FOUNDED 1925
INCORPORATED BY
ROYAL CHARTER 1961

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

THE RADIO AND ELECTRONIC ENGINEER

The Journal of the Institution of Electronic and Radio Engineers

VOLUME 34 No. 2 AUGUST 1967

Computers and the Engineer

DURING the recent Institution Conference on 'The Integration of Design and Production in the Electronics Industry', several of the papers discussed how that most versatile of the electronics industry's many inventions—the electronic computer—can be employed in a major role in design, production and management. Although the applications described at that Conference were concerned with the electronics industry itself, the principles involved can usually be transferred to other industries.

Computer-aided design of electronic circuits is, to the electronic engineer, probably one of the most fascinating uses of the computer. The intellectual processes which a circuit designer uses are not all easy to specify in a manner which can be handled by a computer. Some of the basic creative steps must still be taken by the designer working by intuition or past experience (though perhaps the latter might be stored in the computer). It is where the analysis of alternatives of network configuration is required that the computer comes into its own, working more speedily and more accurately. There are other exciting possibilities, described in a Conference paper by N. E. Wiseman,† of bringing the designer into the 'design loop' via light pens and cathode-ray tube displays. The computer design of optimum layout for extensive arrays of printed wiring or integrated circuits is also the subject of much activity in both industrial and non-industrial research laboratories.

Production and management techniques developed over recent years, such as critical path methods and p.e.r.t., can readily be programmed for solution using digital computers and enable resources and facilities to be planned in the most effective way. On-line production control is a field of activity which, as was obvious from discussions at the Nottingham Conference, will greatly repay exploitation for enterprises concerned with batch or mass production. The business simulator for investigating the feasibility of future operations is another application still in its infancy.

Underlying all these existing and potential applications is the need for far more work to be done in two important areas: direct user communication with a possibly remote computer and multi-access techniques to make the fullest use of large, centrally sited machines. These desirable developments are being encouraged by the British Ministry of Technology, which has already placed several contracts, mostly directly concerned with computer-aided design. In addition a Computer-Aided Design Committee with members drawn from Government establishments, universities, computer manufacturers, industrial users of computer-aided design and other interested bodies, has recently been set up by the Minister of Technology to ensure that the fullest use is made of the computer as a design tool.

Management and design engineers may need reassuring that although the computer does the work, it is they who make the decisions. To quote from another Conference paper, 'The system is there as a tool to assist them in their work rather than a device to rob them of a job. The feature most likely to convince the potential user that the system is "friendly" is to design it in such a way that two-way communication is easy and that there is no unnecessary complication in the man/machine interface.'‡

Problems of communication are found throughout technology and the successful development of computer-aided design techniques presents a considerable challenge to the engineer.

F. W. S.

† 'Some applications of computers in electronics design', *I.E.R.E. Conference Proceedings No. 9*, Paper No. 13.
‡ A. B. Triggs, 'The rise of computers in production and management', *I.E.R.E. Conference Proceedings No. 9*, Paper No. 24.

INSTITUTION NOTICES

Institution Premiums and Awards

The Council of the Institution announces that the following awards are to be made for outstanding papers published in *The Radio and Electronic Engineer* during 1966:

CLERK MAXWELL PREMIUM

'True I.F. Logarithmic Amplifier using Twin-Gain Stages' by A. Woroncow and J. Croney (September).

HEINRICH HERTZ PREMIUM

'Wideband Coaxial Variable Attenuators using p-i-n Diodes' by J. R. James and Professor M. H. N. Potok (March).

J. LANGHAM THOMPSON PREMIUM

'Analysis and Synthesis of Feedback Compensated Third-order Control Systems via the Coefficient Plane' by D. R. Towill (August).

P. PERRING THOMS PREMIUM

'A Subjective Investigation of some Errors in the Chrominance Signal Decoding Circuits of Colour Television Receivers' by K. E. Johnson (June).

DR. NORMAN PARTRIDGE MEMORIAL PREMIUM

'The Floating Transcription Arm: A New Approach to Accurate Tracking with very Low Side-thrust' by A. R. Rangabe (October).

ARTHUR GAY PREMIUM

'Prediction and Engineering Assessment in Early Design' by W. P. Cole (January).

A. F. BULGIN PREMIUM

'Gallium Arsenide Varactor Diodes' by C. A. P. Foxell and K. Wilson (April).

MARCONI AWARD

'Gain and Stability of Tuned Transistor R.F. and I.F. Amplifiers' by M. V. Callendar (October).

LORD RUTHERFORD AWARD

'Acoustic Amplification in Semiconductors' by R. W. Harcourt, J. Froom and C. P. Sandbank (March).

LORD BRABAZON AWARD

'Improved Radar Visibility of Small Targets in Sea Clutter' by J. Croney (September).

The award of the Bose Premium and the Mountbatten Premium is under consideration by the Council of the Indian Division who will make recommendations to the Institution's Council later this year.

The following Premiums and Awards are withheld as papers of suitable standard have not been published during the year: the Charles Babbage Award, the Leslie McMichael Premium, the Hugh Brennan Premium, the Zworykin Award and the Rediffusion Television Premium.

Canadian Division

Members in Canada and the United States are asked to note that the address of the Institution's Canadian Division Office is now:

I.E.R.E. Canadian Division,
Room 300, Burnside Building,
151 Slater Street, Ottawa 4, Ontario, Canada.

Telephone (no charge): 234–5513.

Conference on Solid State Physics

The 5th I.P.P.S. Annual Conference on Solid State Physics will be held at the University of Manchester, Institute of Science and Technology, from 3rd to 6th January 1968. Contributions are invited on any topic of current interest in this field, and offers of papers should be sent before 27th October to Professor N. H. March, Department of Physics, The University, Sheffield 10.

Further information and application forms may be obtained from the Meetings Officer, The Institute of Physics and The Physical Society, 47 Belgrave Square, London, S.W.1.

Corrections

The following correction should be made to the paper 'The Design of a Magnetic Thin-film Store for Commercial Production' which was published in the March 1967 issue of *The Radio and Electronic Engineer:*

Page 197, equation (6) should read:

$$L = \frac{\varepsilon}{v_0^2 C} \text{ H/cm.}$$

In the notice about the forthcoming Special General Meeting which was published in *The Radio and Electronic Engineer* for July 1967 (p. 4), a line of type was omitted from the sentence describing the proposed changes to the grade of Associate. The last sentence of the second paragraph of the notice should read:

'In addition the requirements for entry to the class of Associate will be amended to permit the election of senior technicians possessing educational qualifications of a standard not less than that of Higher National Certificate or the City and Guilds Full Technological Certificate, or such other similar qualifications as the Council may prescribe.'

U.D.C. 681.39

Automatic Recognition of Low-quality Printed Characters using Analogue Techniques

By

J. R. PARKS, Ph.D.
(*Graduate*)†

Summary: After comment on the obvious disparity between the print quality accepted as normal for human reading and the much higher quality needed for machine reading, a system is proposed and results of a simulation study presented for reducing this gap. The system employs a modified form of auto-correlation to extract characteristic properties from which specimen unknown characters can be identified by a simple recognition logic. In order to avoid the problem of representing the continuous grey-tone density scale of low quality print as a two state (binary) variable, analogue techniques are proposed for use in the system.

Using a computer simulation of the proposed system a limited number of printed samples of exaggerated quality distribution are used to explore the limits of performance of the system. Numerical results are given and factors effecting performance discussed.

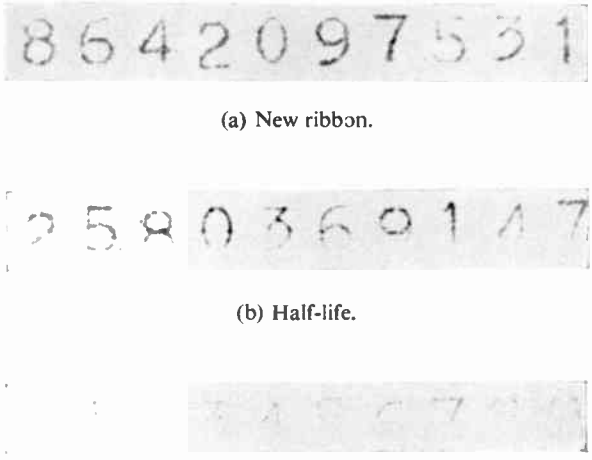
List of Symbols

$A(\mathbf{a}, \mathbf{b}, \mathbf{c}\ldots)$, etc.	function describing the analogue operator AO	n	number of sampling points in an AO
$A_i(c)$	value of $A(\mathbf{a}, \mathbf{b}, \mathbf{c}\ldots)$ when applied to character class c	$p(r)$	probability of exactly r picture elements being black
a, b, c, etc.	translation vectors applied to the original character $d(\mathbf{r})$	R	region containing $d(\mathbf{r})$
$d(\mathbf{r})$	density function representing the input character	r_{ab}	correlation between vectors a and b
k	number of specimens of selected character class	$r_{ij}(c)$	correlation between ith and jth AO measures for character class c
m	dimensionality of the decision space – number of AO measures employed in a recognition logic	S_b	score for character class b
		$\mathrm{var}_i(c)$	variability of measure i for character class c
		ρ	optical density = (1 − reflectance)
		$\mu_i(c)$	mean value of measure i for character class c

1. Introduction

Commercially-available character recognition equipments are limited in their application by their inability to operate with the conventionally accepted range of print quality. Other restriction, e.g. intolerance of skew, inability to read more than one font, etc., are also significant but they are less fundamental than the quality limitation. While some of these other problems are attracting considerable attention, there is little indication that significant advances have been made in the development of techniques expressly directed at the problem of low print quality.

Examples of low-quality printing are shown in Fig. 1. This quality of printing is typical of a variety of low-priced impact printing mechanisms (e.g. cash registers, typewriters, ticket-issuing machines, high-speed computer output printers, etc.). It is, frequently,



(a) New ribbon.

(b) Half-life.

(c) Expired ribbon.

Fig. 1. Examples of low-quality printing.

† National Physical Laboratory, Autonomics Division, Teddington, Middlesex.

not economically possible either to modify or maintain these types of equipments to ensure that the quality of printing which they produce is capable of being read, with an acceptable error and reject rate, by available character recognition (c.r.) equipments.

Two typical print quality specifications are summarized in Table 1. These specifications relate to c.r.

Table 1

Summary of typical print quality specifications. Voids and background marks are defined as closed areas of greater than a stated superficial area whose density falls below or above the permitted character limb density

	A	B
Maximum permissible overall range of limb thickness	2 : 1	1·8 : 1
Maximum range of ink density (on white paper reflectance = 90%)	100–50%	100–55%
Maximum gross variation of density within single character	100–50%	100–55%
Minimum (bulk) background reflectance	70%	60%
Maximum permissible size of voids in character limbs	0·007 in diameter	half an area 0·015 in square
Maximum size of background marks	0·007 in diameter	half an area 0·015 in square
Skew	±2°	±2°
Maximum variation in background reflectance	20%	10%

equipments employing single non-stylized fonts in conventional size. The results of a print quality survey by Gerlach[1] show that, during the normally accepted life of a printing ribbon of some $1{\cdot}25 \times 10^5$ impressions, a range of light absorption within character limbs from 100% to 20% is obtained. In addition offset of ink in the character background may have a density up to 50% for a new ribbon. Gerlach's results are summarized graphically in Fig. 2(a) and are interpreted as a contrast ratio,

$$C = \frac{\text{character limb density}}{\text{background density}}$$

in Fig. 2(b). The print specifications quoted in Table 1 apparently demand a minimum contrast ratio of 4 : 1 which, according to Gerlach's results, cannot be guaranteed at either end of a ribbon's life. Although a ribbon can be discarded prematurely, little can be done to improve the condition at the beginning of a conventional ribbon's life.

It is also apparent from Gerlach's results that character limb width can vary over a range of the

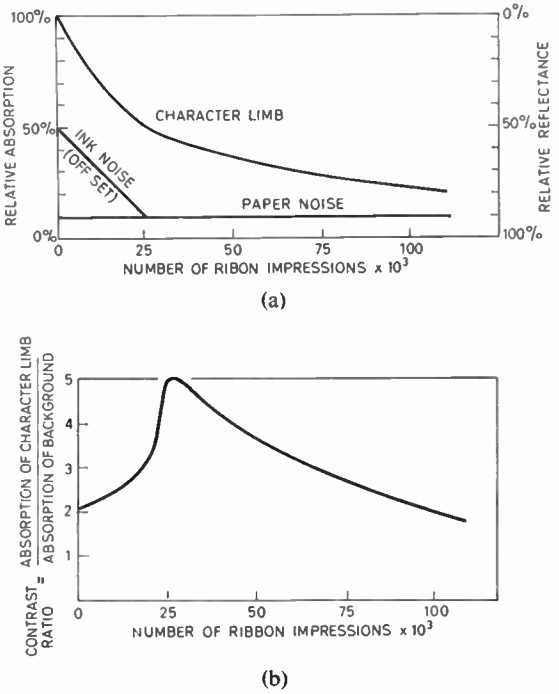


Fig. 2. Ink density as a function of ribbon life (after Gerlach[1]).

order 3 to 1. Again this exceeds the range recommended for use with available systems. Severe non-uniformity of line density is also reported, examples of which are shown in Fig. 1.

The results quoted are related specifically to cash register printing mechanisms and help to define the desirable performance of c.r. equipments operating on material from such sources. Since similar phenomena are likely to influence the quality obtained from most wet-ribbon printing mechanisms, Gerlach's findings will be assumed to apply in the wider context. In the case of typewritten material it should be noted that, due to changing key pressures, ink density may additionally vary from character to character.

It is obvious from this discussion that the print quality tolerances recommended by various equipment manufacturers, if a stated performance with their devices is to be obtained, do not correspond to the quality variation obtained under currently accepted normal working conditions. This does not mean that existing equipments cannot be usefully applied but demonstrates one probable reason why available character recognition equipments have not yet found wide application except in situations in which the print quality can be closely controlled.

To be commercially acceptable a c.r. equipment must have reject and error rates better than 1 in 1000 and 1 in 10 000 characters respectively, (these rates are probably the minimum acceptable limits, actual rates required will vary with the application. A more

realistic performance for a general-purpose machine would be 1 in 10^5 errors and 1 in 10^4 rejects.) Since reject and error rates obtainable in practice are clearly a function of print quality the performance of any system is limited by its ability to process marginal quality material.

It is a fundamental precept of available systems and the majority of theoretical studies published that printed characters can be represented adequately on a binary, two-state, density scale. This is only possible when the quality of printing is such that the reflectance of the character and its background are reasonably constant and clearly discriminable. In typical low-quality conditions these criteria are unlikely to be met and any scheme designed to dichotomize the reflectance scale of presented characters is required to discriminate the character from its background without any prior knowledge of the position or condition of either. Any dichotomizing system, however advanced, must therefore employ some *ad hoc* decision mechanism to indicate whether a particular picture element of a presented pattern belongs to the character or to its background. This operation inevitably further degrades the character. The effect of applying a dichotomizing threshold to a low-quality character is shown in Fig. 3. The maximum and minimum density of the character and its background have been determined and the interval thus defined sub-divided on an eight increment linear scale, a threshold has been inserted half way along this interval, i.e. between divisions 3 and 4.

