conditions. Many detectors could be operated simultaneously, to obtain at once the whole angular distribution of reaction products. But although the detectors are compact and cheap the low noise amplifiers are not, and there remains the difficulty of storing the large amount of data which they could provide. One system, under development at A.E.R.E., is by means of a multi-track magnetic tape on which digital pulses are stored to represent the pulse height, detector position and any other variable. The tape can subsequently be replayed to enable display of the spectrum in any one counter. Disadvantages are the somewhat slow overall counting rate and the lack of immediate display. The high cost of operation of modern particle accelerators justifies considerable attention to methods of easing and speeding the handling of experimental data. We may foresee the use of electronic computers for the immediate reduction of the mass of data to a form which can be readily assessed and stored. Compared with the speed of the rest of the system magnetic tape appears limited and a more direct input method to the computer would be required. Multi-dimensional pulse-height analysers offer an alternative method of data accumulation, but the problem arises of reading out and storing the data in a convenient form. A method which has not, perhaps, received sufficient attention is by direct photography of a digital display. This is simplified by the newly developed cathode-ray tubes with a fibre-optic face plate, such as the Dumont K2160. Read-out of a very large amount of information could be carried out in seconds, and advantages of the storage of data in photographic form are that it remains visually assessable throughout, and parallel access is feasible in contrast to magnetic or paper tape. Conversion to sequential pulse form could be made by an iconoscope arrangement.

9. Applications of Semiconductor Counters in Nuclear Physics

Over the past two years the availability of semiconductor counters has revolutionized work with charged particles of energies up to a few tens of MeV.

Their excellent energy resolution enables one to be used in place of a magnetic spectrometer for many experiments on charged particle reactions and scattering. The whole energy spectrum can be stored simultaneously instead of by slow scanning which involves variation of the magnetic field. The linearity of response makes calibration simple. The compact size of semiconductor detectors allows them to be mounted very easily so as to be rotatable about a target, and at angles close to that of the incident beam. This region has not been accessible for magnetic analysis so that little high resolution data exist, but many reactions show interesting behaviour at these large angles. In an excellent account of the relative merits of semiconductor and other counters, D. A. Bromley[60] has described an annular silicon detector through which the particle beam can be directed, to give symmetrical detection at 180 deg. Such an arrangement brings great simplification in the analysis of angular correlation experiments.

A number of detectors may be used for simultaneous measurements at different angles in cases where at present there is a very high wastage of information in the use of only a single counter. The problems of handling the amount and precision of data which can be obtained with many detectors are, however, considerable, and must be given more attention.

The thickness of the sensitive region of the counter can be varied simply by changing the bias, and this enables discrimination between particles of different range, without as much effect on resolution as is caused by stopping foils. The variation which can be made is continuous, and is easily carried out by remote control, a very convenient feature when the radiation flux prevents access to the counter.

The response to background neutrons and gamma-rays is relatively low, though fast neutrons over 5 MeV can cause difficulties owing to reactions induced in the silicon. Damage by over-irradiation is almost confined to experiments in the intense flux within a reactor, but is then a severe limitation.

Junction detectors have inherently a very fast rise-time, but it is somewhat difficult to take advantage of this owing to the small size of the detector output. However, simultaneous fast and slow amplification is possible enabling fast timing together with high

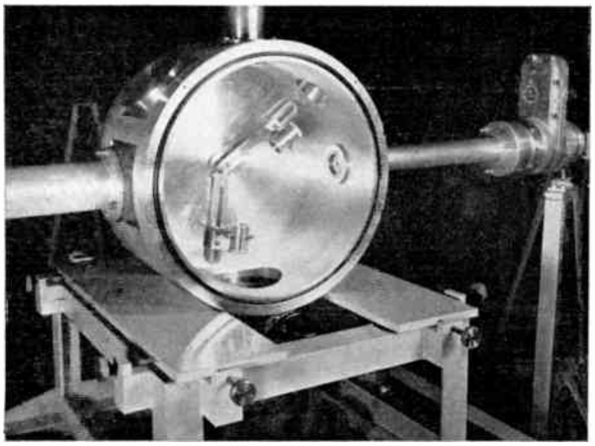


Fig. 9. A scattering chamber designed for use with semiconductor detectors. Two detectors can be set independently at any angle in a plane. The diameter of the chamber is 18 in.

resolution, and this powerful combination remains to be exploited. Even simpler arrangements make certain experiments feasible for the first time. For example, particle gamma-ray coincidence measurements are being made at A.E.R.E. by M. A. Grace to study certain closely spaced levels. Previously the required resolution was obtainable only with magnetic analysers with far too small an acceptance angle for the coincidence rate to be adequate. Similar requirements are satisfied in the application to Coulomb excitation experiments.

A reaction chamber constructed at A.E.R.E. for measurements with a pair of semiconductor counters, which may be set at any required angle, is shown in Fig. 9. The wealth of information which results when a thin target containing various light elements is bombarded in such a chamber is demonstrated in Fig. 10.

Another series of experiments which has become possible with the new detectors is that involving low energy particles, especially alphas and heavier ions. Since the detectors are essentially windowless, i.e. there is usually a negligible energy loss on entering the depletion layer, and as the noise level is low, particles can be detected down to an energy of 50–100 keV. This feature has been valuable for some alpha-scattering experiments on Li^6 below 1 MeV. (Dearnaley and Gemmell, to be published.)

Most applications up to the present have been in the detection of heavy particles such as protons, deuterons and alphas in experiments with particle accelerators or radioactive sources. A spectacular instance has been the discovery of element 103, by Ghiorso *et al.* at Berkeley.[61] The new element named lawrencium, resulting from the capture of boron by californium, was detected by its alpha emission spectrum observed in a silicon counter.

Other applications of the detectors include studies of the fission process by Melkonian[62] and Gooding at A.E.R.E., Miller at Brookhaven and Joyner and others[42] at Oak Ridge. In these experiments a thin film of fissile material such as U^{235} oxide is deposited on the detector surface and fission fragments which enter the counter are detected. With two counters facing each other pairs of fragments can be detected in coincidence. G. D. James is at present carrying out measurements at A.E.R.E. of fission cross-sections in a number of isotopes by a time of flight method using an array of surface-barrier detectors.

Tuzzolino and co-workers[63] at Chicago have designed equipment for the study of primary cosmic ray particles in the energy range 1 to 100 MeV and employing a variety of semiconductor counter arrangements.

Groups at Oak Ridge, Chalk River, Berkeley, University of Yale, and the University of Manchester are making use of silicon counters in experiments with heavy ions.[64] The short range of these particles makes it simple to obtain adequate depletion layer depth even at 100 MeV and above. Bromley[65] has described a particle-separation system for this work consisting of a gas proportional counter and silicon junction detector for simultaneous dE/dx and E (rate of energy loss and energy) measurements. By gating

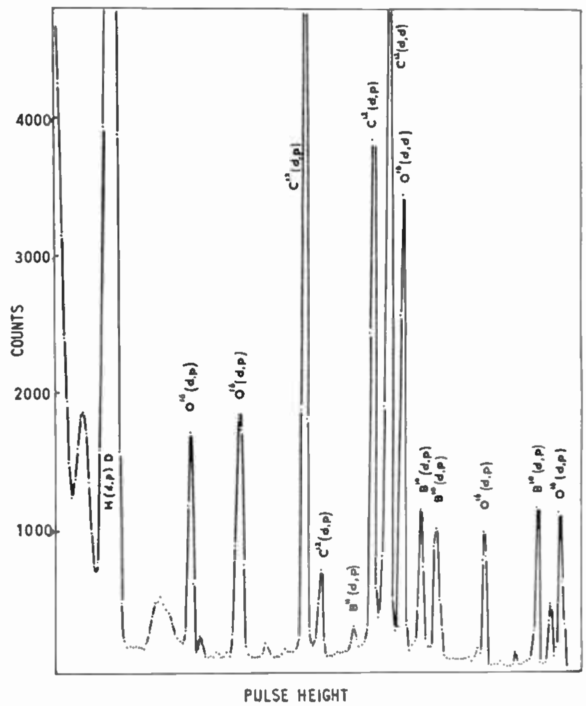


Fig. 10. A charged particle spectrum obtained in the bombardment of a thin target containing carbon, oxygen and boron with 5 MeV deuterons.

on suitable combinations of pulse-heights from the two counters separate detection of isotopes such as C^{11} and C^{12} can be achieved. A similar system at A.E.R.E. is nearing completion for work with lithium ions accelerated in the 5 MeV Van de Graaf generator.

The response of junction counters to electrons of up to 1 MeV has been studied by McKenzie and Ewan[66] and the efficiency at lower energies is very high. An energy resolution of 7 keV was achieved for 55 keV electrons. This performance together with the compact size of the detectors suggests a number of possibilities in radiobiology and medical work with β-emitting tracers. Some specialized detectors for this field are being prepared in collaboration with Dr. G. A. Dissanaike of the Radiotherapeutics Department, Cambridge. The semiconductor detectors may also have applications in counting x-rays, but for these and for higher energy β-particles the lithium-drifted junction counters have the advantage of a much greater sensitive depth.

Particles of high energy lose only a fraction of their energy in the depletion layer and hence do not give a linear response with energy. J. W. Mayer[28] has, however, pointed out that with a lithium-drifted junction giving a depletion layer 3 mm thick the particles may be directed parallel to the junction to give almost any length of counter. By this means protons of energy over 100 MeV could be detected linearly at the expense of angular acceptance. There are many experiments with cyclotrons at this energy for which this would be valuable. At higher energies still counters behave as dE/dx detectors, and particles in the minimum ionization region all give an energy loss of about 400 keV/mm in silicon. The pulses are therefore of small size and require considerable amplification, which leads to difficulties due to electrical pick-up and microphony near large accelerators. There is a spread in pulse height of about 30%[67] due to statistical fluctuations in this small energy loss, the Landau effect.' Bulk conduction counters seem to hold more promise for future applications in this field.