If a technique capable of processing patterns in their original, continuous density, form can be devised then some improvement in performance under marginal print quality conditions should result. It is the purpose of this paper to describe a possible hardware system and to demonstrate, with the aid of a computer simulation of the system, that an improvement

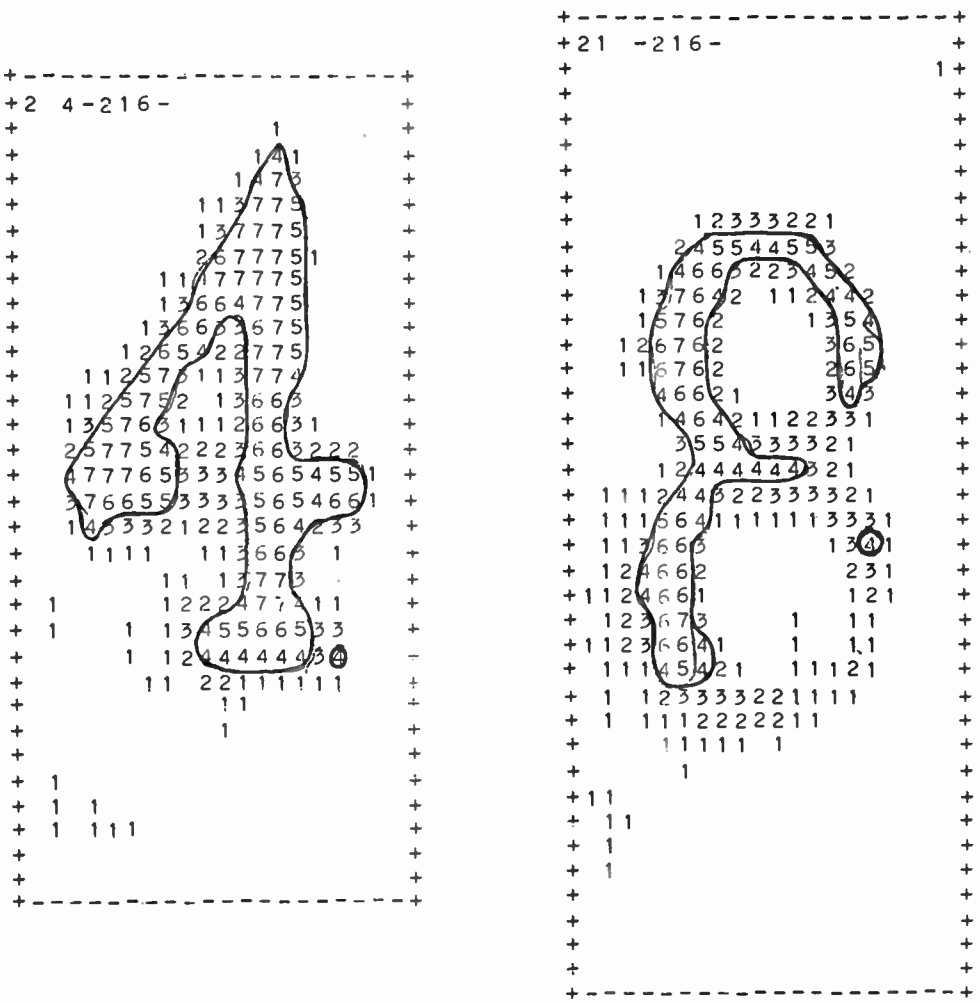


Fig. 3. Dichotomizing threshold applied to poor-quality printed characters.

in performance is obtained by using analogue rather than the more conventional two-level digital processing techniques.

2. Current Approaches to Pattern Recognition Problems

The general problem of pattern recognition can conveniently be split into two parts: (a) the analysis of presented patterns to determine their significant properties, and (b) the identification of the presented patterns from the results of such an analysis. This division is convenient from the point of view of discussion but the two parts are clearly not independent.

The second of these problems, classification, has received considerable theoretical examination assuming as a rule that the problems of devising useful pattern analysis techniques have been resolved. At the present state of the art no formal methods exist for determining what are the significant properties of any class of patterns and those devices which have been proposed use intuitively justified methods to define such properties.

A variety of different techniques have been explored including the use of standard mathematical techniques such as Fourier analysis,[2] central moments of mass[3] and auto-correlation techniques.[4,5] These techniques have not in their 'pure' form shown any great promise, indeed experiments using Fourier and moments methods of analysis have revealed specific fundamental limitations; the former in its inability to discriminate between character pairs which are mirror images about certain axes, the latter being unduly sensitive to disturbing influences towards the periphery of the area examined and relatively insensitive to genuine geometric differences in character geometry near its centre of gravity. Auto-correlation techniques on the other hand have been described in a variety of modified forms with more promise.

Clowes and Parks[6] and Clowes[7] described a modified form of auto-correlation procedures which can detect and discriminate a limited number of geometric properties of a pattern and indicate their orientation. These techniques are fundamentally position invariant so that the need to accurately locate presented characters is eliminated. The inherent noise reduction properties of the auto-correlation transform is retained in its modified forms which should therefore be applicable in a low print quality situation and might additionally be insensitive to changes of character style.

One auto-correlation operation which has been used is illustrated in Fig. 4 in which three (or more generally n) replicas of an input character c are superimposed after rigid translations **a**, **b**, have been applied. A characteristic function for any character c

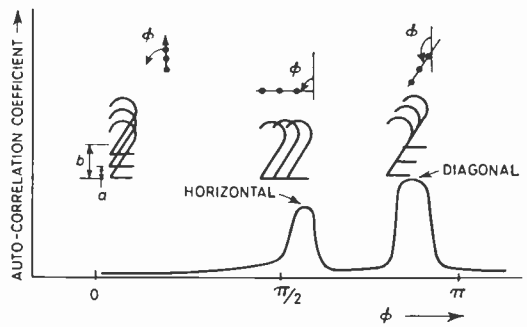


Fig. 4. Relationship between maxima in the auto-correlation function and characteristic straight line features in Fig. 2.

is obtained by measuring the total coincidence of all three replicas as translations **a**, **b**, are rotated synchronously through an angle ϕ. The resulting function has maxima and minima indicating the presence or absence in the character of properties defined by the translations **a**, **b**. In identifying a source character it is only the positions (orientations) of such minima and maxima which are used in discriminating between character classes. It will be noticed that if the character style is known *a priori* then the position of maxima and minima in the function can be predicted for each character class and only those sections of the function need be evaluated.

Another method of analysing characters has been to employ so-called n-tuples. Patterns are characterized by the state, i.e. 1 or 0 of sets of n sampling points distributed about the matrix on which the character is represented.[8,9] The identity of the analysed character can be determined upon examination of the state of the n-tuples.

The earlier n-tuple schemes required that the input character, in binary form, be accurately positioned on a quantizing matrix, this process is clearly not positional invariant. Modified schemes in which the n-tuples are scanned exhaustively over the input character retain this invariance however. Such a technique has been described capable of classifying mixed font numerals and alphabetics[10] and also of recognizing hand-printed characters and simple line drawings.[11]

3. System Principle

Auto-correlation pattern processing techniques are fundamentally analogue although the author is not aware of their use on continuous density patterns. The result of an auto-correlation operation is a second continuous function which, although invariant of the position of the original character, is still not particularly convenient for use in a decision network.

Scanning n-tuples can be likened to the determination of instantaneous values of an auto-correlation

function (Fig. 5) since, for a given style, the position of maxima and minima in the function are predictable and the intermediate parts of the function are largely redundant.

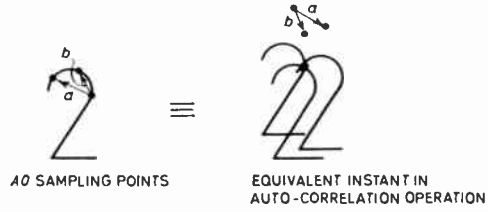


Fig. 5. AO sampling points detecting a curve. In practice the points are scanned exhaustively over the area containing a character and hence will detect the property defined regardless of its position in the area. The equivalence of the AO to a particular point in the auto-correlation function is also shown. (The reversal of the shift vectors is inherent in the two methods of illustrating the same operation.)

The scanning n-tuple searches for a set of specific Boolean functions anywhere within the matrix supporting the input pattern (as an array of '1's and '0's). In practice the mere occurrence of an n-tuple has been used as a property measure.[10,11] However, the aggregate activity of each particular n-tuple is a more reliable measure less sensitive to fortuitous n-tuple activity arising from 'noise' in the source pattern.

If, in addition to summing the activity of n-tuples, the Boolean logical operations are replaced by arithmetic operations, e.g. multiplication, then a continuous density function may replace the binary function representing the input characters and the operation performed by this analogue operator, AO, is clearly analogous to computing an instantaneous value of the auto-correlation function.

The advantages gained by using AO's are significant. These are: (*a*) the complete auto-correlation function is not generated but only the parts specified by individual AO's and (*b*) the measures obtained are analogue in nature and indicate the degree to which AO's detect the properties of the characters analysed, rather than the less informative indication that an n-tuple has occurred at least once.

The proposed process can be expressed formally as follows, if the input character c is defined as a density function $d(\mathbf{r})$ over a region R which supports character c then the function $A(\mathbf{a}, \mathbf{b}, \ldots)$ performed on $d(\mathbf{r})$ is

$$A(\mathbf{a},\mathbf{b},\mathbf{c},\ldots) = \sum_{\mathrm{R}} d(\mathbf{r}).d(\mathbf{r}+\mathbf{a}).d(\mathbf{r}+\mathbf{b}).d(\mathbf{r}+\mathbf{c})\ldots \mathrm{d}S \qquad \ldots\ldots(1)$$

where the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$ define the relative positions of the points of the AO (or n-tuple, if $d(\mathbf{r})$ is a binary function) as in Fig. 5.

So far AO's have been described as reacting to the presence of character structure jointly at each of its n sampling points. Clearly in order directly to detect, say, the open loop at the top of a figure three and distinguish it from the closed loop at the top of an eight the gap in the loop in the three must be detected explicitly. This can be achieved by using sampling points which respond to the character background, i.e. absence of character structure, instead of the character structure itself.

In order that sampling points reacting to the character background may be included, eqn. (1) must be modified to include terms of inverted sense. These terms will be defined as $d'(\mathbf{r})$, where conventionally $d'(\mathbf{r}) = 1 - d(\mathbf{r})$. However, it will be shown later that for engineering reasons an alternative form $d'(\mathbf{r}) = 1/d(\mathbf{r})$ is preferred. The general form of eqn. (1) then becomes

$$A(\mathbf{a},\mathbf{b},\mathbf{c},\ldots) = \sum_{\mathrm{R}} d(\mathbf{r}).d(\mathbf{r}+\mathbf{a})\ldots d'(\mathbf{r}+\mathbf{f})\ldots \mathrm{d}S \qquad \ldots\ldots(2)$$

The distribution of the sampling points of AO's are restricted so that the maximum area spanned by any AO is of the order of half the character space, i.e. that area which will just enclose any of the characters to be recognized. In this way AO's detect local properties of the character and are individually likely to be more reliable, in the presence of mutilation or graduation in ink density across the character, than AO's which extend over the full character space.

The constraint is in accord with similar constraints imposed on the scanning n-tuples (logic circuits) used by Kamentsky and Liu.[10] The most useful n-tuples reported by Bledsoe and Bisson[12] in their experiments are compact although not explicitly constrained in any way.

The remainder of this paper is devoted to the study of an engineerable system for computing functions of the form of eqn. (2) using a computer simulation. Detailed description of the design of individual system components will be published at a later date and only general engineering pronciples will be discussed here.

4. A Practical System for Evaluating AO's

In studying any system for automatic character recognition by computer simulation it is desirable that it should be based upon principles which are practicable with available engineering techniques. Any simulation which does not meet this condition is largely of academic interest only.

Techniques for manipulating two-dimensional density patterns are limited in the main to the manipulation of video waveforms derived from a suitable character scanner. Ideally the patterns would be manipulated as true two-dimensional patterns, but this is only possible by employing optical or electron-optic devices. The former require elaborate mechani-

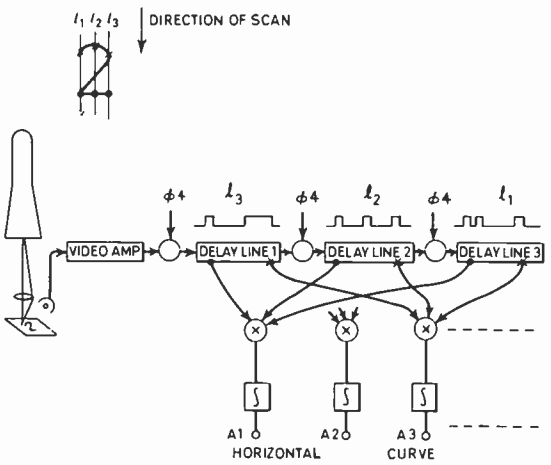


Fig. 6. A system for computing AO's based on a triple line scan showing waveforms in the delay lines at the instant of completing the third scan line. Units for detecting two properties, 'horizontal' and 'curved', are illustrated. The gate signal ϕ_4 is used to prevent spill-over between lines at the completion of a triple line scan.

cal motions of optical components and the use of photographic transparencies resulting in a limited speed of operation, the latter employing electron-optic techniques which are still experimental or even speculative.

A technique for effectively producing displaced character images based upon conventional raster scanning methods was therefore devised. The arrangement to be described here is shown in Fig. 6. In this system the analogue output waveform from a flying-spot scanner is fed to three serially connected delay lines, DL1, DL2 and DL3, individual lines providing a delay equal to the line scan period. At the completion of three-line scan l_1, l_2 and l_3 the corresponding video waveforms are distributed along delay lines 3, 2 and 1 respectively. During the next, the fourth-line scan period, the video signal will propagate through the length of one delay line and any arbitrary point on a delay line will have effectively scanned the character in a manner similar to but displaced from the actual scanning spot. The nature of the displacement is determined by the delay. If this delay is less than a line-period then the displacement is by the appropriate fraction of a scan line-length in the direction of the scan. Similarly, if the delay exceeds a line-scan period the displacement contains a component normal to the scan direction in addition to a component in the line-scan direction.

In the system being described the line-scan direction is vertical and an integral number of line-scan delays result in horizontal displacement. The mode of scanning proceeds in groups of three line-scans, line-triples, separated by a fourth line-scan period; the individual lines in the line-triple are widely spaced across the character scanned (see Fig. 6(a)) and suitably positioned tapping points along the three delay lines define the sampling points of an AO. Clearly, within the constraints imposed by the line-triple scan format groups of tapping points may be selected to define a large number of AO's.

The numerical value of each AO is computed by multiplying together the outputs from the selected tapping points and integrating the product during the course of twenty line-triple scans each displaced a small increment horizontally. Thus during the course of a complete scan containing the stated number of line-triples the sampling points of all AO's are effectively scanned exhaustively over the character space.

The restriction on sampling points to lie on three vertical character sections arises from both technological and economic considerations. In order to include in the space scanned the largest area permitted by a typical (cash register) print format while retaining sufficient resolution to reproduce the thinnest character limbs, the scan line must be resolvable into at least forty picture elements. The technological problem of providing a tappable analogue delay system with a resolution much better than 1% are formidable and a three-line delay was therefore adopted requiring the storage of some 120 picture elements. From an economic point of view three points are the minimum number which can discriminate curved and straight character limbs, three vertical sections are therefore the minimum number for the discrimination of horizontally oriented features.