Finally, there are a number of applications which have made use of modified forms of junction counter, and together they illustrate the versatility of the surface barrier technique.

9.1. *Silicon* dE/dx *Detector*

It is possible to make a very thin transmission detector in which a particle loses in general only a small fraction of its energy. The size of this fraction enables discrimination between different types of particle of the same energy, in simultaneous dE/dx and E measurements. Another possibility with thicker wafers of about 0·5 to 1 mm thickness is the construction of a stack of counters in series so that linear response can be obtained up to energies no longer limited by the depth of a single depletion layer.

Some time ago at A.E.R.E., A. B. Whitehead and the author prepared such counters 0·005 cm thick in which complete collection throughout the slice was obtained at only 3 volts bias. These detectors were prepared by evaporating gold on to both faces, and suffered from excessive leakage current and noise due to injection of carriers from the positive electrode. Since then, P. T. Andrews[27] at Liverpool has demonstrated the low injecting properties of an evaporated contact of aluminium, magnesium or beryllium, and has prepared a number of similar counters which give an excellent and reproducible performance. H. Wegner[68] at Los Alamos has also achieved comparable results with a thin diffused junction counter, but the performance of this detector is at present by no means easily reproducible.

9.2. *Semiconductor Fast Neutron Detectors*

If some or all of the energy of a fast neutron is transferred to one or more charged particles these can be detected in semiconductor counters which in this way are suitable for neutron spectroscopy or flux monitoring. Suitable reactions for neutron detection are $Li^6\,(n, \alpha)\,H^3$, $B^{10}\,(n, \alpha)\,Li^7$, $He^3\,(n, p)\,H^3$ and the proton recoil scattering process $H^1\,(n, n)\,H^1$. If a thin layer of one of the first three isotopes is exposed to neutrons the two resulting charged particles can be detected in coincidence in a pair of closely-spaced detectors. In the proton-recoil method only one charged particle results, so that a single detector is sufficient. The first method has been described by Love and Murray[69] at Oak Ridge, who utilized the $Li^6\,(n, \alpha)$ reaction. Comparative considerations by

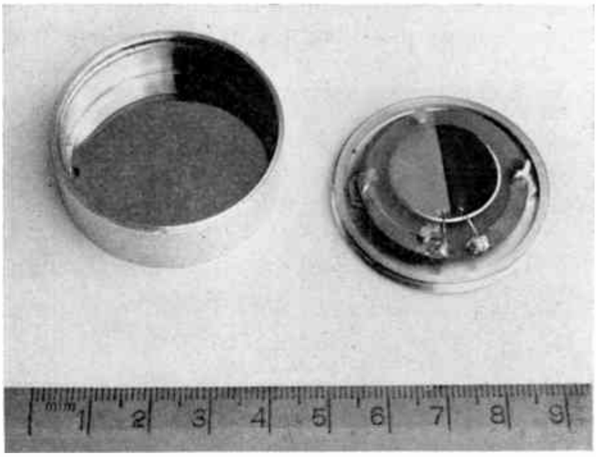


Fig. 11. Semiconductor proton-recoil fast neutron counter. Neutrons eject protons from the polyethylene which covers one of the two semi-circular counters. This counter detects the protons plus background while the other uncoated counter measures background only.

A. T. G. Ferguson and the author led to the conclusion that the best method for neutron spectroscopy is by means of the He^3 (n, p) process, while for flux monitoring, as distinct from energy measurements, the proton recoil method gives the greatest efficiency. An account of the construction and performance of such devices has been published.[52] Figure 11 shows as an example a proton-recoil counter in which two semi-circular surface barrier detectors have been prepared by evaporation on to a single disc of silicon. One is coated with a thin foil of polyethylene which contains an accurately known number of hydrogen atoms. The other detector is bare and enables simultaneous measurement of the background due to the reactions induced in the silicon, and to electron pulses from any background gamma-ray flux. Subtraction of the spectra yields the true proton recoil spectrum. With mono-energetic neutrons the flux can be calculated, while for a complex spectrum the device can be used as a monitor, giving an output proportional to the total flux. This instrument is particularly suitable for the neutron energy range 2 to 6 MeV, which is not adequately covered by present flux monitors.

9.3. *Detector for Polarization Studies*

The investigation of the polarization of the products of a nuclear reaction is rapidly becoming more widely used as a particularly sensitive method for determining nuclear properties. The usual arrangement is to measure the asymmetry produced in a second scattering of the reaction products by a material for which the polarization properties are well-known, such as helium or carbon. Detectors are placed in the scattering plane and symmetrically to left and right of the second scatterer, or "analyser". At A.E.R.E., D. Mingay has been carrying out a series of such experiments, and in collaboration with the author the triple silicon counter shown in Fig. 12 was constructed.

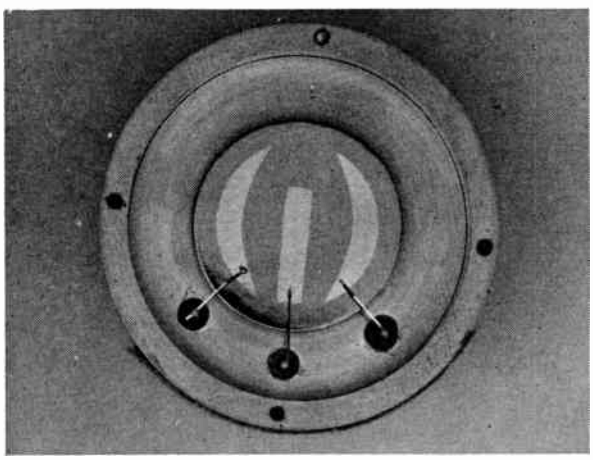


Fig. 12. A triple counter for measurements on the polarization of charged particles from nuclear reactions.

This is placed only 1–2 cm behind an analysing foil of carbon so that the two side counters detect particles scattered through 40–60 deg, while the centre counter acts as a monitor. The detectors are self-collimating and give a relatively low background from the intense neutron and gamma-ray from the initial reaction, in this case C^{12} (d, p). The device has proved far superior to any other type of detector so far employed in this important field.

9.4. *Multiple Detectors*

In experiments with magnetic spectrometers the position of the detected particle on the focal plane of the instrument determines its momentum with high accuracy, so that energy resolution is of less importance than spatial resolution. For this type of experiment it is possible to construct a multiple detector by the evaporation of gold on to an etched silicon crystal through a comb of fine wires. Separate contacts are made to the end of the strips with fine leads in the usual way. The author has prepared such units at A.E.R.E. with a separation of only 0·1 mm between detectors of width 0·9 mm which operated without interaction when used in the normal counting circuit. W. C. Parkinson (University of Michigan) has prepared similar detectors from many small bars of semiconductor. A severe problem arises of how to handle the output from such an arrangement. Parkinson has used separate transistor amplifiers for each detector, while Bilaniuk[70] has successfully applied a method in which a delay line is formed by the capacitance of the detectors linked by small inductances. A fast amplifier at each end of the line gives the difference in time of arrival of the pulses at the two ends, and therefore the location of the counter in which they originated. Bilaniuk claims that such a method is suitable for use with up to 100 detectors, and as such would supplement the nuclear emulsions at present employed with broad-range spectrographs and give an immediately available response.

10. Conclusions

Semiconductor detectors of the junction type have already been shown to possess considerable advantages mainly in the field of low energy nuclear physics, in which any new experiment with charged particles should not be designed without full consideration of the possibilities they afford. In turn this has influenced experimental equipment and electronic apparatus, particularly the design of amplifiers. Further problems such as those of data handling have not yet been pursued sufficiently to take advantage of all that the new detectors can do. The flexibility of the method for making surface barrier counters and its simplicity should lead to a wide variety of applications. One purpose of this paper has been to bring out some of the possibilities of this technique and to encourage its use more widely.

Bulk conduction counters are at present at an earlier stage of development, but in many ways are simpler to understand than the junction. The possibility of precision measurements with gamma-rays and high energy particles which they offer makes it very important that they be given due attention.

The study of semiconductor detectors and their response to nuclear radiations is proving a valuable new method for carrying out experiments in the solid-state physics of semiconductors and in radiation damage. Advances in such knowledge and in the technology of semiconductor preparation will in turn enable new developments to be made in the field of radiation detectors.