Again for technological reasons, the method of multiplication chosen was based on familiar logarithmic (log, add and antilog) techniques. Such techniques are the only economical means for multiplying together an unspecified number of variables of wide bandwidth. In a logarithmic system the use of a reciprocal density scale, $1/d(\mathrm{r})$, to represent inverted sense patterns is preferred to the more obvious complementary form, $1-d(\mathrm{r})$, since the reciprocal scale can be realized by simply subtracting the signal, already logarithmic in form, at selected sampling points before anti-logging. To provide true complementary density functions would require two complete delay line systems.

In order to assess the potential performance of such a system a number of experiments using a computer simulation were carried out with the objective of showing that (*a*) recognition performance is improved by the use of continuous analogue rather than binary density function, (*b*) the modified inverse sense function, $1/d(\mathrm{r})$, is acceptable, and (*c*) that a useful performance is obtained under low quality print conditions.

Fig. 7. Specimen characters—design sets.

[N.B. Due to the number of reproductions some additional loss of print quality has unavoidably occurred in both these illustrations.]

Fig. 8. Specimen characters—test sets.

5. Simulation Study

The DEUCE computer has been programmed to perform transforms of the type shown in eqn. (2). Specimen characters for use with the simulation were transcribed on to paper tape and represented on a 40×20 quantizing matrix. The transcription was performed automatically by sampling each matrix element in the output from a flying-spot scanner and converting the amplitude of the sampled element to digital form using an analogue to digital converter. This equipment recorded the sample amplitude as a seven bit binary number providing a resolution of about 1%. For economy in the use of storage capacity and computing time the simulation program operated on only the three most significant digits resulting in an eight level quantizing scale. A reconstruction of a typical character quantized in this way is shown in Fig. 3. A binarizing threshold at half density level is included to illustrate the additional mutilation which could result from this process. A total of 31 sets of characters 0–9 produced in a variety of ways from a pair of well-worn typewriters were transcribed in this way. This material was divided into two sets (*a*) 'the design set' Fig. 7 and (*b*) 'the test set' Fig. 8. The 'design set' was used in the design of AO's and the determination of the recognition criteria. These were then applied to the 'test set' in order to determine the performance of the selected AO's and supporting recognition logic. It will be noticed that the 'test set' of characters contains an exaggerated range of print quality and a variety of printing conditions. This method of test was preferred to the more conventional use of a large number of typical specimens in order to determine the limits of print quality within which the system will operate. The alternative method of trying to estimate an overall error rate on 'typical material' is not possible when using a computer simulation owing to the computing time required to process a significant number of samples, e.g. 10^5 or more.

6. Consideration in the Design of AO's

AO's were selected manually by superimposing and comparing members from the numeral set in the form of photographic transparencies. Initially AO's were selected which, while obeying the 'local property' criterion, i.e. not extending over more than half the character area, and which tend to divide the set of character classes into two large groups as defined by the observed fitting or otherwise of the AO's sampling points. No conscious effort was made at this stage to eliminate correlated behaviour between individual AO's although geometrical similarity was avoided. Ultimately it was also necessary to introduce AO's designed to resolve specific ambiguities discovered when attempting to discriminate particular character classes. The number of points used in any AO was restricted to four points of normal-sense, $d(\mathrm{r})$, plus a maximum of one point of inverse-sense, $1/d(\mathrm{r})$.

The limit of four normal-sense points in an AO was imposed after the following consideration.

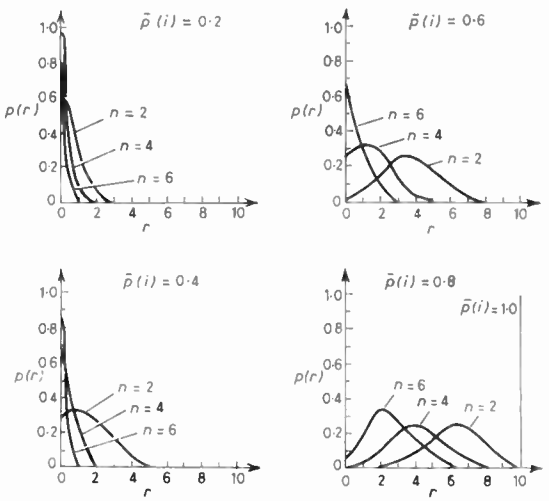


Fig. 9. Showing the probability of r fits $p(r)$ for a range of values of n against the character limb mean density $p(i)$.

(i) According to the expression (1) the value of an AO measure is proportional to the amount of coincidence of the AO with a particular character and the nth power of the character density ρ, where n is the number of normal-sense points in the AO. The area of coincidence is proportional to line thickness or line thickness squared, according to the exact form of the AO. Variation in line thickness will account for some of the variation observed in AO's applied to different specimen characters. An automatic gain control system in the character scanner compensates for variations in mean character density.

Thus the value of $A(\mathbf{a}, \mathbf{b}\ldots)$ is proportional to $w^2\rho^n$ from eqn. (2), but individual values of $A(\mathbf{a}, \mathbf{b}\ldots)$ can readily be shown to be subject to a factorial change in magnitude of approximately $(1-2\delta w)(1-n\delta\rho)$ where δw, $\delta\rho$ are the proportional variations in line thickness w and character density ρ values. The term in δw is uncontrollable and not readily assessed and cannot therefore be compensated. The term in $\delta\rho$ is similarly uncontrollable but its effect on the value of A can be limited by using the minimum useful value of n.

(ii) In practice character density is not uniform as implicitly assumed above and may have a mean density $\langle\rho\rangle$ but, in an extreme case, at the microscopic level may consist of a random distribution of isolated dots of unit density whose sampled probability $p_b = \langle\rho\rangle$.

For any value of $\langle\rho\rangle < 1$ there is a finite probability that any AO measure may be zero. The probability of an AO achieving any proportion, $p(\mathbf{r})$, of its most probable value can readily be shown to conform to a binomial distribution as a function of $\langle\rho\rangle$ and n. Typical distributions are shown in Fig. 9. The most probable fraction of the true values is in all cases identical to $\langle\rho\rangle^n$ but it will be observed that for decreasing values of $\langle\rho\rangle$ or increasing n the widths of the distributions increase rapidly and include $p(\mathbf{r}) = 0$ at a significant level in a number of cases. This case, although extreme, is representative of the condition when, due to the use of worn or dried ribbon, or low impact pressure, the ribbon weave is apparent in the character limbs.

From both these considerations it is clear that n should be kept as small as possible; however, if n is reduced below three the ability of AO's to discriminate curved and straight character limbs is lost. It was therefore concluded that n, the number of normal-sense points, should be limited to 3 or 4. The effect of adding inverted-sense points is less than that of points sampling the character limbs since the background density is in general less variable and an additional sense-inverted sampling point was allowed when required.

Having generated a set of experimental AO's in this way a sufficient, but minimal, sub-set capable of efficiently discriminating each member of the numeric set from the remainder has to be selected.

7. Design of Recognition Logic

In the current context the character property measures obtained are continuous and are presented upon an arbitrary scale determined by limb thickness and mean ink density. Thus no significance can be attached to the *absolute* magnitude of individual measures and a recognition logic for use in this situation must be based upon the *relative* magnitudes of AO's.

The set of m measures derived from a specimen character can be thought of as defining a point and hence a vector in an m-dimensional orthogonal signal space, m-space. The vector is used to define a direction in m-space its magnitude being largely irrelevant.

The vector representing an unknown character can be compared with standard vectors derived from characters of known identity, the unknown character being allotted to the class to which the 'nearest' standard vector belongs. In the present context the concept 'nearest' will be interpreted as minimum angular separation or the equivalent maximum correlation. Euclidean distance measurement cannot be used for assessing 'nearness' because of the amplitude uncertainty of the unknown character vector.

The cross-correlation coefficient r_{ab}, equal to the cosine of the angular separation θ_{ab}, of any pair of vectors a and b in m-space is given by

$$r_{ab} = \cos\theta_{ab} = \frac{\sum_m a_i . b_i}{|a| . |b|} \qquad \ldots\ldots(3)$$

If the vector a represents the set of measures obtained from the unknown character and the vector b represents the measures obtained from a known character class ($b = 0, 1, \ldots 9$), then the unknown character c_a is identified with that character c_b which maximizes r_{ab} (or minimizes θ_{ab}). The normalizing term a in eqn. (3) can be dropped since it is common for all c_b. The term b of eqn. (3) is known *a priori* and a score for each possible character class, S_b, which is proportional to the correlation r_{ab} can be determined as

$$S_b = ar_{ab} = \frac{1}{b}\sum_m a_i . b_i = \sum_m a_i . b_i' \qquad \ldots\ldots(4)$$

This is a simple linear function in which the coefficients of each terms have been normalized such that

$$(\sum b_i'^2)^{\frac{1}{2}} = 1$$

Upon substituting the values of a obtained from an unknown specimen character in eqn. (4) the most probable identity of the unknown is given by the largest value of S_b.

Linear decision functions of this general type have been applied widely and are relatively easy to solve using conventional analogue computing techniques and a maximum response selecting circuit (Fig. 10).

If it is assumed that the great majority of individual vectors from characters of the same class are highly correlated and define a narrow cone in m-space then

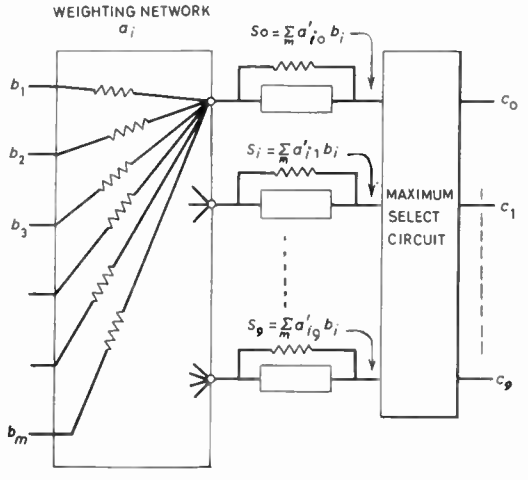


Fig. 10. Analogue computing network for the solution of linear decision functions.

the unknown vector may be compared with a single vector coincident with the axis of each cone defined for each character class. In selecting a set of characteristic property measures for use in a decision mechanism the twin requirements that the cone representing a given character class shall have the smallest possible angle and that the angular separation of the cone axes representing different character classes shall be maximized have to be satisfied. An arbitrarily chosen set of measures is unlikely to satisfy these requirements and in general a small but not necessarily exclusive sub-set of the arbitrary set is required for each character class to be recognized.

From an arbitrary set of measures the optimum sub-sets can be selected only by performing an exhaustive evaluation of all possible sub-sets. This task is clearly enormous for more than a trivially small number of possible measures. To obtain a usable solution in a short time for the current experiments recourse was therefore made to heuristic methods.

The steps in the selection of AO's for use in decision functions were as follows:

(1) AO's were ranked in order of their ability to separate the character classes. For this purpose the scale upon which each AO was represented was arbitrarily partitioned at points corresponding to 0·2 and 0·8 of this scale defined by the largest mean value for any character class. The following crude 'information value', I, was then used to rank the AO's.

$$I = p_l \log_2 p_l + (1-p_l)\log_2(1-p_1) + \\ + p_u \log_2 p_u + (1-p_u)\log_2(1-p_u)$$

where p_1 is the probability that the AO will achieve a mean value less than the lower partition, (0·2), and p_u is the probability that the AO will achieve a mean value exceeding the upper partition (0·8) over all character classes. These probabilities were estimated from the distribution of AO mean values for each character class. The possibility of statistical distributions about the means infringing the partitions was ignored.

In partitioning AO's in this arbitrary manner it was assumed that those components of the vector defining the unknown character having large or small values make a greater contribution to the discrimination between character classes than the intermediate values. Such intermediate values, although contributing some information, are in general subject to a larger statistical variation than the extreme values since they arise from 'chance fits' between the AO and the character scanned and are best neglected. Only AO measures having high or low mean values were therefore used in decision functions for particular character classes and the weight a, was restricted to values of $+1$ or -1 respectively.†

(2) Any groups of AO's whose behaviour is highly correlated over all character classes were bracketed together. In any such group the AO exhibiting the lowest variability was preferred.

(3) From the ranked list of AO's individual AO's were selected for their ability to discriminate each member of the character set. Using decision functions of the form of eqn. (4) 'scores' for all character classes were determined from the AO mean values. These scores can conveniently be displayed in a 'score matrix' (see Table 2), with which probable confusions can be readily detected and the effects of adding additional AO's to particular decision functions readily observed. AO's were initially selected on the basis of their ranking alone but eventually also on their ability to resolve persistent ambiguities apparent in the score matrix as it was evaluated. This iterative process was terminated when either no further improvement seemed possible or when the design set of material was correctly identified.

It was found in practice that to eliminate persistent ambiguities specially designed AO's of comparatively low information value had to be employed. This experience runs contrary to the assumption of some workers that the application of a maximum information criteria is a sufficient strategy to select the most valuable properties.

In order to assist the selection of AO's for use in the decision function and to determine the behaviour of AO's when applied to low quality material the following statistical measures were computed for the design character set (sets 3 to 12):

(i) The mean value of each AO, A_i, applied to k members of character class, c:

$$\mu_i(c) = \frac{1}{k}\sum_{i=1}^{k} A_i(c)$$

(ii) The variability, $\mathrm{var}_i(c)$, of A_i, within each character class, c:

$$\mathrm{var}_i(c) = \frac{\frac{1}{k}\left(\sum_{i=1}^{k}(A_i(c)-\mu_i(c))^2\right)^{\frac{1}{2}}}{\mu_i(c)}$$

(iii) The correlation, r, between all pairs of AO's, AO_i and AO_j, for each character class, c, $i = k$.

$$r_{ij}(c) = \frac{\sum_k [A_i(c)-\mu_i(c)][A_i(c)-\mu_j(c)]}{\left(\sum_k (A_i(c)-\mu_i(c)^2)\right)^{\frac{1}{2}} \cdot \left(\sum_k (A_j(c)-\mu_j(c))^2\right)^{\frac{1}{2}}}$$

† The use of weights of fixed values $+1$ or -1 is not strictly in accordance with eqn. (4), but is a simplifying approximation when only AO's having large ($>0{\cdot}8$) or small ($<0{\cdot}2$) values are used.

Table 2

Score matrix on design sets—No errors

Possible classification	Applied characters 0	1	2	3	4	5	6	7	8	9	Character sets 3–12
0	1	0	−0·3	0	−0·2	0·2	0·3	0	0·1	0·2	
1	0·4	1	−0·5	0·4	0·2	0·5	0·2	0·3	−0·1	0·1	
2	−0·9	0	1	0·4	0·2	0·2	0·1	0	0·2	−0·2	
3	−0·4	0	−0·1	1	−0·6	−0·4	0	0	−0·4	0	
4	0·1	0	0·3	0·1	1	0·2	0·3	0·2	0·3	0·2	
5	0	0	0	0	0	1	0	0	0·2	0	
6	0·4	0	−0·5	−0·2	0·3	0·5	1	0	0·1	0·3	
7	−0·1	0	0·3	0·1	0	0·1	0·2	1	0·1	0·4	
8	0·2	0·4	0·6	0·4	0·1	0·3	0	0·1	1	0	
9	0·3	0	0·1	0·1	0	0·1	0	0·5	0·1	1	

These measures are tabulated for a typical character class in Table 3.

From these measures it is clear that in spite of a considerable variability in individual AO values, due to variation in character limb thickness and ink density, their behaviour correlates highly within any character class.

It is appreciated that the heuristics used are crude and inelegant but they have proved adequate for preliminary studies and have obtained useful results. It is planned to apply more elaborate and rigorous techniques to obtain more nearly optimum feature selection and design in the near future.

8. Results

Several recognition logics were obtained and tested under the following constraints:

(*a*) only AO's containing no inverse-sense sampling points were used;

(*b*) AO's containing complementary, $1-d(\mathrm{r})$, inverse-sense sampling points;

(*c*) AO's containing the same inverse-sense sampling points as (*b*) but using the reciprocal, $1/d(\mathrm{r})$, form;

(*d*) as (*b*) but using binary instead of continuous density characters.

Test (*a*) was included to show that recognition performance is substantially improved if sampling points of inverse-sense are included and test (*d*) to confirm that representing the specimen characters on a continuous density scale results in a reduced error-rate.

The recognition logics used in (*b*) and (*c*) were derived from the logic used in (*a*), which employed the best logic obtained in ten trials, but including additional features containing inverse-sense sampling points. The recognition logic used in (*d*) is the same

Table 3

Mean values, variabilities and correlations of a number of typical AO's applied to character class 2 (10 samples)

	1	3	7	10	13	16	20	36	37	45	49	mean value	variability
1	1	0·9	1	0·9	0·5	0·9	1	1	0·7	0·8	0·6	77	0·2
3		1	1	0·8	0·7	0·1	0·9	0·9	0·8	0·8	0·7	26	0·4
7			1	0·9	0·5	1	1	0·9	0·7	0·7	0·7	25	0·3
10				1	0·3	0·8	0·8	0·9	0·5	0·5	0·5	14	0·3
13					1	0·7	0·4	0·3	0·7	0·4	0·9	1	0·9
16						1	0·9	0·9	0·8	0·8	0·7	28	0·4
20							1	1	0·6	0·7	0·5	29	0·3
36								1	0·7	0·7	0·4	107	0·4
37									1	0·9	0·6	25	0·6
45										1	0·4	11	0·6
49											1	70	0·6

Average correlation 0·84. Average variability 0·41 (for all means > (0·1 of largest mean)).

Table 4

Results of experiments (*a*), (*b*), (*c*) and (*d*)

Test	Error rates % Design sets	Test sets	Comments
(*a*)	2	25·7	No inverted sense points
(*b*)	0	10	$1-d(\mathrm{r})$ form of sense inversion
(*c*)	0	10·4	$1/d(\mathrm{r})$ form of sense inversion
(*d*)	6	Not tested	As (*b*) but on binarized material

as in (*b*) and (*c*) but using logical NOT for sense inversion. The various scaling factors used in correcting for variations in the arbitrary scale of measures at various stages in the computation of score matrices were adjusted to suit the conditions of each test. The results of these tests are given in Table 4.

From these results it is clear that the departure from the rigid ideas of auto-correlation permitting the inclusion of inverted sense points in the computation of AO's produces a great improvement in the recognition performance obtained. It will also be observed that the two forms of sense-inversion employed are only marginally different. Much the same group of individual characters were mis-recognized by both forms. The difference is not, therefore, thought to be significant so that both forms of inversion can be considered to be equivalent in effect.

The effect of representing characters on a dichotomized density scale, (*d*), is to greatly increase the error rate. The dichotomizing threshold was placed midway between the maximum and minimum peak densities in the characters, i.e. midway between peak black and peak white.

In addition to the reduction in performance in (*d*) it was also observed that the variability of measures obtained from binarized characters was some 10% greater than the variability obtained under any of the other conditions which showed no significant variation.

It will be recalled, and is obvious from the reproduction of the printed material used (see Figs. 7 and 8), that the method of testing the simulated system was to present it with a greatly exaggerated range of quality within a limited number of samples rather than to attempt to obtain an accurate figure for error rate when using a large bulk of 'typical' material. The economies realized by this method of test, which determines comparative limits for system performance under various conditions, are obvious but do not give a real estimate of error rate for material falling within prescribed quality tolerances. Suffice it to say in respect of error performance therefore that those characters which were mis-recognized when tested by the criterion of print quality laid down by various manufacturers as outlined earlier were exceeded by a large margin in all cases. Minimum variation in print parameters causing individual characters to be mis-recognized are listed in Table 5. Many of the characters correctly recognized fall very far outside even these limits.

From eqn. (2) it will be obvious that the magnitude of AO's is approximately proportional to the *n*th power of character density, where *n* is the number of sampling points in the AO of normal-sense, and hence variation in AO measures is likely to be at least *n* times the variation in character density. The variability of AO measures will be reduced with a corresponding improvement in the recognition performance if some root of the AO measure is taken.

Table 5

Print quality assessment of mis-recognized characters. The characters listed represent the minimum degradation of print quality occurring in error characters. (Main contributory factors are in bold type)

Error character	3_{18}	5_{20}	8_{30}	0_{30}	7_{13}
Factorial change in character area relative to nominal	1·0	0·9	1·4	1·4	1·05
Maximum limb density	0·57	0·55	0·95	0·95	0·50
Factorial variation of character limb density	**2·2**	**2·0**	**1·1**	**1·1**	**2·8**
Maximum background reflectance	0·87	0·87	0·87	0·87	0·87
Factorial change in background reflectance (bulk background reflectance 0·87)	1·06	1·0	1·7	1·6	1·05

To determine the effect and value of such an operation the following empirical experiment was performed. A recognition logic was designed, using the methods described above, based upon the *n*th root of AO measures. The performances of this logic and the logic used in test (*b*) above were determined when applied to AO measures of which various roots had been taken. The results of these experiments are shown in Table 6 from which it will be observed that the minimum error rate is obtained from both logics when the *n*/2th root of AO measures is used. The error rate being several times lower under this condition than that obtained with the measures used to design the recognition logics.

Table 6

Results of experiments using various roots of AO measures

	Error rate % (test sets 13–33)		
Recognition logic	$\sqrt[n]{A}$	$\sqrt[n/2]{A}$	$\sqrt[n/3]{A}$
Based on A	13·5	4·3	11·4
Based on $\sqrt[n]{A}$	20·5	5	12·9

Examination of the statistics of AO measures showed that in the *n*th root case the variability of large valued measures was considerably reduced but that low values have a higher variability when compared with AO measures of the unmodified form. On the experimental material used the *n*/2th root appears to achieve the best compromise between stability of high and low AO measures.

9. Further Experiments

Up to this point the distribution of AO sampling points has been restricted to three vertical scan lines, this restriction was imposed for engineering convenience and system simplicity, three is the minimum useful number and an increase would permit greater use to be made of horizontal character structure. This will be essential if the character set is to be extended significantly beyond the numerals.

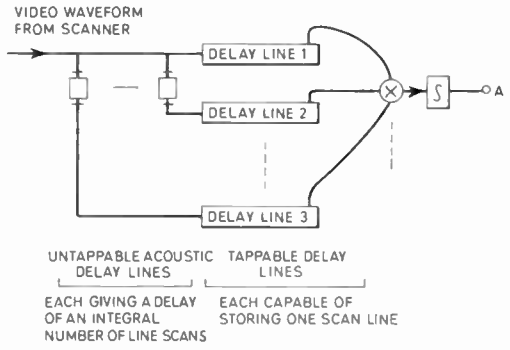


Fig. 11. Revised delay line systems.

Increasing the number of scan lines available to the AO computing networks requires a remodelling of the delay line system. The revised system is shown in Fig. 11 from which it will be seen that the tappable delay lines are now connected in parallel rather than in series, additional acoustical delay lines are then used to make-up the time difference between the various scan lines. The disadvantage of increased complexity of the delay line system brings with is some advantages in that the scan can now be a simple non-interleaved raster scan and the effective utility of the systems is increased by a factor of four since the charge-discharge cycle of the delay lines proposed in the original system is eliminated. This factor can be used either to achieve an increase in operating speed and/or a reduction in systems bandwidth.

Having increased the number of lines available to five, a selection of AO's were generated and a recognition logic determined for samples of typical output from I.C.T. card operated tabulators. AO sampling points were again generated by inspection of the character shapes. Selection of particular AO's for use in the recognition logic was performed automatically using an iterative process.

The recognition logic was based upon a modified form of cross-correlation[13] in which the weighting coefficients are initially made proportional to the mean AO values obtained from an analysis of a design set of characters as before but instead of selecting individual AO's to form a recognition logic all AO's are included in the cross-correlation network. Initially a large number of AO's are included, this number must be whittled down to an acceptably small set of, say, twenty. The sifting operation was performed by an iterative process which assessed the value of each AO in the decision mechanism by increasing each AO coefficient in turn by 40%. The effect on the recognition logic performance, in terms of error rate or tendency to error, is determined after each adjustment. If the performance is improved the adjustment is retained otherwise it is removed and the next measure adjusted. At the completion of each cycle the direction of adjustment alternates. Thus after a number of iterative cycles some measures will have been increased in significance and others reduced. The AO's of low significance can then be discarded and the set of AO's gradually reduced in size.

The method described briefly above has been successfully implemented in a computer simulation and will be described in more detail subsequently when fully developed.

10. Conclusion

From these experiments it is concluded that the use of analogue techniques combined with the facility for including inverted sense sampling points in AO's improves the performance of the proposed system beyond that obtainable with systems employing operators lacking either attribute.

Although the experiments described here have been directed exclusively at the problem of recognizing low quality printed numerals, for which two numeral styles have been used, preliminary experiments confirm that other styles can be successfully recognized with little modification of the existing AO's and decision functions. This observation suggests that the system has some potential as a multifont device, a possibility supported by the results of Kamentsky and Liu[10] who, using scanning n-tuples, have successfully recognized a variety of alphabets of different fonts.

One criticism justifiably levelled at character recognition system proposals based solely upon computer simulation studies is that however good the results obtained these cannot be directly related to the performance of a system of hardware or even that the system can never be realized in hardware form other than a digital computer. The simulation studies described here were undertaken simultaneously with an engineering study of the proposed system. Comparison of AO's evaluated by the simulation and experimental hardware agree within close limits, confirming that the simulation is valid in terms of practical engineering. As a result of these simultaneous studies a full-scale engineered version of the system is now under construction and will be reported upon in due course.

11. Acknowledgments

The author gratefully acknowledges much encouragement and advice from his colleagues and in particular Drs. M. B. Clowes, T. Lukes and A. M. Uttley and Mr. M. D. Armitage. He also wishes to acknowledge the co-operation of Dr. B. A. Wichmann, who devised more formal approaches to the design of recognition logics.

The work described above has been carried out as part of the research programme of the National Physical Laboratory.

12. References

1. R. K. Gerlach, 'Wide-tolerance optical character recognition for existing printing mechanisms', pp. 93–114, 'Optical Character Recognition' (Spartan Books, Washington, 1963).
2. C. H. Jones, J. F. Kellaher, J. M. Dillard and S. L. Bernstein, 'Fourier Transform Methods for Pattern Recognition', AFCRL–62–512, Air Force Cambridge Research Laboratories, Bedford, Mass., U.S.A., 1962.
3. F. L. Alt, 'Digital pattern recognition by moments', pp. 153–180, 'Optical Character Recognition' (Spartan Books, Washington, 1962).
4. W. Doyle, 'Operations useful for similarity invariant pattern recognition', *J. Ass. Computing Machinery*, **9**, p. 259, April 1962.
5. L. P. Horwitz and G. L. Shelton, Jr., 'Pattern recognition using autocorrelation', *Proc. Inst. Radio Engrs*, **49**, p. 175, January 1961.
6. M. B. Clowes and J. R. Parks, 'A new technique in automatic character recognition', *Computer J.*, **4**, pp. 121–128, February 1961.
7. M. B. Clowes, 'The use of multiple autocorrelation in character recognition', pp. 305–318, 'Optical Character Recognition' (Spartan Books, Washington, 1962).
8. W. W. Bledsoe and I. Browning, 'Pattern recognition and reading by machine', pp. 225–232, Proc. 1959 Eastern Joint Computer Conference.
9. W. H. Highleyman and L. A. Kamentsky, 'Comments on the *n*-tuple pattern recognition method of Bledsoe and Browning', *Trans. Inst. Radio Engrs on Electronic Computers*, EC-**9**, p. 263, February 1960.
10. L. A. Kamentsky and C. N. Liu, 'Computer automated design of multifont print recognition logic', *IBM J. Res. Development*, **7**, pp. 2–13, January 1963.
11. L. Uhr and C. Vossler, 'A pattern recognition program that generates, evaluates and adjusts its own operator', pp. 555–569, Proc. 1961 Western Joint Computer Conference.
12. W. W. Bledsoe and C. L. Bisson, 'Improved memory matrices for the *n*-tuple pattern recognition method', *Trans. I.R.E.*, EC–**11**, p. 414, June 1962.
13. B. A. Wichmann, Private communication.

Manuscript first received by the Institution on 25th November 1966 and in final form on 14th April 1967. (Paper No. 1133/C96.)

Letter to the Editor

A Suggested Prefix and Symbol for the Sub-multiple 10^{-21}

SIR,

The October 1962 meeting of the International Committee on Weights and Measures adopted two new prefixes, femto and atto, for denoting 10^{-15} and 10^{-18} respectively. One might have expected that these two prefixes would have sufficed for some considerable time to come; however, in a recent article on satellite communications† it was estimated that the noise power at the receiver for the Mars probe *Mariner IV* was 5×10^{-21} watt.

This suggests that there is a need for a prefix and symbol for the sub-multiple 10^{-21} and the prefix banto, symbol b, is proposed. Thus, for the instance quoted, the power would be stated as 5 bW.

The prefix banto is derived from the word bantam which has been used elsewhere (e.g. boxing) to convey the idea of smallness. In the English-speaking community the adoption of the prefix banto and its symbol b would lead to a convenient mnemonic as the symbols for femto, atto, banto would be f, a, b, respectively.

G. MAY (*Associate*)

Ministry of Technology,
Radio Department,
Royal Aircraft Establishment,
Farnborough, Hampshire.

18th July 1967.

† W. Arens, "Übertragungsverfahren bei Satellitenverfindungen" (Transmission methods for satellite links), *Nachrichtentechnische Z.*, **20**, No. 1, pp. 7–11, 1967 (R.A.E. Library Translation 1236).

U.D.C. 537.226

New Thin-film Resistive Memory

By
J. G. SIMMONS,
B.Sc., Ph.D., F.Inst.P.†
AND
R. R. VERDERBER,
B.Sc., M.Sc.‡

Summary: A new thin-film metal–insulator–metal device is described. After the insulator has undergone a forming process, which consists of the electrolytic introduction of gold ions from one of the electrodes, its conductivity is observed to have increased quite markedly. In addition the sample displays negative-resistance and memory phenomena. It is shown that under the appropriate switching conditions the device can be used as a non-volatile analogue memory with non-destructive read-out. The theory of operation of the device is also presented.

1. Introduction

Since the advent of microelectronics and thin-film circuitry there has been a search for a thin-film active device.[1–6] Most of these efforts have been oriented towards a thin-film device which performs essentially the same function as existing solid-state devices. In the event of such devices being proven commercially feasible one is immediately faced with formidable obstacles, and the reason for this is two-fold. Firstly, materials in thin-film form have electrical characteristics that are usually inferior to the crystalline forms; secondly, the silicon technology is in such an advanced state that it is difficult to see thin-film devices competing effectively with equivalent solid-state devices.