11. References

1. F. J. Walter, J. W. T. Dabbs and L. D. Roberts, "Large area germanium surface-barrier counters", *Rev. Sci. Instrum.*, **31**, p. 756, 1960.
2. J. M. McKenzie and J. B. S. Waugh, "Silicon junctions as particle spectrometers", *Trans. Inst. Radio Engrs* (*Nuclear Science*), **NS-7**, p. 195, 1960.
3. J. L. Blankenship and C. J. Borkowski, "Silicon surface-barrier nuclear particle spectrometer", *Trans. I.R.E.*, **NS-7**, p. 190, 1960.
4. G. Amsel, P. Baruch and O. Smulkowski, "Detecteur de particules lourdes à jonction *n-p* au silicium", *Nuclear Instrum. Methods*, **8**, p. 92, 1960.
5. J. W. Mayer, "Performance of Ge and Si surface barrier diodes as alpha particle spectrometers", *J. Appl. Phys.*, **30**, p. 1937, 1959.
6. G. Dearnaley and A. B. Whitehead, "The Solid-state Surface Barrier Charged Particle Detector". A.E.R.E. report R-3437, 1960.
7. G. Dearnaley and A. B. Whitehead, "The semiconductor surface barrier for charged particle detection", *Nuclear Instrum. Methods*, **12**, p. 205, 1961.
8. P. J. Van Heerden, "The Crystal Counter", (North-Holland Publishing Co., Amsterdam, 1945).
9. A. G. Chynoweth, "Conductivity crystal counters", *Amer. J. Phys.*, **20**, p. 218, 1952.
10. S. M. Ryvkin, "Secondary and through photocurrent in semiconductors", *Soviet Physics—Technical Physics*, **1**, p. 2357, 1956.
11. J. D. Van Putten and J. C. Vander Velde, "Homogeneous solid state ionization chamber", *Trans. I.R.E.*, NS-8, p. 124, 1961.
12. P. E. Gibbons and D. C. Northrop, "A silicon gamma-ray spectrometer", *Nature*, **188**, p. 803, 1960.
13. O. Simpson, "Bulk conduction counters in silicon and Ga As", *Proc. Nuclear Instruments Symposium, Harwell*, September 1961. (To be published.)
14. W. R. Harding, C. Hilsum, M. Moncaster, D. C. Northrop and O. Simpson, "Gallium arsenide for gamma-ray spectroscopy", *Nature*, **187**, p. 405, 1960.
15. J. W. Mayer, "Pulse formation in semiconductor detectors", *Proc. Asheville Conf.*, NAS-NRC publication, **871**, p. 1, 1961.
16. W. J. Beyen, "Remarks on photoconductivity and gold-doped silicon", *Proc. Asheville Conf.*, NAS-NRC publication, **871**, p. 121, 1961.
17. K. G. McKay, "A germanium counter", *Phys. Rev.*, **76**, p. 1537, 1949.
18. F. J. Walter, J. W. T. Dabbs, L. D. Roberts and H. W. Wright, "A Study of Germanium Surface Barrier Counters". O.R.N.L. report CF 58-11-99, 1958.
19. L. Koch, "*n-i-p* structure as detectors of nuclear radiation", *Proc. Asheville Conf.*, NAS-NRC publication, **871**, p. 51, 1961.
20. E. M. Pell, "Ion drift in an *n-p* junction", *J. Appl. Phys.*, **31**, p. 291, 1960.
21. J. Benveniste, R. Booth and A. C. Mitchell, "Observations on the Response of a Lithium-drifted Detector to Protons", UCRL report 6441, 1961.
22. P. F. Donovan, "Paint-on particle detectors", *Proc. Asheville Conf.*, NAS-NRC publication 871, p. 268, 1961.
23 F. J. Biondi, (ed.) "Transistor Technology" Vol. 3 (Van Nostrand, New York, 1958).
24. T. M. Buck, "Surface effects on silicon particle detectors", *Proc. Asheville Conf.*, NAS-NRC publication 871, p. 111, 1961.
25. R. J. Archer, "Optional measurement of film growth on silicon and germanium surfaces in room air", *J. Electrochem. Soc.*, **104**, p. 619, 1957.
26. J. L. Blankenship, C. J. Borkowski and R. J. Fox, "Silicon surface barrier nuclear particle detectors", *Proc. Internat. Conf. Nuclear Electronics, Belgrade 1961.* (To be published.)
27. P. T. Andrews, "Thin silicon surface barrier counters", *Proc. Nuclear Instruments Symposium, Harwell*, September 1961. (To be published.)
28. J. W. Mayer, N. A. Baily and H. L. Dunlap, "Characteristics of ion drifted *p-i-n* junction particle detectors", *Proc. Conf. Nuclear Electronics, Belgrade*, May 1961. (To be published.)
29. J. H. Elliott, "Thick junction radiation detectors made by ion drift", *Nuclear Instrum. Methods*, **12**, p. 60, 1961.
30. M. H. Sullivan and J. H. Eigler, "Electrodeless nickel plating for making contacts to silicon", *J. Electrochem. Soc.*, **104**, p. 226, 1957.
31. H. Statz, "Semiconductor surface effects", *Proc. Asheville Conf.*, NAS-NRC publication 871, p. 99, 1961.
32. C. T. Sah, R. N. Noyce and W. Shockley, "Carrier generation and recombination in *p-n* junctions", *Proc. Inst. Radio Engrs*, **45**, p. 1228, 1957.
33. F. S. Goulding and W. L. Hansen, "Leakage current in semiconductor junction radiation detectors and its influence on energy-resolution characteristics", *Nuclear Instrum. Methods*, **12**, p. 249, 1961.
34. E. Fairstein, "Considerations in the design of pulse amplifiers for use with solid state radiation detectors", *Trans. I.R.E.*, **NS-8**, p. 129, 1961.
35. E. Baldinger and W. Czaja, "On the energy expended per electron-hole pair produced in *p-n* junction detectors", *Nuclear Instrum. Methods*, **10**, p. 237, 1961.
36. W. Shockley, *Proc. Internat. Conf. Semiconductors, Prague*, July 1960. (To be published.)
37. J. Tauc, "Electron impact ionization in semiconductors", *J. Phys. Chem. Solids*, **8**, p. 219, 1959.

38. V. S. Vavilov, "On photo-ionization by fast electrons in germanium and silicon", *J. Phys. Chem. Solids*, **8**, p. 223, 1959.

39. W. Czaja, "Zur Theorie der Arbeit pro Elektron-Loch-Paar in Halbleitern de IV Gruppe", *Helvetica Physica Acta*, **34**, p. 760, 1961.

40. G. L. Miller and W. M. Gibson, "Charge collection in semiconductor particle detectors", *Proc. Internat. Conf. Nuclear Electronics, Belgrade*, May 1961.

41. W. Shockley and W. T. Read, "Statistics of the recombination of holes and electrons", *Phys. Rev.*, **87**, p. 835, 1952.

42. W. T. Joyner, H. W. Schmitt, J. H. Neiler and R. J. Silva, "Energy spectra of correlated fragment pairs from the spontaneous fission of Cf^{252}", *Trans. I.R.E.*, **NS-8**, p. 54, 1961.

43. G. L. Miller, *Proc. Asheville Conf.*, NAS-NRC publication 871, p. 27, 1961.

44. J. Rhodes, W. Franzen and W. E. Stephens, "Ionization measurement of B^{10} (n, α) Li^{7}", *Phys. Rev.*, **87**, p. 141, 1952.

45. T. E. Cranshaw and J. A. Harvey, "Measurement of the energies of alpha-particles", *Canadian J. Research*, **26A**, p. 243, 1948.

46. J. C. D. Milton, A. R. Rutledge and P. I. K. Lennox, Chalk River (Canada) Report No. CRP–632, 1956.

47. H. C. Schweinler, "Energy loss of moving charged particles in a valence or ionic crystal", *Proc. Asheville Conf.*, NAS-NRC publication 871, p. 91, 1961.

48. J. L. Blankenship, "Silicon Diode Barrier Equation Nomograph", O.R.N.L. LR-45902, 1960.

49. C. T. Raymo and J. W. Mayer, "Transient response of *p-n* junction detectors", *Trans. I.R.E.*, **NS-8**, p. 157, 1961.

50. G. Amsel, P. Baruch and O. Smulkowski, "High resolution study of nuclear reactions by *p-n* junction detectors", *Trans. I.R.E.*, **NS-8**, p. 21, 1961.

51. R. W. Klingensmith, "The effect of a high radiation environment on gold-silicon charged particle detectors", *Trans. I.R.E.*, **NS-8**, p. 112, 1961.

52. G. Dearnaley and A. T. G. Ferguson, "Semiconductor neutron counters", *Proc. Nuclear Instruments Symposium, Harwell*, September 1961. (To be published.)

53. W. Deuchars and G. P. Lawrence, "Interaction of 14 MeV neutrons with a silicon semiconductor nuclear particle detector", *Nature*, **191**, p. 995, 2nd September 1961.

54. R. J. Howerton, "Tabulated Neutron Cross-Sections". U.C.R.L. Report No. 5226, October 1959.

55. A. J. Tuzzolino, E. L. Hubbard, M. A. Perkins and C. Y. Fan, "Photoeffects in silicon surface-barrier diodes". (To be published) 1961.

56. C. Cottini, E. Gatti, G. Giannelli and G. Rossi, "Minimum noise preamplifier for fast ionization chambers", *Il Nuovo Cimento*, **3**, p. 423, 1956.

57. A. B. Gillespie, "Signal, Noise and Resolution in Nuclear Counter Amplifiers". (Pergamon, London, 1953.)

58. J. L. Blankenship and C. J. Borkowski, "Performance of silicon surface barrier diode detectors", *Trans. I.R.E.*, **NS-8**, p. 17, 1961.

59. R. L. Chase, W. A. Higinbotham and G. L. Miller, "Amplifiers for use with *p-n* junction radiation detectors", *Trans. I.R.E.*, **NS-8**, p. 147, 1961.

60. D. A. Bromley, "Semiconductor detectors in nuclear physics", *Proc. Asheville Conf.*, NAS-NRC publication 871, p. 61, 1961

61. A. Ghiorso, T. Sikkeland, A. E. Larsh and R. M. Latimer, "New element, lawrencium, atomic no. 103", *Phys. Rev. Letters*, **6**, p. 473, 1961.

62. E. Melkonian, "Application of surface barrier solid state ionization chambers to measurements on fission", *Nuclear Instrum. Methods*, **11**, p. 307, 1961.

63. R. Takaki, M. Perkins and A. Tuzzolino, "A gold-silicon surface barrier proton range telescope", *Trans. I.R.E.*, **NS-8**, p. 64, 1961.

64. M. L. Halbert, J. L. Blankenship and M. V. Goldman, "Response of surface-barrier semiconductor counters to N^{14} ions", *Bull. Amer. Phys. Soc.*, **5**, p. 38, 1960.

65. D. A. Bromley, *Proc. Nuclear Instruments Symposium, Harwell*, September 1961. (To be published.)

66. J. M. McKenzie and G. T. Ewan, "Semiconductor electron detectors", *Trans. I.R.E.*, **NS-8**, p. 50, 1961.

67. G. L. Miller, B. M. Foreman and L. C. L. Yuan, "Applications of solid state detectors to high energy physics", *Trans. I.R.E.*, **NS-8**, p. 73, 1961.