Thin films, however, often manifest electrical characteristics not seen in the crystalline state; these characteristics can be exploited to produce devices which perform functions not available in existing solid-state devices. Thus we see thin-film devices supplementing, rather than competing with existing solid-state devices.

Our studies on thin-film metal–insulator–metal sandwiches have revealed that after undergoing a forming process, the insulator develops appreciable conductivity and a V–I characteristic which manifests a pronounced d.c. negative resistance region. In addition the device can, by the application of suitable voltage pulses, be made to exhibit a continuum of reversible, resistive, memory states. It is the object of this paper to describe in some detail the characteristics and mode of operation of this device, and to illustrate its potential.

2. Fabrication

The devices are fabricated by the vacuum deposition of successive layers of aluminium, silicon monoxide, and gold on to a 3 in × 1 in (7·5 cm × 2·5 cm) glass (Corning 7059) substrate at a pressure of approximately 10^{-6} torr. A typical array is illustrated in Fig. 1. The thickness of the aluminium and gold electrodes is not critical, but is usually greater than 2000 Å in order to limit the lead resistivity to a few ohms. The silicon monoxide is deposited at a rate of about 4 Å/s, and its thickness must be controlled to obtain layers between 300 and 3000 Å thick.

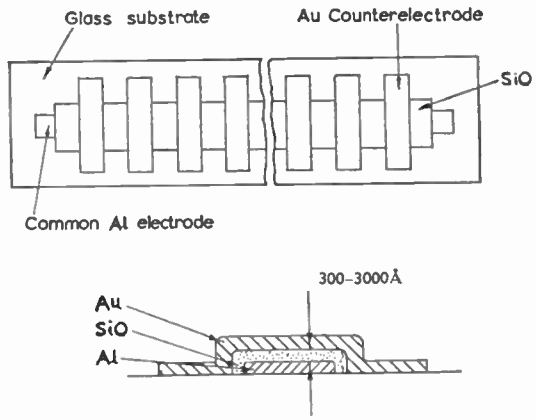


Fig. 1. Device configuration.

3. Electrical Characteristics

All the electrical characteristics described below were obtained on devices encapsulated in a partial vacuum of up to about 500 μm Hg.

3.1. *Forming Process*

It is necessary to 'form' a virgin sample in order to obtain the desired characteristic. This forming process consists of cycling the device between 0–12 V with the gold electrode *positively* biased. Initially, before cycling, the device has a very high impedance, as is to be expected from what is essentially a capacitor. The effect of forming is to induce a change in the conductivity of the device.

The forming mechanism is readily shown to be an electrode effect and attributable to the gold rather

† Formerly with Standard Telecommunication Laboratories Ltd., Harlow, Essex; now at the Department of Physics, University of Lancaster, Lancaster.

‡ Standard Telecommunications Laboratories Ltd., Harlow, Essex.

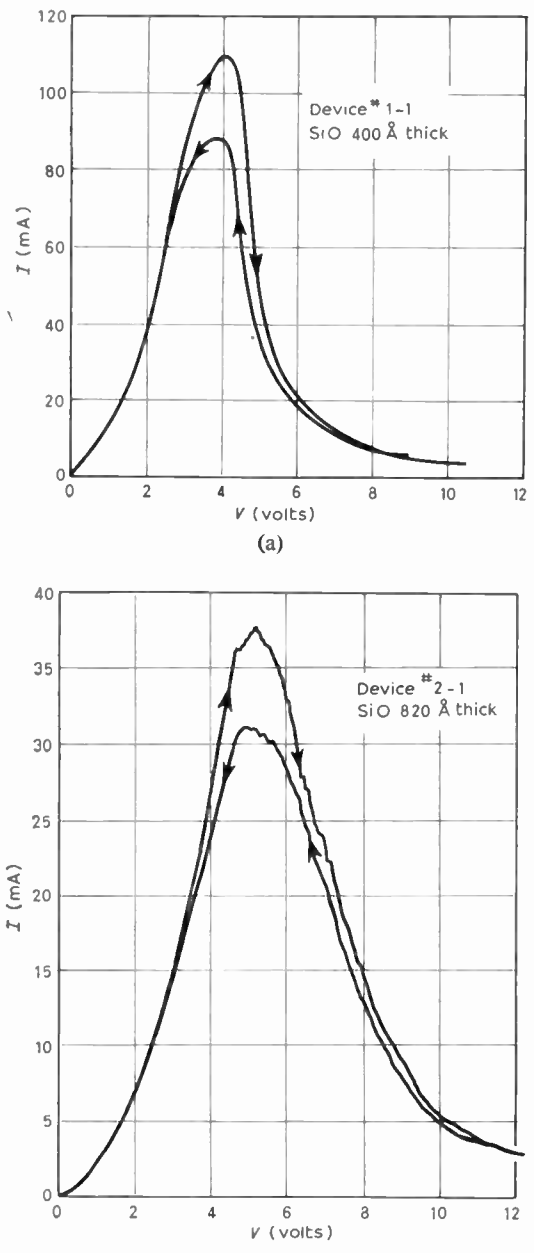


Fig. 2. Current-voltage (d.c.) characteristic of (a) 400 Å thick insulator (b) 820 Å thick insulator device.

than the aluminium electrode, because an Al–SiO–Al device will not form, whereas an Au–SiO–Au device will. Accumulated data[7] offer incontrovertible evidence that the forming mechanism is due to the injection into the silicon monoxide of positively charged gold ions from the gold electrode. It has also been observed that trivalent metals, e.g. aluminium and indium, tend to be neutral (will not form), and that monovalent metals, e.g. gold, silver and copper, tend to be active (will form).

3.2. *D.C. Conductivity*

The d.c. characteristic of a formed sample is shown in Fig. 2. The two particularly interesting features of this characteristic are: (i) the pronounced negative-resistance region in the 4–8 V range, which has a peak-to-valley current ratio of typically 100:1 but can be as high as 1000:1 and (ii) the voltage at which the peak current occurs is virtually independent of insulator thickness, and insulator and electrode materials. The device characteristic is symmetrical about the origin.

The characteristic in the voltage range zero to approximately 2·75 V, which we will designate as the threshold voltage, V_T, is extremely stable and also virtually temperature independent, i.e. the impedance increases by only approximately 10% when the device temperature is reduced from room to liquid-nitrogen temperature, and in this region the V–I characteristic is described by the following relationship:

$$I = K \sinh kV \qquad \text{.....(1)}$$

where K and k are constants.

3.3. *A.C. Characteristic*

In the voltage range $0 - V_T$ the a.c. ($\gtrsim 1000$ Hz) characteristic follows the d.c. characteristic quite faithfully, but beyond V_T, deviates from it in a manner shown in Fig. 3. In the range $V_T < V \lesssim 8$ V a distinct a.c. V–I characteristic can be associated with every voltage amplitude, the overall impedance of the characteristic for a given voltage excursion increasing with increasing voltage amplitude. These characteristics do not manifest negative resistance, although the locus of their end points for increasing values of voltage beyond 4 V generates the d.c. characteristic. The device resistance to a.c. is thus a function of

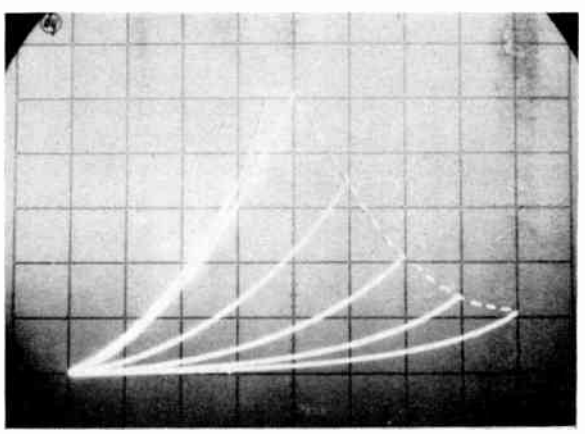


Fig. 3. Dynamic current-voltage characteristic (1000 Hz.). Dotted line is d.c. characteristic for device. (Y-axis: 5 mA/div., X-axis: 1 V/div.)

voltage amplitude, V_a, but the V–I relationship can still be expressed by an equation similar to (1) but with K and k functions of voltage amplitude thus:

$$I = K(V_a) \sinh k(V_a)V, \qquad \text{.....(2)}$$

in which both $K(V_a)$ decrease in magnitude with increasing V_a. Typical values for $K(V_a)$ are given in Table 1.

Table 1

Values of $K(V_a)$ and $k(V_a)$ as function of V_a

V_a	$K(V_a) \times 10^3$	$k(V_a)$
3	5·3	0·75
4	4·2	0·60
6	1·1	0·50
8	0·19	0·46

As an example, for a maximum voltage amplitude of 6 V the V–I characteristic of the device described by Table 1 is

$$I = 1{\cdot}1 \times 10^{-3} \sinh 0{\cdot}5\,V \qquad \text{.....(3)}$$

Beyond about 10 V or so, K and k are independent of voltage excursion.

4. Memory Characteristics

4.1. *Memory States and Dead Time*

If a voltage of 8 V is applied to the device and then reduced to zero in a time of about 0·1 ms, i.e. an 8 V pulse with a trailing edge faster than 0·1 ms, and then the voltage reapplied but of magnitude less than the threshold voltage V_T, it is found that instead of describing the characteristic OA, the device characteristic has changed to that described by the curve OE′, as shown in Fig. 4. The device will remain in this induced or high-impedance memory state indefinitely, provided that no direct or alternating peak interrogation voltages in excess of V_T are applied; thus the memory state has non-destructive read-out. The high-impedance memory state can be stored without electrical power applied for indefinite periods of time and is thus also non-volatile. Erasure of the above memory state is effected by applying a voltage in excess of V_T, whereupon the device reverts to its original low-impedance state corresponding to the curve OA.

As described the device has potential as a binary memory. In fact, it also functions as an analogue memory, for it can be stored in any of the continuum of impedance states existing between the high- and low-impedance memory states described above. Thus the state OC′ shown in Fig. 4 is stored by the application of a 6 V pulse with trailing edge faster than 0·1 ms. These memory states are also non-volatile and can be read out non-destructively, provided that voltage in excess of V_T is not applied to the device.

It will be apparent from the foregoing that the device is stable in the region of the V–I characteristic bound by the high-impedance, low-impedance and threshold voltage characteristics. The region bound by the high impedance, threshold voltage and negative resistance characteristics is unstable, but any point on the negative resistance characteristic is stable. These results are illustrated schematically in Fig. 5; note that the threshold voltage is not constant voltage but increases slightly with decreasing impedance of the various memory state.

Switching from the low to the high-impedance state can be accomplished in a time as short as two nanoseconds, and from the high to the low-impedance

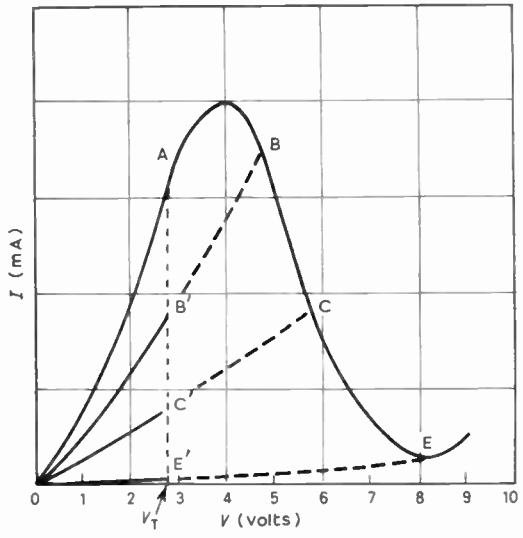


Fig. 4. Schematic illustrating various memory states and the threshold voltage.

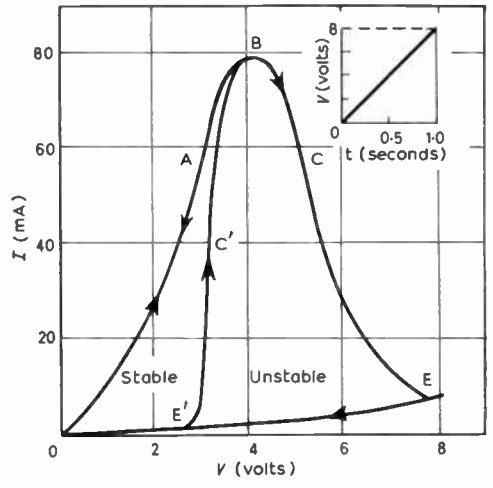


Fig. 5. Schematic illustrating the stable and unstable areas of the V–I characteristic.

state in one hundred nanoseconds although 100 ns and one microsecond respectively are typical. It would appear, then, that a repetition switching rate of approximately once in 100 ns is possible. However, this is not the case, because it is found that although the device may be switched from the high to the low-impedance state and back again in approximately 100 ns, any immediate attempt to switch it again is inhibited for a period of time that varies from microseconds to milliseconds dependent upon the temperature and the method of fabrication. This phenomenon is designated the dead time. The dead time is responsible for the fact that the a.c. *V–I* characteristics shown in Fig. 3 do not switch to the low-impedance state as the applied voltage exceeds V_T, and this is because the a.c. cycle-time is greater than the dead time.

4.2. *Further Observations on the Analogue Memory States*

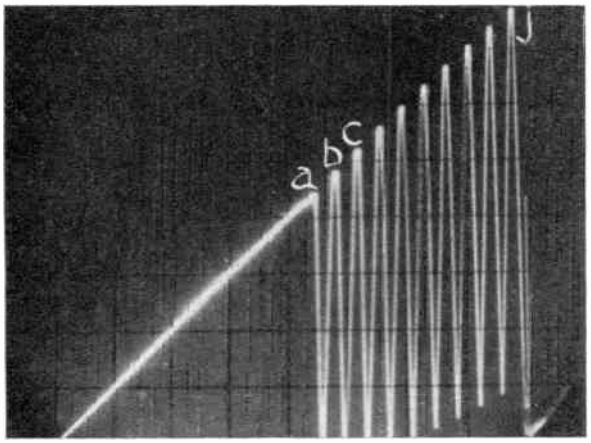
By applying the modulated triangular wave-form shown in Fig. 6(a) to the device the *V–I* characteristics shown in Fig. 6(b) are obtained. These characteristics are explained as follows. The device is initially set into its high-impedance state; thus during the initial period of voltage application, where the voltage rises linearly to value V_a which is slightly greater than V_T, the high impedance *V–I* characteristic OA′, Fig. 6(b), is generated. If the threshold voltage were constant for the whole continuum of memory states, the device would switch directly from the high to the low-impedance state as the linearly rising voltage exceeded V_T. However, since the threshold voltage increases for each low impedance state, as the voltage exceeds the high-impedance threshold voltage the device is forced into a continuous switch mode, switching in a series of infinitesimally small steps to successively lower impedance states, and path A′A is traced out. This process prevails until all memory in the sample has been erased.