68. H. E. Wegner, "dE/dx and E semiconductor detector systems for 25 MeV He^{3} and alpha particles", *Trans. I.R.E.*, **NS-8**, p. 103, 1961.

69. T. A. Love and R. B. Murray, "The use of surface-barrier diodes for fast-neutron spectroscopy", *Trans. I.R.E.*, **NS-8**, p. 91, 1961.

70. O. M. Bilaniuk, A. K. Hamann and B. B. Marsh, "A multichannel Unit of Solid-state Counters for Nuclear Spectroscopy", University of Rochester report AT (30–1) 875, May 1960.

Manuscript first received by the Institution on 8th November 1961 and in final form on 23rd January 1962. (*Paper No. 749.*)

U.D.C. 621.396.21

A Method of I.F. Switching for a Microwave Diversity System

By

D. R. BESTER, B.Sc.
(*Graduate*)†

Summary: This paper deals with the various factors involved when considering the use of diversity on a microwave radio link. Various systems of i.f. switching are briefly surveyed. An i.f. switching circuit designed for use on such systems is described which uses high speed diode switches working at the i.f. of 70 Mc/s. The switch control circuits are partially transistorized.

1. Introduction

Wideband microwave radio links are now used extensively to carry telephone channels, television signals, teleprinter and high speed data signals from one point to another. These links operate with carrier frequencies in the 2000 Mc/s, 4000 Mc/s and 6000 Mc/s bands. Transmissions are limited to line of sight paths because of the high frequencies used, and the transmitting and receiving stations are usually about 30 to 40 miles apart, the actual distance depending upon topography and the availability of sites. A complete system consists of terminal stations and unattended repeater stations.

When such a system is used to carry telephone channels, the channels are usually combined by a frequency division multiplex (f.d.m.) system. By this means, 60 channels are made to occupy a frequency band from 60 to 300 kc/s, 300 channels occupy a band from 60 to 1364 kc/s, and 960 channels occupy a band from 60 to 4287 kc/s.[1] These bands are commonly called the baseband signal, and this signal is brought to the terminal equipment of the radio system. In general, conversion to and from baseband only occurs at terminal stations. The repeaters usually deal only with r.f. and i.f. carriers, although provision may be made for the insertion of supervisory signals, or channels for the use of maintenance engineers.

Any system may use more than one radio carrier. On small systems two radio carriers, modulated with the same baseband information, are often used; one carrier then acts as a stand-by in case of a failure on the other. On high capacity systems the frequency plan may allow up to eight carriers to be transmitted, each modulated with different baseband information. In this case it is usual to use one or two of the carriers as a stand-by in case of a failure on one of the others. The carriers may be either vertically or horizontally polarized, depending upon the frequency plan for the system as a whole. Where the system uses an intermediate frequency carrier in its radio equipment, the i.f. chosen is 70 Mc/s in accordance with the recommendations of the C.C.I.R.[1] Thus the i.f. switch, described later, deals with carriers at a centre frequency of 70 Mc/s.

2. Factors Affecting the Received Carrier Level

The atmosphere acts as a dielectric whose characteristics alter with temperature, pressure, and moisture content. Since all of these factors may vary along a multi-channel radio link, there may be several distinct transmission paths differing in length. In general, the carriers arriving at the receiver aerial from these paths will add, and will produce a varying carrier level, dependent upon their relative amplitudes and phases. This gives rise to the most common form of fading and, since the relative phases will depend upon relative path lengths and the wavelength of the carrier, this phenomenon is frequency selective. Crawford and Jakes[2] describe work which was done to compare the frequency selectivity of such fading and this work has since been analysed statistically by Kaylor,[3] with a view to using frequency diversity on such links. It was found that this type of fading rarely produced very deep frequency selective fades, unless the general level over the frequency band was depressed by about 10 dB owing to other causes, such as super-refraction or reverse refraction.

2.1. *Refraction*

Normally, the refractive index of the atmosphere decreases as altitude increases. This has the effect of refracting a microwave carrier so that it curves in the same direction as the earth's surface. On a profile map of a radio link, where it is easier to draw the microwave path as a straight line, this effect is allowed for by superimposing all topographical features upon a paper which is corrected to 4/3 of the earth's radius. The ratio, $K = 4/3$, is taken as representing the effect of a normal atmospheric refractive index. However, the atmospheric refractive index changes hour by hour. If it decreases more rapidly with altitude, then the wave curves more towards the earth, and the

† The General Electric Company Ltd., Telephone Works, Coventry, Warwickshire.

effect is considered as equivalent to an earth "flattening". Under these conditions the effective earth's radius increases. In practice a limit of $K = \infty$ is normally taken to allow for these effects.

Under certain conditions, particularly when temperature inversions occur, or when ground fogs are present, the refractive index increases with increase of altitude and the wave curves in the opposite direction to the earth's surface. Referring to a profile map, the earth is considered to "bulge" and it is usual for a value of $K = \frac{2}{3}$ to be taken to represent the extreme of this condition. Under such conditions all clearances over intermediate objects are reduced. The carrier level at the receiver aerial falls (if there is an obstruction), and since this inversion will affect all frequencies, it will produce a general decrease in level over a frequency band. Systems are designed so that, under severe conditions of earth bulging, a workable signal is still received. The various diurnal effects of this type of fading have been dealt with fully by Gough.[4]

If a cold front moves across the path of a microwave link it introduces an atmospheric discontinuity in the path of the carrier. This will possibly cause some reflection of the wave, but the main effect will be one of refraction, and a great part of the carrier may be refracted off course, causing a fall in received level. This effect has been known to cause a complete loss of carrier during the time taken for the cold front to pass across the path, and it may affect the links in a system in succession causing a system-failure for a long period of time.

2.2. *Reflecting Layers*

A horizontal reflecting layer in the atmosphere may reflect some of the signal off course.[2] A severe form of fading may be caused by two such layers acting in a similar way to a waveguide, producing the phenomenon known as "ducting".

2.3. *Interference from Reflected Waves*

The main cause of very deep fades is the addition of the direct carrier with a reflected carrier; these two carriers may be in anti-phase. Moreover if the reflection occurs from a flat surface such as water, the level of the reflected wave may be equal to that of the direct wave. This form of fading can be very severe on a link where a great part of the path lies over water, salt flats, or bare flat terrain. It may also be due to a reflected wave from some building or topographical feature near to the receiver aerial. In general, if the intervening terrain is woody or rough, the reflected wave will be dispersed sufficiently to cause little interference. On certain links the reflected wave may be blocked by some topographical feature, so that no interference takes place.

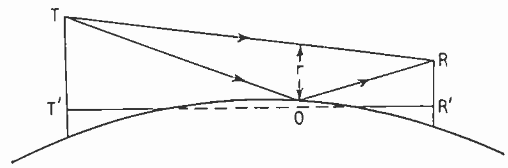


Fig. 1. The effect of a reflected wave.

3. The Need for Diversity

3.1. *Strength of Received Carrier*

When interference occurs from a reflected wave, the strength of the resultant received carrier may be found as follows. Referring to Fig. 1, if the direct path is TR and the reflected path is TOR, let there be a difference in path length δ between them. If it is assumed that perfect reflection occurs at O, it follows that:

(a) The reflected carrier level will be equal to that of the direct wave.

(b) A phase change of 180 deg will occur at O. Owing to the length of the link compared with aerial height, reflection angles are usually of the order of fractions of a degree. At these small reflection angles, which are much less than the pseudo-Brewster angle (approximately 6 deg for water and 14 deg for wooded country) horizontally and vertically polarized waves will have a phase change of approximately 180 deg.

Then the difference in phase of the two carriers is $\frac{2\pi\delta}{\lambda} - \pi$ radians, where λ is the wavelength of the carrier. The carriers add vectorially giving a resultant level E, where E is derived as follows:

Letting $\theta = \frac{2\pi\delta}{\lambda}$ and E' be the amplitude of the direct and reflected waves,

$$E' \sin(\theta - \pi) = E \sin(\theta/2 - \pi/2)$$

hence
$$E' \sin \theta = E \cos \theta/2$$

Therefore
$$E = \frac{E' \sin \theta}{\cos \theta/2}$$
$$= \frac{2E' \sin \theta/2 \cos \theta/2}{\cos \theta/2}$$
$$= 2E' \sin \theta/2$$

Therefore the strength of the received carrier is given by

$$E = 2E' \sin \frac{\pi\delta}{\lambda} \qquad \text{......(1)}$$

3.2. *Fresnel Zones*

Fresnel zones are circular areas surrounding the direct line of sight path of such a radius that the path

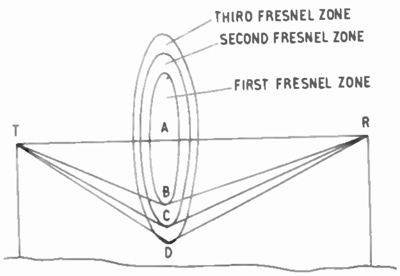


Fig. 2. Fresnel zones.

length from the perimeter is a multiple of half wavelengths longer than the direct path. Figure 2 illustrates this. If TAR is the line of sight path, then TBR represents a path which is half a wavelength longer than TAR. The circle radius AB is called the first Fresnel zone and signals with paths through this zone will tend to reinforce the direct signal.

The path TCR is one wavelength longer than the direct path. The area outside the first Fresnel zone, but within the circle radius AC, is called the second Fresnel zone. Signals which pass through the second Fresnel zone tend to cancel the direct signal. Similarly we may draw third, fourth, fifth Fresnel zones and so on. Since the area of the zones become progressively smaller, only the first few zones are considered of practical importance. In general, signals from the odd zones increase, while those from even zones decrease, the received signal level.

It will be seen from eqn. (1) that E equals $2E'$ when δ equals a half wavelength. This is a maximum value of level and it corresponds to some value of r (Fig. 1) where r is the radius of the first Fresnel zone. If we first consider T and R at some points T′ and R′ where there is no line of sight path, then the received carrier is low, as shown at X on Fig. 3. This received level is due to diffraction of the transmitted wave into the shadow area. As T′ and R′ rise towards T and R the received carrier level rises until a direct grazing path occurs. The free space level does not occur, however, until the clearance is about 0·5 of the radius of the first Fresnel zone.

Now if T′ and R′ are raised again, a direct and reflected wave will occur and the received level will rise to a maximum which occurs when the height of the direct wave above the reflection point is r. If T′ and R′ are raised still further, the reflected path length will be such that the received level falls, until when δ equals one wavelength the direct and reflected rays cancel. When this occurs, the clearance r' is equal to the radius of the second Fresnel zone. When T′ and R′ rise yet further, the received level will rise again to a maximum when δ equals one and a half wavelengths, and the clearance corresponds to the radius of the third Fresnel zone, and so on.

It is clear that the above argument is true if T′ is fixed and R′ is continually raised. That is to say, R′ will pass through consecutive maxima and minima corresponding to first, second, third Fresnel zone clearances, and so on. If two aerials are spaced at the correct vertical distance with respect to each other, one will receive a very high level of carrier while the other may receive none at all.

Now, if the effect of earth bulging and flattening, due to a change in the atmospheric refractive index is considered, it will be seen that both the paths of the direct wave and the reflection point will be affected. As the two path lengths change, the space pattern above the receiver position will alter, and a condition can occur where the aerial which was receiving a high level carrier is in such a position that the received level is zero.

At the same time, the aerial which was receiving no carrier at all may now be in a position to receive the maximum carrier level. When such a path occurs, the advantage of using diversity is obvious, since the aerial receiving the higher level carrier may always be selected.

In a similar manner, the advantage of frequency diversity, using one aerial but two differing carrier frequencies, may be shown. The level of the received carrier depends directly on the wavelength of the carrier; thus by choosing the correct frequencies one can show that, when the difference in path lengths is such that the carriers cancel at one frequency, the carriers will be in phase at the second frequency.

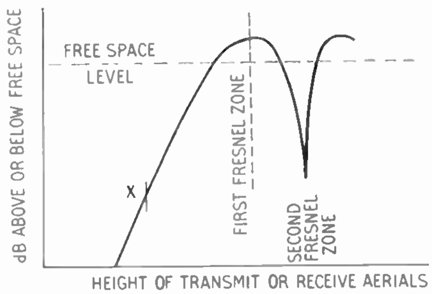


Fig. 3. Variation of received carrier level with height of transmit or receive aerials.

It has been assumed in the previous paragraphs that the aerials will not discriminate between the direct and reflected waves. In the majority of cases this is correct. Although the aerials are designed to be as directional as possible, when path lengths of the order of 35 or 40 miles are considered, the vertical angle between the direct and reflected wave is seldom more than a fraction of a degree, and the aerial cannot differentiate between them. This would not be true in an extreme case where the reflection point is very close to the receiver aerial, and in these rare cases re-orientation of the aerial may have considerable effect.

It is worth noting that the vertical spacing between maxima and minima depends on the path configuration and length. On some paths under extreme conditions this spacing may be very small. Von Hagen, MacDiarmid and Goldenberg[5] report that, on a path across the St. Lawrence River, this spacing was of the order of 10 ft, and in fact a form of space diversity even existed across the aperture of the 10 ft aerial being used.

3.3. *Aerial Spacing for Space Diversity*

Referring to Fig. 4 suppose that

(a) AXB is the reflecting plane of the wave.

(b) H is the height of the transmit aerial above the plane.

(c) h is the height of the receive aerial above the plane.

(d) d is the length of the direct path.

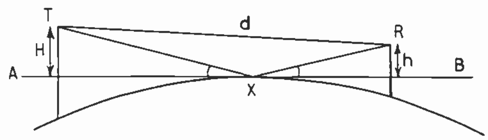


Fig. 4. Reflecting plane of interfering wave.

Then, by deriving equations for the path lengths followed by binomial expansion and approximation,[6] it can be shown that the difference in path length δ is given approximately by

$$\delta = \frac{2Hh}{d}$$

Suppose that the receive aerials have heights h' and h''. Then for the first aerial,

$$\delta' = \frac{2Hh'}{d}$$

$$h' = \frac{d\delta'}{2H}$$

It is required that the second path shall differ by a half-wavelength, therefore

$$h'' = \frac{d(\delta' + \lambda/2)}{2H}$$

Hence $$h'' - h' = \frac{d\lambda}{4H}$$

where all quantities are expressed in the same units. This equation becomes:

$$\text{aerial spacing} = \frac{43{\cdot}4\lambda d}{H} \quad \text{feet}$$

where d is in miles

λ is in centimetres

H is in feet.

It should be noted that H is the height above the reflecting plane, *not* the height of the transmitting aerial above the ground.

4. Diversity Systems

4.1. *Frequency and Space Diversity*

Having decided that a particular path requires a form of diversity one may choose between space and frequency diversity. For frequency diversity, the same baseband information is transmitted over the path on two u.h.f. carriers at different frequencies. The carriers are received on one aerial, separated at u.h.f. by a diplexer, and are passed to their respective frequency changers and i.f. amplifiers. In space diversity two receiving aerials are spaced vertically with respect to each other, by a spacing determined by the equation given in Section 3.3. The received carriers are taken directly to their respective frequency changers and i.f. amplifiers. The carriers in either case may be combined or switched, as shown later.

Frequency diversity has the main disadvantage that it requires two different carrier frequencies to transmit the same information, and hence the number of carriers available in a frequency plan to transmit different baseband information is halved. This restricts the capacity of the system, and in general a system which is working with the maximum number of u.h.f. carriers or which will be required to do so in the future, will use space diversity. Frequency diversity is usually used over a path where the causes of fading are more dependent on frequency, such as fades due to atmospheric refraction etc.[3] On a path where the main cause of fading is the anti-phase wave caused by reflection from water, salt flats etc., one would generally use space diversity.

4.2. *Switching and Combining*

Whichever method of diversity is used, the carriers may be either individually selected by switching, or they may be combined. These procedures may be carried out at u.h.f., i.f., or baseband.

In a switching system, the carrier with the higher level is automatically selected. A switch panel for doing this is described later. A combining system combines the received carriers and, in theory, a gain over a switched system can be achieved. However, the carriers must be combined in phase, and small differences in phase between the carriers must be detected. A typical i.f. combining system would include a network capable of detecting phase difference; the output from this network would then be used to change the phase of one local oscillator.

5. I.F. Switching on a Diversity System

The main advantages of switching over combining at i.f. are:

(a) A switching system is generally comparatively simple. In a combining system the means of detecting phase difference and controlling the phase of one carrier may be complex.

(b) A switching system can be manufactured in the form of separate racks and does not require any modification to existing standard transmitter-receiver racks. The system is not critical with respect to inter-rack cabling.

(c) If carriers are combined when they are considerably out of phase, the resultant distortion may render the system unusable. This could occur with the majority of component faults on a combining system. If a switching system is used, most component faults will merely make the switching inoperative and one carrier will be permanently selected. Thus, the system is still usable although it is reduced to a state where no diversity exists.

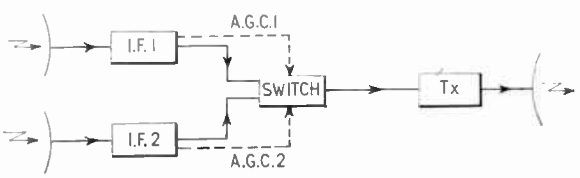


Fig. 5. Dual space diversity block schematic.

5.1. *Space Diversity*

A general method of i.f. switching, on a dual space diversity system, is shown in block schematic form in Fig. 5. The aerials receive carriers at the same frequency, I.F. 1 and I.F. 2 represent the frequency changers and i.f. amplifiers for the two aerials, and they develop a.g.c. voltages depending on the level of carrier fed into them. These a.g.c. voltages are taken to the i.f. switch panel. The i.f. switch selects the carrier corresponding to the more negative a.g.c. voltage, and this carrier is fed to a limiter panel on the transmitter rack. The diagram shows a schematic for a repeater station. If the diversity occurs at a terminal station, the output from the switch is taken directly to a limiter-demodulator panel.

Figure 6 shows the block diagram for a triple space diversity system which would be used on a very poor path producing a wide range of reflection conditions. This time, three aerials are spaced vertically with respect to each other, with different spacing between pairs. Switch 1 compares the carrier levels received by I.F. 1 and I.F. 2, and feeds the higher level carrier to switch 2. The a.g.c. selector compares the a.g.c. voltages of these two amplifiers and feeds

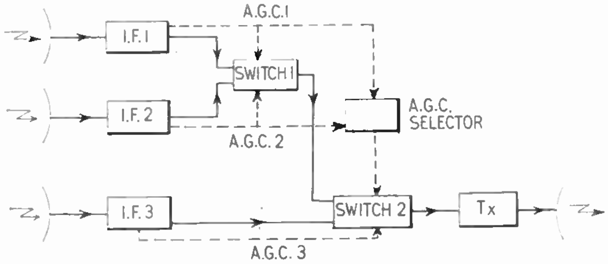


Fig. 6. Triple space diversity block schematic.

the more negative voltage to switch 2. Switch 2 now compares the higher level carrier from aerials 1 and 2 with the carrier on aerial 3, and feeds the selected carrier to the transmitter.

A triple diversity system across the Bay of Fundy, Canada, has been described by Sheffield.[7] This system covers a 49-mile path across a bay which is notorious for its high tides. The General Electric Company have installed a 300-channel triple diversity link along this same path (August 1959) using the method given above. On the G.E.C. link, the aerial heights at Otter Lake are 100, 120 and 180 feet, and those at Aylesford are 65, 83, and 143 feet.