On complete erasure of the memory the applied voltage drops to zero, thus generating the low-impedance *V–I* characteristic OA. The device is now subjected to the voltage pulse 'b' (Fig. 6(a)) which in rising to its maximum value generates the *V–I* characteristic OB′ which is simply the characteristic OA extended by an amount AB′. However, the operating point B′ is unstable, since it lies off the d.c. negative resistance curve, which is shown dotted in Fig. 6(b) (cf. with the negative resistance characteristic of Fig. 7(b)). The device thus switches from B′ to B—the latter point being located on the d.c. negative-resistance characteristic—and the reciprocal of the slope of the line B′B is equal to the external resistance of the circuit. As the pulse 'b' falls from its maximum value to zero, it generates the *V–I* characteristic BO which is of slightly higher impedance than OA; thus the pulse 'b' generates clockwise the hysteresis loop OB′BO. The same sequence of events observed for pulse 'b' occur during the application of the subsequent voltage pulses 'c'. . . 'j', but since each successive pulse is slightly larger than the last one, they each generate a characteristic *V–I* hysteresis loop of which the overall impedance increases with increasing pulse amplitude. The reason that the device does not switch when the rising voltage of each pulse exceeds V_T is that the dwell time at zero voltage is less than the dead time.

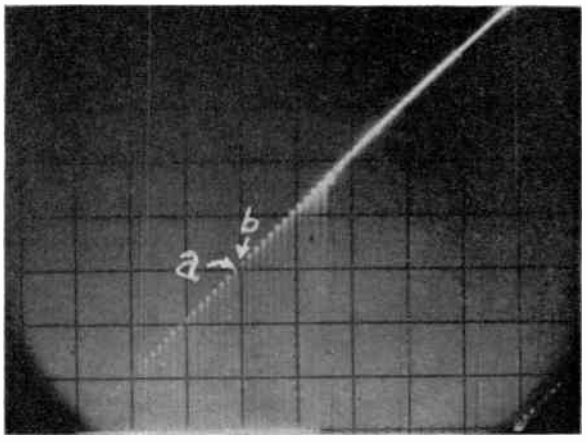
Figure 6(b) illustrates the *induction* of several of the continuum of possible memory states by successive pulses of increasing amplitude and can be described as a subtraction process, since the device is being driven into memory state of *decreasing* conductivity. Figure 7(b) displays the *addition* analogue of the subtraction process. To obtain these characteristics the device is first set into the high-impedance state and the modulated triangular waveform shown in Fig. 7(a) applied to it. The point A on the threshold characteristic, which is indicated by the dotted line (cf. the threshold characteristic of Fig. 6(b)), corresponds to the operating point for the maximum voltage of pulse 'a', which is just equal to the threshold voltage of the high-impedance state. Thus voltage pulses of amplitude less than pulse height 'a' (Fig. 7(a)), simply interrogate the high-impedance state OA. The pulse 'b', however, has a voltage amplitude slightly in excess of that of pulse 'a', and generates the characteristic OB′, which is simply an extension of the high-impedance characteristic OA. The point B′ is, however, in the unstable region of the *V–I* characteristic (see Fig. 5) so the device switches from the operating point B′ to B which is located on the threshold voltage characteristic and is thus stable. As the pulse 'b' falls from its maximum value to zero, the characteristic BO is generated and is of slightly lower impedance than OA; thus the pulse has generated the anti-clockwise *V–I* loop OAB′BO. The same sequence of events observed for pulse 'b' occur during the application of the subsequent voltage pulses, but since the amplitude of each successive pulse is slightly greater than the last, each of them generates a characteristic anti-clockwise *V–I* loop of which the overall impedance is lower the greater the pulse amplitude. After the application of the last pulse the device is subjected to the linear rising voltage portion starting at $V \simeq 4{\cdot}5$ V (see Fig. 7(a)), which is responsible for generating the negative resistance characteristic shown in Fig. 7(b).

4.3. *Preferred Method of Switching*

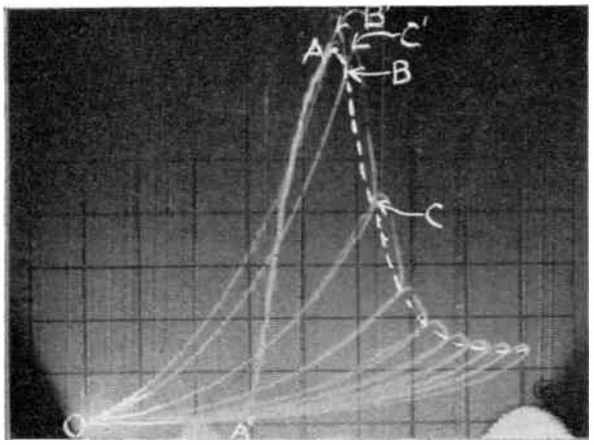
The use of variable voltage pulses is not a particularly suitable way of switching the device into the analogue memory states. A superior method is to use a constant voltage pulse of height just in excess of



(a) Voltage waveform (X-axis: 20 ms/div., Y-axis: 1 V/div.).



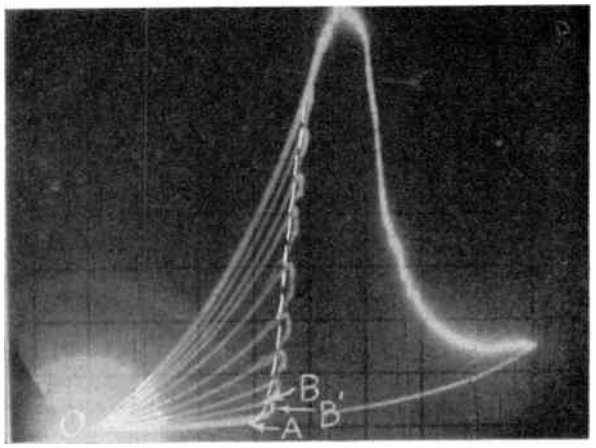
(a) Voltage waveform (X-axis: 2 ms/div., Y-axis: 1 V/div.).



(b) *V–I* characteristic for waveform (a)

(X-axis: 1 V/div., Y-axis: 2 mA/div.). Dotted line corresponds to d.c. negative resistance characteristic.

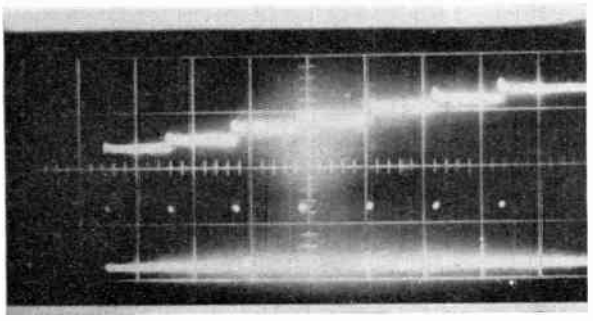
Fig. 6.



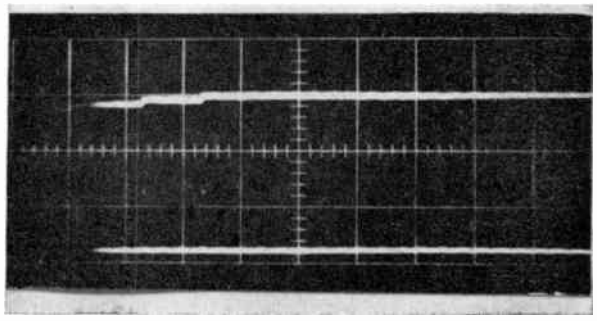
(b) *V–I* characteristic using waveform shown in (a)

(X-axis: 1 V/div., Y-axis: 2 mA/div.). Dotted line corresponds to threshold voltage characteristic.

Fig. 7.



(a)



(b)

Fig. 8. Oscillogram showing various analogue states obtained using a constant voltage pulse and pulse width of: (a) 10 μs, (b) 3 μs (X-axis: 1 ms/div., Y-axis: 1 mA/div.) Reading voltage was one volt.

the threshold voltage and width of less than the switching time, which when applied to the sample will cause it to undergo only a partial switch.[8] By using a sufficiently narrow pulse many pulses are required to switch the device from the high to the low-impedance state and vice versa. In Fig. 8(a) we have illustrated an oscillogram resulting from applying a series of pulses each one of width 10 μs to a 'slow' device, it is seen that for the six pulses the sample impedance has changed from $2{\cdot}4 \times 10^{-3}\ \Omega$ to $3{\cdot}4 \times 10^{-3}\ \Omega$. Pulse widths of ~ 1 μs are required to obtain similar incremental steps in going from the low to the high-impedance state on the same device. In Fig. 8(b) we illustrate the effect of voltage pulses of width of 3 μs on the same device; the incremental current changes are about one-third of those shown in Fig. 8(a). Figure 8(b) also shows the stability of the analogue state with time after the application of the third pulse.

5. Theory

The remaining sections of the paper are devoted to the theory of operation of the device. It would, however, be out of place here to give a detailed account of the theory of the conduction and memory mechanisms, as it is fairly complex and has been treated fairly exhaustively elsewhere;[9] we will therefore limit the discussion to an outline of the ideas involved in the explanation of the phenomena.

6. Energy Band Scheme

In Section 3.1 it was pointed out that the experimental evidence suggests that the insulator is formed by the electrolytic induction of gold ions from the gold electrode. If the ions were injected into a perfect lattice, they would have a discrete energy level associated with them. However, the energy levels associated with ions in a highly disordered or amorphous structure, which our evaporated films must have, will not be discrete, since there is no consistency between nearest neighbours and next-nearest neighbour configuration etc., but instead will form an impurity band in the forbidden band of the insulator, as shown in Fig. 9(a).

The energy band model of the metal–insulator–metal system is drawn in Fig. 9(a) and shows the metal–insulator barrier, ϕ_o, and the height of the top of the localized allowed states, ϕ_i. The positive depletion region shown in the diagram is formed since the work function of the metal (Ψ_m) is greater than the work function of the insulator (Ψ_i + electron affinity).

7. D.C. *V–I* Characteristic

7.1. *Low-impedance Characteristic*

When the density of injected ions is sufficiently high, the energy levels communicate, quantum mechanically, with each other, which means that electrons can pass

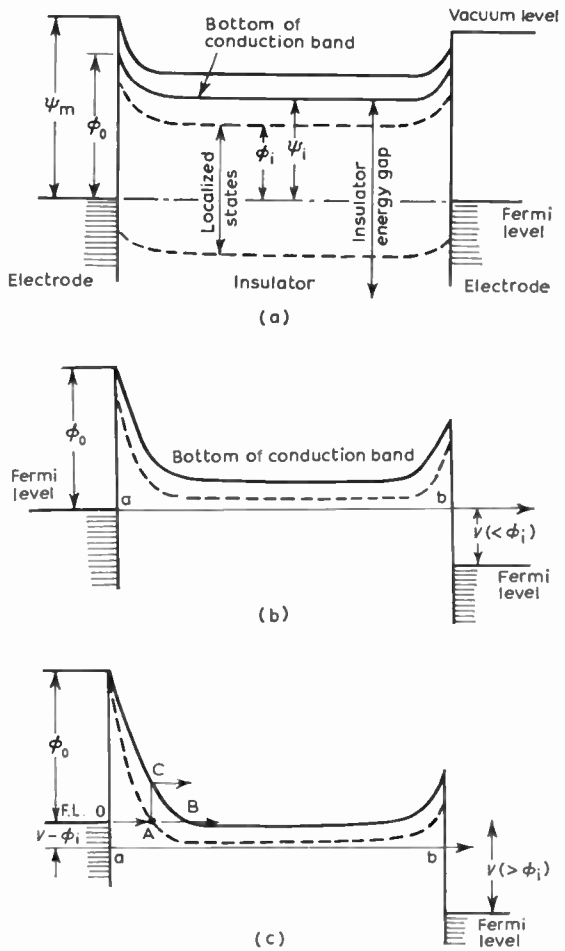


Fig. 9. Energy diagram of formed insulator for (a) $V = 0$, (b) $V < \phi_i$ and (c) $V > \phi_i$.

between adjacent and energetically equivalent levels by means of the tunnel effect as shown in Fig. 9(b). Using this tunnel-hopping process it is possible to derive the *V–I* relation for the device in the low-impedance state, with the result[9]

$$I = A \exp\left[-1{\cdot}025\, s(\phi_0 - F_0 s/2)^{\frac{1}{2}}\right] \times \times \sinh\left[0{\cdot}256\, s^2 V/\lambda(\phi_0 - F_0 s/2)^{\frac{1}{2}}\right] \quad \ldots\ldots(4)$$

where s is the average distance between tunnelling sites, λ is the depth of depletion region, and F_o is the field at the metal–insulator interface. In deriving this equation it was assumed that the electron energy is conserved during the process and that the potential barrier to the tunnelling process exists at the cathode–insulator interface, i.e. the barrier of height ϕ_0. Equation (4) fits the experimental curve below the peak current quite well.

7.2. *Negative-resistance Characteristic*

The mechanism resulting in the negative-resistance region can be best seen by reference to Fig. 9(c). For $V > \phi_i$, because the total energy of the tunnelling

electrons is conserved in the tunnel process, i.e. the electron path is horizontal on an energy diagram; it is only energy levels in the cathode positioned below the lowest point of the top of the localized band that can contribute electrons to the circulating current. Electrons emanating from levels above this point can penetrate the insulator only partially; for example electrons injected from the Fermi level of the cathode will penetrate the insulator only as far as the point A at the top of the localized band, because beyond this point there are no available states for further tunnelling to occur. The greater the bias voltage, the lower are the energy levels below the Fermi level which contribute electrons to the conduction process. Thus the barrier 'seen' by an electron entering from level 'a' (Fig. 9(c)) is given by†

$$\phi'_0 = \phi_0 + V - \phi_i \qquad \text{.....(5)}$$

Substituting ϕ'_o for ϕ_o in eqn. (4) yields

$$I = A \exp\left[-1{\cdot}025\, s(\phi_0 + V - \phi_i - F_0\, s/2)^{\frac{1}{2}}\right] \times$$
$$\times \sinh\left[0{\cdot}256\, s^2\, V/\lambda(\phi_0 + V - \phi_i - F_0\, s/2)^{\frac{1}{2}}\right] \quad (6)$$

From eqn. (6) it is seen that the current in the system decreases monotonically with increasing voltage bias for $V > \phi_i$. In Fig. 11, eqns. (4) and (6) are shown plotted using the parameters $s = 20$ Å, $\phi_o = 3$ V, $F_o = 10^7$ V/cm, $\lambda = 10^2$ Å, $\phi_i = 2$ V, and it is seen that the resulting curve is in good qualitative agreement with experiment. The sharp peak at $V = \phi_i$ is due to the fact that in deriving eqns. (4) and (6) the electrons are assumed to have come from a single level. Since ϕ_i is independent of the insulator thickness, the voltage at which the onset of the negative region occurs, that is $V > \phi_i$ will be independent of the thickness of the device, as is experimentally observed.

8. Memory

8.1. *Stored Charge Hypothesis*

We have seen in all cases when a new memory state is induced, the accompanying *V–I* sweep manifests hysteresis. In changing from a low to a high-impedance state, a clockwise *V–I* loop is generated, and when a memory state is erased, the accompanying *V–I* loop is generated anti-clockwise.

This hysteresis effect is indicative of stored energy in the case of a change from a low to a high-impedance state, and of energy release in the case of a change from a high to a low-impedance state. One mode by which energy can be stored in the insulator is by the storage of electronic charge, and our model of the memory phenomena is based on this hypothesis.