5.2. *Frequency Diversity*

Figure 7 shows a switched frequency diversity system. The two received signals are separated by a diplexer. The levels of the carriers are compared as before, and the higher level carrier is amplified and re-transmitted.

It will be seen that on either system the i.f. switch must

(a) compare the a.g.c. voltages, and

(b) use this information to select the higher level carrier.

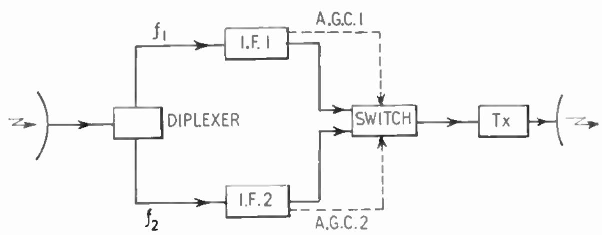


Fig. 7. Frequency diversity block schematic.

6. A Practical I.F. Switching Panel

The following paragraphs describe a practical i.f. switching panel, developed for fast switching on a 960-channel system. The changeover time is not greater than 1 microsecond and the rejection of the unwanted channel is not less than 75 dB.

6.1. *Diode Switch*

The circuit diagram of the diode switch operating at 70 Mc/s, is shown in Fig. 8. When a negative voltage is applied to the d.c. control lines, the diodes D2 and D3 conduct, but D4 is "cut off" by the voltage appearing across D2 and D3 in parallel. At the same time D1 is non-conducting, and a 70 Mc/s carrier input at CX 1 is routed to CX 3 through the diodes D2 and D3. The same direct voltage is applied to the second arm where the diodes are reversed. Thus D8 conducts and switches C across the carrier path, while D6 and D7 offer a high impedance to the carrier, since they are "cut off" by the voltage developed across D8 and R8 in series. The carrier applied at CX 2 is therefore highly attenuated. In order to present a correct termination at CX 2 under these conditions, D5 conducts and switches R6 across the input. The resistor R6 is chosen so that the combination of R6, D5 and R, offers a 75-ohms termination at CX 2.

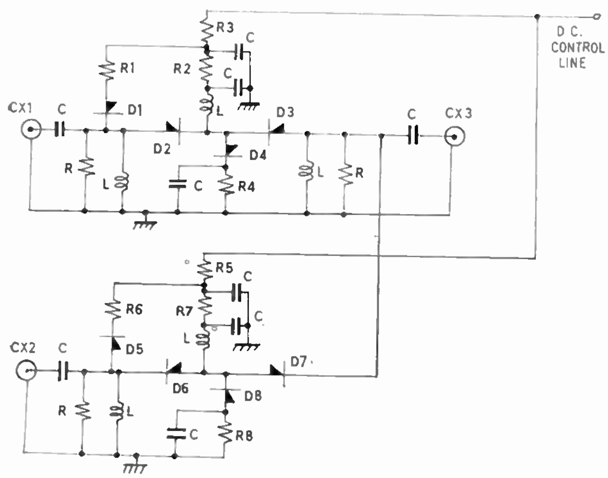


Fig. 8. Schematic diagram of diode switch.

When a positive voltage is applied to the d.c. control line the above conditions are reversed, and a carrier can pass from CX 2 to CX 3 via D6 and D7 conducting. At the same time the combination of D2 and D3 (non-conducting) and D4 (conducting) offer a high impedance to the carrier at CX 1. The diode D1 conducts and terminates CX 1 in 75 ohms.

The inductances L offer a high impedance at 70 Mc/s and act as d.c. paths only. In the strict sense this device is not tuned at 70 Mc/s, although it is obvious that at low frequencies the inductances would not act as high impedances, while at higher frequencies, stray capacitance would attenuate the signal. The switch has a very wide band application and the performances on amplitude response and relative group delay are well in excess of that required at present for 960 channels. The nominal specification for 960 channel operation is that the amplitude response has a variation of less than 0·25 dB, over the frequency band from 57 to 83 Mc/s (70 Mc/s ± 13 Mc/s) while the relative group delay varies by less than 1 millimicrosecond over the frequency band from 61 to 79 Mc/s (70 Mc/s ± 9 Mc/s). The input and output impedances can be controlled to a high degree of accuracy by a suitable choice of R.

All the capacitors marked C are 1000 pF disc type, and offer a very low impedance to the i.f. carrier. The values of R2, 3, 4, 5, 7 and 8 are chosen on d.c. considerations only.

6.2. *Diode Switch Equivalent Circuit* (*Conducting*)

The equivalent circuit for the conducting condition is given in Fig. 9. If R_D represents the resistance of a diode when conducting, R the value of resistance chosen for matching purposes, and Z the value of impedance required at the input and output, it will be seen that the input impedance, Z, is given by

$$Z = \frac{R^2(2R_D+Z)+2R_D ZR}{R^2+R(2Z+2R_D)+2R_D Z}$$

hence

$$ZR^2+R(2Z^2+2R_DZ)+2R_DZ^2 = R^2(2R_D+Z)+2R_DZR$$

and $\quad 2R_DR^2-2Z^2R-2R_DZ^2=0$

Thus
$$R = \frac{2Z^2 \pm (4Z^4+16R_D^2Z^2)^{\frac{1}{2}}}{4R_D}$$

$$= \frac{Z^2 \pm Z(Z^2+4R_D^2)^{\frac{1}{2}}}{2R_D}$$

Substituting the practical values $Z = 75\ \Omega$ and $R_D = 13\ \Omega$, gives a value for R of approximately 440 Ω. In practice the standard value of 430 Ω gives the required match conditions, well within the tolerance limits allowed.

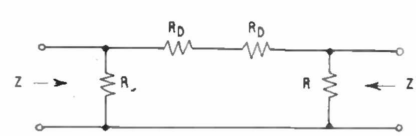


Fig. 9. Equivalent circuit conducting.

The loss on the conducting path may be approximated as follows:

$$V_o \simeq \frac{V_i RZ}{2R_D(R+Z)+RZ}$$

where V_o is the output voltage and V_i is the input voltage. Thus, the forward loss in decibels is approximately given by

$$\text{loss} = 20\log_{10}\frac{(RZ+2R_D(R+Z))}{RZ}\ \text{dB}$$

Substituting the values for Z, R, and R_D, in this equation gives a calculated loss of 2·9 dB. In practice the loss is slightly less than this and is of the order of 2·5 dB. However, it will be seen that the loss depends to a great extent on the value of R_D, and hence on the choice of diode.

It should be noted that C.C.I.R. recommend an output voltage of 0·5 V r.m.s. with an input voltage of 0·3 volts r.m.s. on the panels on the system.[1] Thus if the output of the previous i.f. amplifier is 0·5 V r.m.s., a 3 dB loss on the switch is acceptable, since the following limiter is designed for a 0·3 V input.

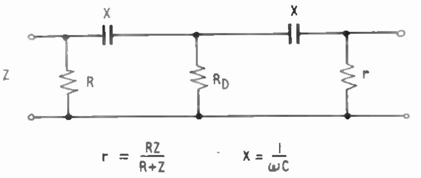


Fig. 10. Equivalent circuit non-conducting.

6.3. *Diode Switch Equivalent Circuit (Non-conducting)*

The equivalent circuit for the non-conducting path is given in Fig. 10. In the non-conducting condition, at 70 Mc/s, the feed across the diode is largely controlled by the effect of diode capacitance and stray capacitances associated with the path. The loss under these conditions is given approximately by the following equations:

$$V_o = V_i\frac{R_D}{\sqrt{X^2+R_D^2}}\times\frac{r}{\sqrt{R_D^2+r^2}}$$

where $r = \dfrac{RZ}{R+Z}$

and $X = \dfrac{1}{\omega C}$ where C is the value of capacitance.

Since X is very much greater than R_D or r, the loss in decibels is given approximately by

$$\text{loss} = 20\log_{10}\frac{X^2}{R_D r}$$

Substituting a capacitance of 1 pF for the capacitance of the diode and strays, at a frequency of 70 Mc/s, gives a value of loss as approximately 76 dB. In practice the specification of loss is 75 dB, and values generally obtained are in the region of 78 to 79 dB. It will be seen that the value of capacitance must be very small and this affects both the choice of diode and the mechanical layout of the switch. The diode requires to have a very low resistance when conducting, together with a low capacitance when non-conducting; this leads to the choice of a gold-bonded germanium diode. The mechanical layout must be such as to avoid stray coupling, and all leads, especially on decoupling capacitors, must be kept as short as possible. The layout is also influenced by the fact that the switches must be easy to produce in large numbers on a commercial basis.

6.4. *I.F. Amplifier A.G.C. Characteristics*

The switch is designed to work from existing i.f. amplifiers which, being valve amplifiers, have a high impedance a.g.c. line. The level of the carriers reaching two i.f. amplifiers are compared by comparing the a.g.c. voltages produced by the carriers; the higher level carrier will produce the more negative a.g.c. voltage. Typical a.g.c. characteristic curves are shown in Fig. 11. Two points should be noted about these curves.

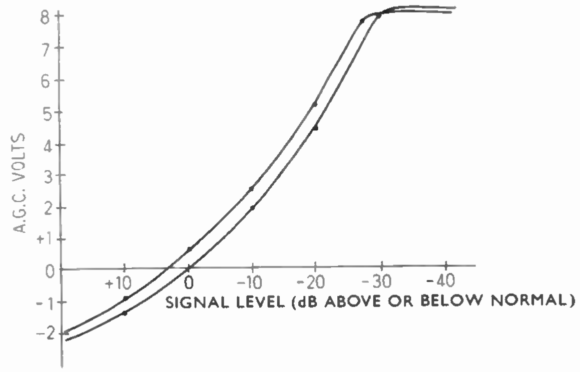


Fig. 11. Typical a.g.c. characteristic curves.