The stored electrons are thought to be those injected from energy levels positioned within $V-\phi_i$ electron-volts of the Fermi level (see Fig. 9(c)); that is, the electrons that can travel through the insulator only as far as the top of the localized band—for example, as far as the points a b c d in Fig. 10(a). This reasoning is supported by the fact that memory storage only occurs for $V > \phi_i$ (see Fig. 10(a)). Immediately after reducing the voltage to zero, sufficiently quickly so that the electrons do not have time to vacate the levels in which they are initially stored, we have the energy scheme illustrated in Fig. 10(b). It will be noted that the electrons are now stored in higher energy states than into which they were injected. Thus, for $V > \phi_i$ *rapid* removal of the voltage results in energy storage, that is, hysteresis in the *V–I* characteristic as observed. The electronic configuration shown in Fig. 10(b) is not a stable one, and the electrons will eventually tunnel out of these storage levels and drift towards the centre of the insulator, because this is the direction of decreasing barrier height and is the preferred direction of motion as far as the tunnelling process is concerned. Once having reached the centre of the insulator they are confined there, because to the left and right of them exists the depletion regions; that is, regions of barriers of increasing potential energy.

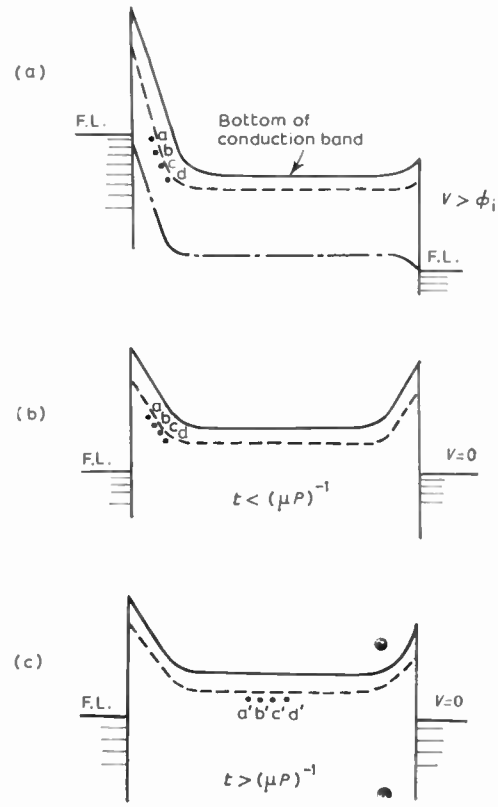


Fig. 10. Energy diagram illustrating memory, (a) charge storage, (b) stored electrons for $V = 0$ and $t < (\mu P)^{-1}$, and (c) stored electrons for $V = 0$ and $t \gg (\mu P)^{-1}$.

† All energies measured in electron-volts.

The effect of the stored charge is to alter the field at the electrode insulator interfaces by an amount ΔF, given by

$$\Delta F \simeq eN_s/K\varepsilon_0 \text{ volts/metre}$$

This has the effect of increasing the limiting barrier ϕ_0 at the metal-interface by an amount $\Delta F\, s/2$, where s^{-3} is the approximate ion density in the insulator; thus with N_s stored electrons the limiting barrier is now given by

$$\phi_0'' = \phi_0 + eN_s s/2K\varepsilon \qquad \ldots\ldots(7)$$

Substituting ϕ''_0 for ϕ_0 in eqn. (4) shows that the current density is lower when charge is stored, that is, the impedance of the system has increased, as observed.

8.2. *Switching Process*

It is only when the Fermi level of the cathode lines up with the levels occupied by the electrons, that is, when $V \simeq \phi_i$ that the stored electrons can escape from the insulator. This is because there is a large difference in the cathode electron density (where the concentration is relatively high) and the stored electron density (which is relatively low), and it is this concentration gradient that is responsible for driving out the stored electrons. Thus, when $V \simeq \phi_i$ which corresponds to the voltage at which the peak current flows in the system (see Fig. 11), the device switches, as is observed.

8.3. *Area of V–I Hysteresis Loops*

In the course of travelling to the centre of the insulator the electrons have to give up energy (this is apparent by comparing Figs. 10(b) and 10(c)) to the lattice in reaching the centre of the insulator. Thus the energy released during memory erasure is less than that stored during memory induction, which explains why the hysteresis loop accompanying erasure is smaller than that accompanying the storage.

8.4. *Dead Time*

The origin of the dead time lies in the fact that the electrons take a finite time to reach the centre of the insulator, i.e. the time it spends in each site $(\mu P)^{-1}$ times the number of jumps, where μ is the attempt-to-penetrate frequency and P is the probability of the transition. If the voltage inducing a memory state is removed and reapplied in a time less than it takes the stored electrons to do this, the device will not switch. In order to give some physical insight into the process by which the stored electrons move, we have to consider the fact that they have to give up energy to reach the centre of the insulator and they do this by way of phonon-assisted tunnelling. This process is a much more improbable process than that of the direct tunnelling process responsible for the circulating electronic motion, since P must now include the Boltzmann factor $\exp(-\Delta E_T/kT)$ where ΔE_T is the energy separation of the tunnelling sites; another way of stating this is that the drift velocity associated with phonon-assisted tunnelling is relatively slow compared to that associated with direct tunnelling. This in turn means that the time taken for the stored electrons at a b c d to reach the points a′ b′ c′ d′ (see Fig. 10(c)) is relatively long compared with the time taken for the circulating electrons to pass through the insulator which is in qualitative agreement with observation.

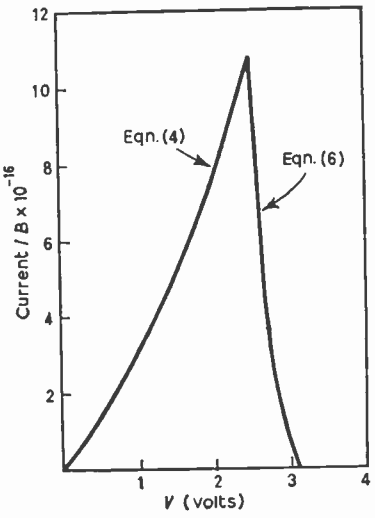


Fig. 11. Theoretical d.c. current-voltage characteristic.

9. Conclusion

We have described a thin-film device which exhibits a pronounced negative resistance region and binary and analogue memory, and it is apparent that there are many possible applications for the device. We have used the device as a buffer memory, bistable element with memory and divide-by-two circuit.[8] The devices are very cheap to produce, typically 95% of all devices have the formed characteristic, and packing densities of up to $10^4/\text{in}^2$ can be visualized.

There are however many problems to be solved before one can confidently start thinking in terms of a productive device. Limitations on the present devices are that they have to be operated in a partial vacuum, although devices have been made which operate, but at present erratically, at atmospheric pressure, and stability is still a problem—however, the device is very much in its infancy, and, as yet, relatively unexplored. Most of our studies have been confined to the Al–SiO–Au system, primarily to gain an understanding of the mechanism of operation. Theory suggests that the effect should be seen in other materials, and, indeed, we have seen the effect in CdS, MgF_2 and Al_2O_3 to mention a few, and the characteristics are remarkably similar to those of the SiO system, i.e. threshold

voltages, voltage at which peak current occurs, etc. Thus, because the devices are exceedingly cheap to manufacture and in view of their unique characteristics, if the present problems can be solved, they appear to have a bright future.

10. Acknowledgments

The support and encouragement of Dr. P. White is gratefully acknowledged. The authors also acknowledge stimulating discussions with Mr. E. A. Fuell and the technical assistance of Mr. S. Shepard who designed and built the waveform generator used to produce Figs. 6 and 7.

11. References

1. C. A. Mead, 'The tunnel-emission amplifier', *Proc. Inst. Radio Engrs*, 48, pp. 359–61, March 1960.
2. C. A. Mead, 'Operation of tunnel emission device', *J. Appl. Phys.*, 32, pp. 646–52, April 1961.
3. P. K. Weimer, 'The "TFT"—a new thin-film transistor', *Proc. I.R.E.*, 50, pp. 1462–9, June 1962.
4. D. V. Geppert, 'A metal base transistor', *Proc. I.R.E.*, 50, pp. 1527–8, June 1962.
5. M. M. Atalla and R. W. Soshea, 'Hot carrier triode with thin-films metal base', *Solid-State Electronics*, 6, pp. 245–50, June 1963.
6. J. P. Spratt R. F. Schwarz and W. M. Kane, 'Hot electrons in metal films: injection and collection', *Phys. Rev. Letters*, 6, pp. 341–2, April 1961.
7. R. R. Verderber, J. G. Simmons and B. Eales, 'Forming process in SiO thin films', *Phil. Mag.* (to be published).
8. E. A. Fuell, private communication.
9. J. G. Simmons and R. R. Verderber, 'New conduction and reversible memory phenomena in insulating films.' *Proc. Roy. Soc.*, A, (to be published).

Manuscript received by the Institution on 19th April 1967. (Paper No. 1134.)

STANDARD FREQUENCY TRANSMISSIONS

(Communication from the National Physical Laboratory)

Deviations, in parts in 10^{10}, from nominal frequency for July 1967

July 1967	24-hour mean centred on 0300 U.T.			July 1967	24-hour mean centred on 0300 U.T.		
	GBR 16 kHz	MSF 60 kHz	Droitwich 200 kHz		GBR 16 kHz	MSF 60 kHz	Droitwich 200 kHz
1	− 300·2	− 0·2	− 0·2	17	− 300·1	0	− 1·8
2	− 300·1	0	0	18	− 300·0	0	− 2·3
3	− 299·8	+ 0·2	+ 0·2	19	− 299·8	+ 0·1	− 2·6
4	− 300·0	− 0·2	+ 0·4	20	− 299·7	+ 0·2	− 2·0
5	− 300·1	− 0·1	0	21	− 299·9	+ 0·1	− 1·7
6	− 300·0	0	− 0·1	22	− 299·9	+ 0·1	− 0·9
7	− 300·2	− 0·1	0	23	− 299·9	+ 0·1	− 0·5
8	− 299·8	+ 0·1	0	24	− 299·9	+ 0·1	− 0·1
9	− 300·1	− 0·1	0	25	− 299·9	+ 0·1	− 0·6
10	− 300·2	− 0·3	− 1·1	26	− 299·8	+ 0·1	− 0·6
11	− 300·0	0	− 0·7	27	− 299·9	+ 0·1	− 0·5
12	—	− 0·1	− 1·2	28	− 299·8	+ 0·2	− 0·6
13	− 300·1	0	− 1·1	29	− 300·0	+ 0·2	− 0·1
14	− 300·1	—	− 1·1	30	− 299·9	0	− 0·2
15	− 300·1	0	− 1·1	31	− 299·9	+ 0·2	− 0·3
16	− 300·0	− 0·2	− 1·8				

Nominal frequency corresponds to a value of 9 192 631 770·0 Hz for the caesium F,m (4,0)–F,m (3,0) transition at zero field.

U.D.C. 621.372

Low-pass Active Filters

By

H. BLACKBURN,
C.Eng., A.M.I.E.R.E.,†

D. S. CAMPBELL, B.Sc.†

AND

A. J. MUIR,
B.Sc., C.Eng., M.I.E.E.‡

Reprinted from the Proceedings of the Joint I.E.R.E.–I.E.E. Conference on 'Applications of Thin Films in Electronic Engineering' held in London on 11th–14th July 1966.

Summary: Microminiature active filters have been constructed using a combination of thin film and integrated circuits. The advantages of such an approach are discussed. The filter required could be met by a fifth-order Chebyshev function and the method of realization was to split the transfer function into quadratic and linear factors with each factor then realized separately. The tolerances required have been examined using a Monte Carlo computer program and the results are discussed. A filter constructed for use in p.c.m. systems is described.

1. Introduction

The trend towards continuous reduction in size of communications equipment and in particular the growing use of microelectronics requires electric filter networks using compatible techniques.

Of the components used in normal LC passive filters, inductors are the obvious choice for replacement when considering the applications of these techniques. If we consider therefore resistors and capacitors as the passive components to be used we find that the range of realizable voltage transfer functions is very limited. If, however, we allow the use of negative components, negative driving point impedances, or other such concepts associated with active devices, it can be shown that all the normal filter voltage transfer functions can be realized using resistors, capacitors and active devices. The two most commonly used methods of realizing these active RC filters use negative impedance convertors or feedback amplifiers as the active devices.

The first method is well described in the literature by Linvill[1] and Yanagisawa[2] with useful practical articles by Storey.[3,4] The second method using feedback amplifiers with prescribed gains has been described by Sallen and Key[5] and Piercey.[6]

It is the design method developed by Piercey[6] which was used to realize the filter to be described. This theory shows how we may partition the voltage transfer function V_{in}/V_{out} for a filter insertion loss approximation into the product of second-order numerator polynomials with the addition of a linear term if the overall order is odd. If the insertion loss has attenuation poles there will be a denominator polynomial which may always be partitioned into second-order terms. Basic sections are then used to generate quadratic or biquadratic factors and since, due to the method of synthesis, these have high input impedance and low output impedance, they may be cascaded to form the required final response.

Considering the realization of a single quadratic factor, $p^2+\sum p+\Delta^2$ by active and passive means, Fig. 1 shows the relative numbers of components required. The simplest form of transistor feedback

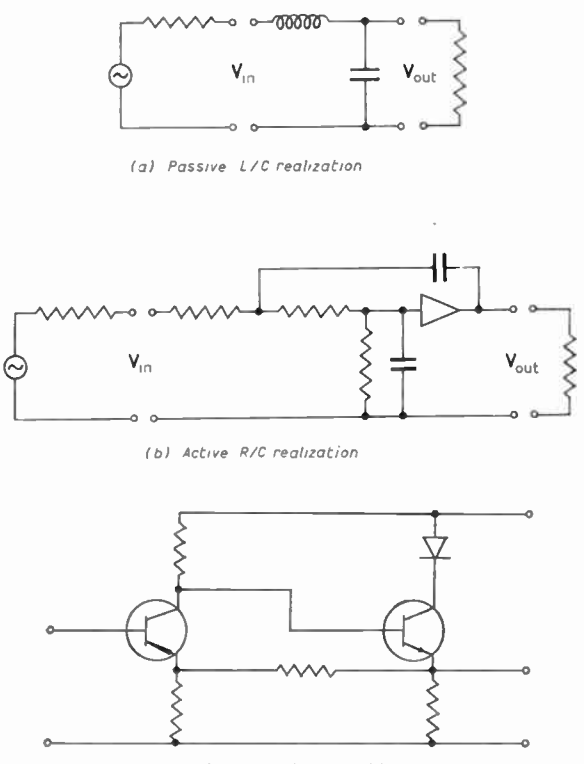


Fig. 1. Realization of quadratic factor

$$\frac{V_{in}}{V_{out}} = \frac{p^2+\Sigma p+\Delta^2}{k_0}.$$

† The Plessey Company Limited, Allen Clark Research Centre, Caswell, Towcester, Northants.

‡ British Telecommunications Research Limited, Taplow Court, Taplow, Berks.