(a) The actual direct voltage obtained for a given input carrier level will vary from amplifier to amplifier. This voltage obviously depends on the gain of each valve (which will vary with age) and the efficiency of the rectifier circuits and amplifiers associated with the a.g.c. Thus, although it is true to say that the shape of the a.g.c. characteristics may be fairly constant from amplifier to amplifier, the curves may be shifted with respect to each other on the voltage axis. This difference in voltage may be as great as 1 V, and it must be allowed for in the initial setting up of the switch panel on the system.

(b) The rate of change of voltage is greater when low level carriers are being received. This is an advantage since it allows greater switching sensitivity when dealing with low-level carriers.

6.5. *Comparison of A.G.C. Voltages*

A valve circuit is used to compare the a.g.c. voltages, since this offers a high impedance to the a.g.c. line. The contacts of a high-speed relay in its anode circuit are used to control the transistor switching circuits (see Figs. 12 and 13).

Each a.g.c. voltage is applied to one grid of a double triode, hence if these voltages differ, the anode voltages will differ, and the high-speed bistable relay will be operated in one direction. The relay remains in this position if the a.g.c. voltages become equal,

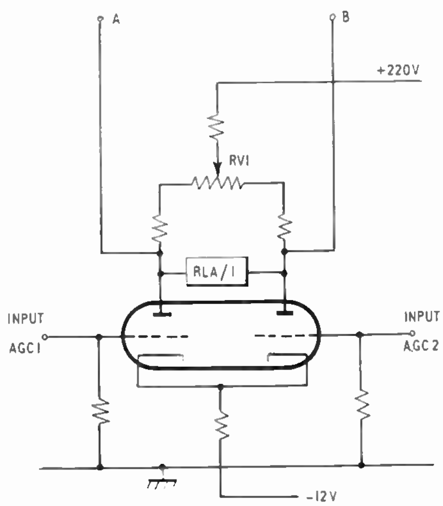


Fig. 12. Schematic of a.g.c. comparison stage.

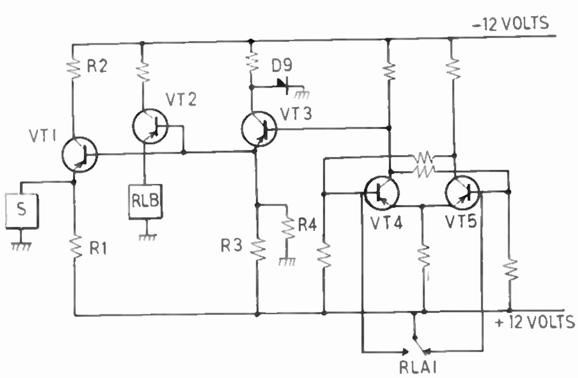


Fig. 13. Schematic of transistor control circuits.

and only changes over when the level of the selected carrier falls below that of the rejected one. The common cathode resistor is taken back to −12 V, to give the maximum possible effect of a long-tailed pair without the introduction of another negative supply.

To allow for the difference between the actual a.g.c. voltages produced by different i.f. amplifiers with the same level of carrier, the potentiometer RV 1 is provided to bias the valve in either direction. When the system is initially set up, an i.f. signal (split by a matched junction) is fed into both of the controlling amplifiers. RV 1 is then adjusted to give zero voltage across the relay, the voltage being measured at external points A and B.

The relay has a transit time of 2 milliseconds. This time delay is quite satisfactory since the switch is normally dealing with signals whose relative levels vary slowly compared with 2 milliseconds, Welber, Evans and Pullis[8] have found that in practice fades have an average rate of change of level of 10 dB per second. Fades at 100 dB per second seldom occur.

The control circuits of the switch are designed to hold one signal until the moving contact of the relay actually makes on the opposite contact. Thus the 2 milliseconds may constitute a delay if the signal abruptly fails due, say, to an equipment fault. These faults are rare.

6.6. *Transistor Control Circuits*

The control circuit consists of a bistable trigger circuit, a buffer transistor and a power transistor which actually controls the switch, together with a means of indication of the input selected. The circuit is given in Fig. 13. The relay contacts, RLA 1, are those of the high speed relay.

The bistable circuits of VT 4 and VT 5 are controlled by the contacts of relay RLA 1. With the contacts in the position shown, the base of VT 5 is connected to +12 V, and hence is more positive than its emitter which is held at about +7 V by VT 4 conducting. Thus VT 5 is not conducting; its collector is at about −7 V and this determines the base potential of VT 4. The transistor VT 4 is conducting with its collector potential about +7 V, this potential being applied to the base of VT 3.

Although VT 5 becomes more negative when the contact of RLA 1 lifts, just before a changeover, it does not become more negative than the emitter and hence VT 5 remains in its non-conducting condition. When RLA 1 makes on the opposite contact, VT 4 is "cut-off", VT 5 conducts, and a change-over occurs. The voltage of VT 3 base is now about − 7 V. It is essential that this changeover should occur as quickly as possible since the ultimate changeover time is to be less than 1 μs. The transistors VT 4 and VT 5 are not allowed to saturate in the conducting condition, to avoid any delay due to hole storage effects.[9] The value of the collector resistors is kept as low as possible to reduce the time constant associated with the collector capacitance. The choice of operating current, which is obviously affected by the value chosen for the collector resistor, is also influenced by the fact that the switch is required to work reliably at temperatures up to 55°C, and the transistor dissipation must be as low as possible.

The transistor VT 1 is a power transistor directly controlling the switch. When the transistor is non-conducting, the switch S forms a potential divider with R1, and about +6 V appears on the control lines of S. Alternatively, if the base of VT 1 is held at about −6·5 V, current is drawn through the switch and R1 by VT 1 conducting, and about −6 V is applied to the switch. Thus the diode switch may be directly controlled by changing the base voltage of VT 1 from −6·5 V to +6·5 V. The currents required for the diodes in the switch give rise to a transistor current of about 400 mA in the conducting condition, and hence produce a large base current in the transistor. This base current is only present for one condition of the switch.

The buffer transistor, VT 3, is used to drive the power transistor—its use is decided by the following factors.

(a) The trigger circuits of VT 4 and VT 5 ultimately control the voltage applied to the switch. The buffer stage eases the design problems of the trigger stages, which would otherwise have to allow for a large base current under one condition, and no base current under the other condition. The current gain of VT 3 gives a considerably reduced effect of base current on VT 4 and VT 5.

(b) The current gain of the power transistor may vary from transistor to transistor, due to production spreads. This may produce large differences in the base current of VT 1 from panel to panel. This variation is taken up by the buffer stage.

The sequence of operation is as follows. When the trigger voltage output is at −7 V, VT 3 conducts and the emitter is at about −6·7 V; VT 1 therefore conducts and applies about −6 V to the switch. When the trigger output voltage is +7 V, VT 3 is not conducting because its emitter is held at +6·5 V by the potential divider R3, R4, hence VT 1 is non-conducting because its emitter is held at +6 V by the potential divider S, R1. The diode D9 is a Zener diode which controls VT 3 collector voltage when VT 3 is non-conducting and so reduces the maximum base to collector voltage of VT 3. The resistor R2 reduces the power dissipation of VT 1 when it is conducting, but VT 1 is not allowed to saturate in order to reduce the hole storage effects on this transistor. The transistor VT 2 merely controls the relay RLB which provides an indication of the input selected.

7. Conclusion

This switch is designed for 960-channel use but its bandwidth would enable it to be used to switch systems with higher channel capacities. Its change-over time of 1 microsecond allows it to be used on circuits carrying high speed digital information.

8. Acknowledgments

The author wishes to express his thanks to the management of the General Electric Co. Ltd., Telephone Works, Coventry, for permission to publish this paper. He would also like to thank Mr. W. Grossett and colleagues in the Radio Transmission Laboratory for helpful advice.

9. References

1. Documents of the IXth Plenary Assembly of the C.C.I.R. Los Angeles 1959.
2. A. B. Crawford and W. C. Jakes, "Selective fading of microwaves", *Bell Syst. Tech. J.*, **31**, pp. 68–90, January 1952.
3. R. L. Kaylor, "A statistical study of selective fading of super-high frequency radio signals", *Bell Syst. Tech. J.*, pp. 1187–1202, September 1953.
4. M. W. Gough, "Microwave line-of-sight propagation", *Electronic Engineering*, **30**, pp. 237–47, May 1958.
   M. W. Gough, "Propagation influences in microwave link operation", *J.Brit.I.R.E.*, **24**, No. 1, pp. 53–72, July 1962.
5. W. von Hagen, A. N. MacDiarmid and L. V. Goldenberg, "Path-loss testing of the Trans-Canada TD-2 route", Canadian Convention of the I.R.E., Toronto, 1957. Paper 49.
6. D. E. Kerr, "Theory of propagation in a horizontally stratified atmosphere", in "Propagation of Short Radio Waves", Chap. 2, p. 27, M.I.T. Radiation Laboratory Series, Vol. 13 (McGraw-Hill, New York, 1951).
7. H. C. Sheffield, "Microwave relay system between Saint John and Halifax", *Electrical Communication*, **32**, No. 4, pp. 215–31, December 1955.
8. I. Welber, H. W. Evans and G. A. Pullis, "Protection of service in the TD-2 radio relay system by automatic channel switching", *Bell Syst. Tech. J.*, **34**, pp. 473–510, May 1955.
9. D. E. Deuitch, "Junction transistor switching characteristics", "Transistors I", pp. 609–39 (R.C.A. Laboratories, 1956).

Manuscript first received by the Institution on 31st August 1961, and in final form on 6th December 1961. Originally submitted as a thesis for exemption from the Graduateship Examination. (Paper No. 750).

GRADUATESHIP EXAMINATION—MAY 1962—PASS LISTS

These lists contain the results of all successful candidates in the May 1962 Graduateship Examination. A total of 595 candidates entered for the Examination, which was held at sixty-nine centres.

LIST 1: The following candidates have now completed the Graduateship Examination and thus qualify for transfer or election to Graduate or a higher grade of membership.

ARAVAMUDHAN, Krishna (S) *Bombay, India.*
ARMSTRONG, John Patrick (S) *Basingstoke, Hants.*
AUDUS, Michael Francis (S) *Abingdon, Berks.*

BARDOS, Peter Andrew (S) *London, W.2.*
BAYLISS, Bryan Howard (S) *Birmingham, 14.*
BINENSTOCK, Joseph (S) *Rehovitch, Israel.*
BOHTAN, Sardar Singh (S) *Poona, India.*
BRIGGS, Terence (S) *Alford, Lincs.*

CARR, Philip Ernest (S) *Broxbourne, Herts.*
CHADDA, Santosh Kumar (S) *Delhi, India.*
CHAPMAN, Christopher John (S) *London, S.E.22.*
CHUNG, Yook Kei (S) *Hong Kong.*
COULTER, John Patrick (S) *Stanmore, Middlesex.*
COXAN, David John (S) *Malvern Link, Worcs.*

DAS, Mukunda Madhale (S) *Calcutta, India.*
DAS, Tapan Kumar (S) *Bangalore, India.*
DICK, Ian Jowett (S) *Cardross, Dumbartonshire.*
DIXEY, Graham Edward (S) *Ickenham, Middlesex.*
DUARTE-CATULO, Fernando Jose Eugenio (S) *Goa, India.*

EDMONTON, Brian, *Norwich, Norfolk.*

FORESTAL, Peter (S) *Hong Kong.*

GIBSON, George Arthur (S) *Aberdeen.*
GOLDMAN, Rudolf (S) *San Paulo, Brazil.*
GREENLAND, Elwyn (S) *Luton, Bedfordshire.*
GUPTA, Jitender Kumar (S) *Delhi, India.*

HAKHVERDIAN, Armik (S) *Teheran, Iran.*
HALE, Derek Stanley (S) *Jersey, Channel Islands.*
HAMMOND, Edward (S) *Stoke Poges, Bucks.*
HARRIS, Herbert Arthur (S) *Johannesburg, South Africa.*
HASAN, Ghulam (S) *Karachi, Pakistan.*
HILL, Eric Sylvester (S) *Wells, Somerset.*
HILL, Michael Edward (S) *Birmingham, 15.*

INKPIN, Jack Dennis (S) *London.*

KENWARD, Michael (S) *Wallington, Surrey.*
KUMAR, Jagdish (S) *Delhi, India.*

MANCHESTER, John Kay (Associate) *London, W.4.*
MARTIN, Alexander Duncan Brown (Associate) *Lisburn, N. Ireland.*
MUKHERJEE, Sudhendu Kumer (S) *Kanpur, India.*

NAGESH, Jai Ram (S) *Punjab, India.*
NG, Ser Choon (S) *Singapore.*

PADMANABHAN, Krishnaswami (S) *Madras, India.*
PARK, Kenneth Charles (S) *Shepton Mallet, Somerset.*
PARKIN, Oswald Theodore (S) *Havant, Hants.*
PEIL, Frederick (S) *Marlow, Bucks.*

RUTGERS, Willem (S) *Netherlands.*

SALTER, Martin Thomas Ardley (S) *Croydon, Surrey.*
SHUKLA, Avanish Chandra (S) *Lucknow, India.*
SIOW, Keng Cheng (S) *Singapore.*
SIVARAMAKRISHNAN, Perudurai Nagaswamy (S) *Bombay, India.*

TAYLOR, John Chisholm, *South Shields, Co. Durham.*
THAKRAL, H. B. Rai (S) *Kew, Surrey.*
THOMAS, Philip Wilfred (S) *Wells, Somerset.*
THOMAS, Victor Francis, *Wheathampstead, Herts.*
TOEG, Haim (S) *Tel Aviv, Israel.*

WHEATON, John Edward George (S) *London, S.W.15.*
WILSON, James Albert (S) *Hazel Grove, Cheshire.*
WOODS, John Frederick (S) *Bishops Stortford, Herts.*

LIST 2: The following candidates have now satisfied the requirements of Section A.

ADEFEMI, Stephen Ayodele, *London, S.E.27.*
ADEYINKA, Jacob Ore (S) *Lagos, Nigeria.*
AGASHE, Srikrishna Narayan (S) *Aurangabad, India.*
ALLEN, Norman Leonard (Associate) *Hockley, Essex.*
ASGILL, Julius James (S) *London, N.7.*
AUQIL, Hafiz Mohd (S) *London, N.W.3.*

BALASUBRAMANYAN, Pallassena Natesa (S) *Poona, India.*
BARRATT, Kenneth Henry (S) *Reading, Berkshire.*
BASSEY, David Andrew (S) *Kaduna, Nigeria.*
BAYLEY, Arthur Ward (S) *Staines, Middlesex.*
BHATTACHARYYA, Amar (S) *London, N.W.11.*
BOWSFIELD, Ralph Clement (S) *Watford, Herts.*

CERESOLE, Peter Lester, *London, S.W.5.*
CHACKAL, Rizk Fathi Elias (S) *London, N.13.*
CHAKRABORTY, Narayan Chandra (S) *Kanpur, India.*
COPPING, John Richard (S) *Portsmouth, Hants.*
CORBETT, Dennis John (S) *Evesham, Worcs.*
CURTIS, Alan Paul (S) *London, W.13.*
CURTIS, Norman Ralph (Associate) *Quebec, Canada.*

EGWU, Godwin (S) *London, N.5.*

FULTON, John, *York.*

GASKILL, William (S) *Liverpool, 9.*
GRIFFIN, Michael Robert (S) *Didcot, Berkshire.*

HARVEY, Robert (S) *Ikeja, Nigeria.*
HEGHOYAN, Manas (S) *London, N.5.*
HILLMAN, Peter, (S) *Rainham, Essex.*
HORTON, Laurence, *London, S.W.2.*
HOWARD, Patrick John (S) *B.F.P.O. 40.*

IJIWOLA, Gabriel Oworuntan (S) *Ikeja, Nigeria.*

JACOBS, Jonathan, *London, N.W.6.*
JAMES, Valiaparambil George (S) *Ernakulam, India.*

KAINE, Donald (S) *Reading, Berkshire.*
KRISHNAMURTHY, Srinivas (S) *Jamnagar, India.*

LAWRENCE, John Leonard Albert (S) *London, N.13.*
LYSAK, Eugene, *London, W.4.*

MASON, Leonard Charles (S) *Kampala, Uganda.*
MATHEW, Puliyelil George (S) *Hyderabad, India.*
MATTA, Sushil Kumar (S) *New Delhi, India.*
MENG, Joo Tng (S) *Singapore.*
MILES, David Frank (S) *Smethwick, Staffs.*
MITCHELL, Thomas William (S) *Falkirk, Stirlingshire.*
MITHANI, Kishore Shantilal (S) *London, S.W.3.*
MITTAL, Satish Chandra (S) *Bombay, India.*
MOORE, Roger Anthony (S) *Onitsha, Nigeria.*
MORRISON, Andrew Stephen, *Danbury Common, Essex.*

NAYEEM, Mohammed Abdul (S) *Hyderabad, India.*
NISSIM, Moshe (S) *Ramat Gan, Israel.*

OKAFOR, Godson Nnonye (S) *Lagos, Nigeria.*

PADMANABHAN, Vadakkath (S) *Cochin, India.*
PADMANATHAN, Muruguppilai, (S) *Ceylon.*
PATEL, Ebrahim Mohamed Moosa (S) *London, S.E.9.*
PITAWALA, Abeywardena (S) *Sutton, Surrey.*
PRESCOT, Edward Horace Albert (S) *Oulton Broad, Suffolk.*

RAMAMURTHY, Aiyasamy (S) *Bangalore, India.*
RAGHAVEN, Cheruparambath (S) *Ernakulam, India.*
RANDALL, James William (S) *Portsmouth, Hampshire.*
RENDALL, Gavin Mowat (S) *Lagos, Nigeria.*
REVITCH, Alexander (S) *Tel Aviv, Israel.*
ROSSI, Agostino, *London, N.5.*

SAGE, Michael John (S) *Harrow, Middlesex.*
SARIKONDA, Jai Ram Raj (S) *Bangalore, India.*
SARKISSIAN, Arsham (S) *Teheran, Iran.*
SCOTT, Peter (S) *Barnet, Hertfordshire.*
SHOTTER, Robert Raymond Joseph (S) *Paisley, Scotland.*
SHENOI, C. S. Sundaresa (S) *Cochin, India.*
SRINIVASAN, S. (S) *Madras, India.*
SRIVASTAVA, Havis Chandra (S) *Lucknow, India.*
STOTT, Alan Ernest (S) *Manchester 8.*
SUDHAKARAN, Melangath (S) *Cochin, India.*

TAYLOR, Robert William (S) *London, W.14.*
THAKE, Philip Edward, *Leicester.*
THIRUMALAI, Norton Seshikanth (S) *London, N.16.*

VAN RIXTEL, Rawdon Paul (S) *London, N.10.*
VERMA, Vidya Dhar (S) *Madras, India.*
VOOR, Taivo (S) *London, W.2.*

WADDINGTON, John Bryan (S) *New Malden, Surrey.*
WAKERLEY, Philip Albert (Associate) *Salisbury, S. Rhodesia.*
WATERLANDER, Cornelius Johannes (S) *Sydney, Australia.*
WHELAN, John (S) *Stevenage, Herts.*
WONG, Alan Yui Keong (S) *London, N.5.*

YATES, Robert Stuart (S) *Stafford.*

(S) denotes a Registered Student.