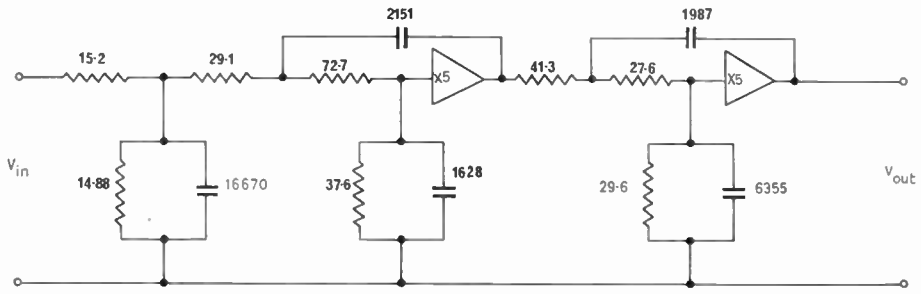


Fig. 2. 3·4 kHz low-pass filter (component values in kΩ and pF).

$$\frac{V_{in}}{V_{out}} = 5{\cdot}59(p+0{\cdot}36232)(p^2+0{\cdot}22393\,p+1{\cdot}03578)\,(p^2+0{\cdot}58625\,p+0{\cdot}47677)$$

amplifier is also shown. Thus we see that although the number of reactive elements is the same the the number of resistors has increased from two to seven with additional transistors and a diode. With a more complicated network of say three quadratic factors the additional components would now consist of nineteen resistors, six transistors and three diodes. It will be seen therefore that the use of capacitors as the sole reactive element has led to a considerable increase in the numbers of resistors and associated interconnections with the addition of active devices. The use of thin film components and silicon integrated linear amplifiers suggests itself as an advantageous method of construction and the realization of a low-pass filter for a multi-channel speech system using these techniques will be described.

The required insertion loss approximation was met by a fifth-order Chebyshev function, shown with the circuit and component values in Fig. 2.

2. Active Filter Analysis

The circuit consists of two basic sections of the type shown in Fig. 1(b) and a passive RC divider for the linear factor in the linear term.

An important consideration in the design of active filters is the sensitivity of response to gain and component tolerances. When a given polynomial has been factorized special consideration must be given in the design to the most sensitive or high Q-factors, i.e. those having roots closest to the vertical axis of the P plane. In the design being considered the third section, realizing the high Q-factor, was synthesized to have minimum sensitivity which set the gain of the amplifier. In the second section the amplifier was set, for convenience, at the same gain as the first amplifier and the section then designed for this gain.

Internal impedance levels must also be carefully scaled so that the range of component values is suitable for thin film construction.

To investigate the component tolerances required and to compute the response changes resulting, a computer program using the Monte Carlo method of random number tolerancing was prepared. This program analyses a basic circuit block treating the amplifier gain as a component for tolerancing purposes.

Tolerances may be applied as desired to each component and for a selected number of analyses or

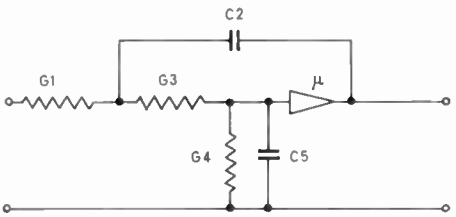


Fig. 3. Standard deviation of output when one component is allowed to vary by ± 1 %.

$$p^2+0{\cdot}59\,p+0{\cdot}48\ldots.\ \text{(i)}$$
$$p^2+0{\cdot}22\,p+1{\cdot}04\ldots.\text{(ii)}$$

Standard deviation dB.

	Varying components	$\omega = 0$	3·3 kHz $\omega = 0{\cdot}94$	3·5 kHz $\omega = 1$
(i)	G1	0·021	0·032	0·036
	C2	—	0·005	0·004
	G3	0·014	0·080	0·078
	G4	0·036	0·073	0·074
	C5	—	0·115	0·109
	μ	0·048	0·116	0·109
(ii)	G1	0·011	0·174	0·159
	C2	—	0·214	0·204
	G3	0·027	0·190	0·317
	G4	0·037	0·200	0·191
	C5	—	0·052	0·187
	μ	0·048	0·324	0·445

builds the standard deviation at predetermined frequencies is computed. The location of the most sensitive component may thus be noted and the filter response suitably placed on the frequency scale to meet the desired specification.

Figures 3 and 4 show some Monte Carlo results for the factors of the filter under consideration, the number of builds for each factor being 100. Figure 3 shows that the most sensitive component in the complete filter is the amplifier gain of the high-Q-factor section so that careful adjustment of this would be worthwhile. Figure 4 shows the response contours in the pass and stop bands for filters made with 1% and ½% component tolerances.

3. Preparation of Thin Film Circuits

The specifications required by Figs. 3 and 4 for component tolerances show that an important characteristic of the components should be a very low drift of value with time, and low values for the temperature coefficients of resistance and of capacitance (t.c.r. and t.c.c.). A combination of vacuum-evaporated nickel–chromium alloy resistors and silicon oxide capacitors is capable of meeting these requirements.

The films were deposited on cleaned Corning 7059 glass substrates by vacuum evaporation in a system pumped to 5×10^{-5} torr by an oil diffusion pump.

The means for defining component patterns were based on the use of photographic reduction of large-scale (×25) taped layouts of the circuit shape, which were then used to define areas for photoetching patterns in 0·0025 in thick molybdenum foil or directly to etch patterns in the thin film. The etching process used was to apply a thin layer of Kodak KPR photoresist to the foil or film, expose it to ultra-violet light through the photographic reduction plate, and then to develop and etch.

3.1. *Resistors and Connectors*

Previous work had shown[7] that the highest stable value of sheet resistance for a NiCr film was 300 ohms/square when deposited on a heated glass substrate. To allow a safety margin and to obtain repeatably low t.c.r. values, a value of 200 ohms/square was used. Other work[8] had shown that low t.c.r. values (< 20 parts in 10^6 per deg C) could be obtained using flash evaporation of nickel alloy powder which was slowly sprinkled on a very hot (1800°C) tungsten strip. As the composition of the deposited film is the same as that of the powder, this process gave much more repeatable results than were obtained than by other evaporation means. To obtain good adhesion and high stability the substrates were heated to 300°C. The resistance of the film was monitored during evaporation.

The need for the lands and connectors to have low resistance, easy solderability and resistance to oxidation, dictated the choice of gold for this purpose. This was deposited from a directly heated molybdenum boat on to the substrate which was already covered with the NiCr resistor layer, the NiCr being necessary to provide good adhesion. A substrate temperature of > 200°C was necessary for good adhesion but < 300°C in order to prevent diffusion of gold on to the nichrome

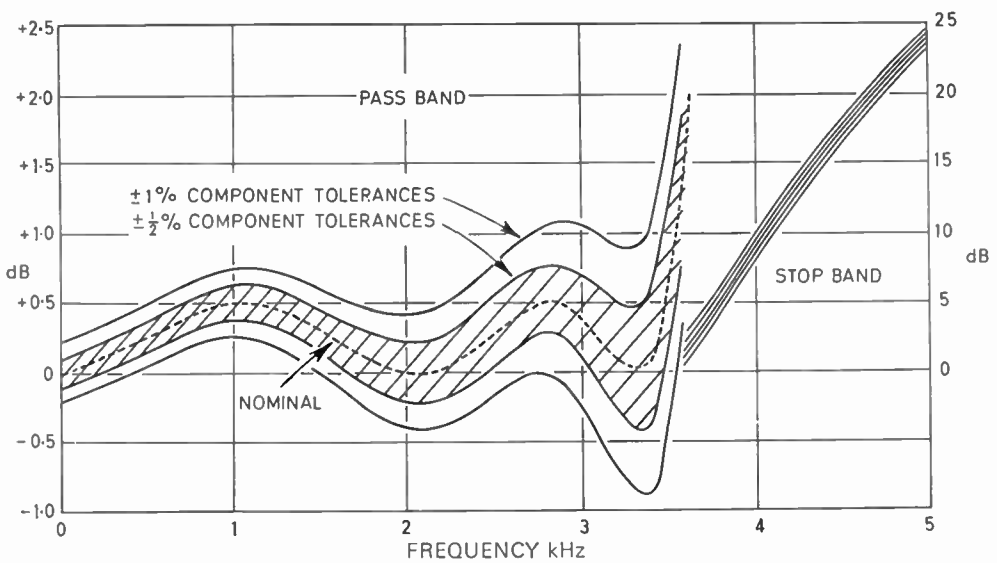


Fig. 4. Standard deviation of output with ±½% and ±1% tolerances.

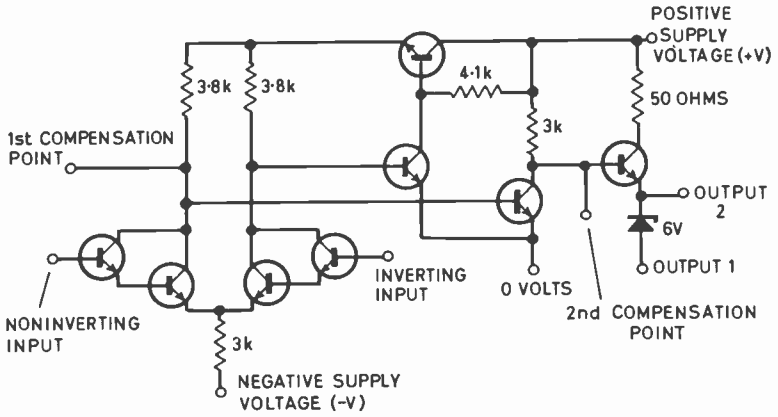


Fig. 5. SL751 amplifier.

during deposition. The gold was deposited to a resistance of 0·5 ohms/square, i.e. a thickness of 2000 Å.

3.2. *Capacitors*

The capacitors took the form of a layer of silicon oxide between electrodes of aluminium. The silicon oxide used was Kemet vacuum-baked silicon monoxide powder and it was evaporated from a directly heated tantalum boat. The film was deposited at a rate of 10 Å/s on substrates heated to 300°C. The thickness of film used was 4000 Å and this gave a capacitance of 10,000 pF per cm^2 of active area. The thickness was monitored during evaporation using an Edwards quartz crystal film thickness monitor in which the mass of film deposited altered the resonant frequency of the crystal. The aluminium electrodes were evaporated from tungsten spirals to a resistance of 0·5 ohms/square.

The component patterns were obtained by evaporating through molybdenum masks. All the depositions were carried out in one pump-down using an Edwards microcircuit mask changer jig.

4. Component Tolerances

The component tolerances as deposited were much too wide for the filter application, resistors being to $\pm 10\%$ and capacitors to $\pm 5\%$ even under carefully controlled laboratory conditions, and so means for adjustment were necessary in all components.

Resistors were adjusted by scribing into the side of the resistor with a diamond stylus. For convenience a shoulder was added to one side of the resistors during the original layout; by this means a cut could be made into it parallel to the long side of the resistor and adjustment to $\pm 0{\cdot}1\%$ was possible. Using this technique resistors can only be increased in value. It was not practicable to isolate areas of the capacitor in the same way so small capacitors were connected in parallel to the main one. These areas were arranged so that the connecting bars to the small capacitors did not pass over the bottom electrode. By scribing through the connectors the small capacitor can be disconnected and adjustment to about ± 2 pF accuracy was possible.

5. Component Stability

Component stability was dependent on two factors, short-term changes due to temperature coefficient effects and long-term drifts due to electrical changes with time. As previously described,t.c.r. values were reproducibly obtained using flash evaporation and values of $+50 \pm 25$ parts in 10^6 per deg C were obtained with a set NiCr ratio. Capacitors could also be obtained with low t.c.c. values; typical values being $+70 \pm 20$ parts in 10^6 per deg C. If a change of $0{\cdot}1\%$ was allowed for t.c.c. change, this limits the temperature range to ± 15°C about room temperature due to both resistor and capacitor value variations.

Life testing results on components are at present available on only small numbers of components. Results on resistors held at 20°C and 180°C had drifts of $< +0{\cdot}05\%$ and $+5\%$ after 5000 and 1000 hours respectively. Capacitors at 25°C show capacitance change of less than $\pm 0{\cdot}15\%$ in 3500 hours at voltage levels up to 5 volts and less than $-0{\cdot}25\%$ in 3000 hours at 125°C. It is likely that the long-term stability could be improved by prolonged annealing and running in of the components. Further work on this and on t.c.r. and t.c.c. is continuing.

6. Amplifiers

As previously discussed the gain of the amplifier has to be extremely well defined. The gain of an individual transistor is unpredictable and varies considerably with temperature, and it was thus

apparent that a suitable amplifier could only be obtained using an amplifier circuit with a very high gain and using close tolerance feedback resistors to reduce the gain to the desired low value. In this way variations in transistor gain would be almost eliminated. Whilst it would be possible to design a circuit using separate transistors, it would require five or six to meet all the requirements specified, and this would have been prohibitively expensive. A much more economical result was obtained by using silicon integrated circuits; these were ideally suited to providing high gain and to being used with thin film feedback resistors and stabilizing capacitors.

Several different hybrid amplifiers have been designed with great success. The one used for this application (Plessey Semic Ltd. No. SL751) is a general-purpose high-gain amplifier. It is not a complete functional unit but was designed for use in hybrid systems using thin films or conventional components. The most important section of the amplifier, whose equivalent piece-part circuit is shown in Fig. 5, is a precision comparator 'front end'; this uses 'Darlington' compound pairs in a long-tailed pair configuration. Following the comparator is a common emitter stage and an emitter follower output stage. A 6V Zener diode is used to shift the output level voltage.

An open-loop gain of ~70 dB is obtained and to obtain a gain of ×5 the circuit of Fig. 6 was designed. The absolute values of resistors were not critical, only the ratio of the two feedback resistors being important. The technology used in silicon integrated circuits can only give ratios of resistance to a tolerance in the region of ±5% and so it was necessary to use external thin film resistors for this purpose. The capacitors are required to prevent oscillation occuring.

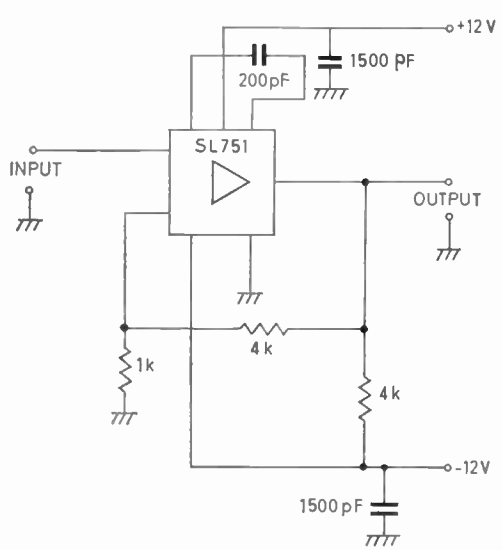


Fig. 6. SL751 amplifier plus feedback resistors and stabilizing capacitors.

The layout of solid circuits in flat-pack form was particularly well suited to incorporation into thin film circuits, although it will soon be practicable to use flip-chip circuits in the same way as is becoming common practice with transistors.[9]

The thin film resistors and four capacitors were made in the same way as before, the flat pack being attached by soldering. The amplifier gain was set by adjusting the ratio of feedback resistors to a ratio of 4:1 and then testing on an attenuator set. Further adjustments could then be made to obtain the correct gain.

The ratio of the two resistors was very stable indeed as close proximity almost entirely eliminates the effect of t.c.r. and of long term drift. The input impedance of the complete amplifier was >10 Mohm and the output impedance was less than 1 ohm, and so it can be regarded simply as a voltage gain block.

7. Complete Filter

For convenience the filter was made up with each factor occupying a 1 inch square substrate, not including the amplifiers. In this way it was possible to check that each factor was correct before assembling the complete filter.

The complete filter was laid out on a sheet of printed circuit board which served to interconnect the circuits and also to give mechanical strength to the complete filter. Figure 7 shows the substrates and Fig. 8 the completed filter. The filter is encapsulated in silicone rubber with a hard outer layer of epoxy resin.

Figures 9 and 10 show the response curves obtained from the filter; it is apparent that the pass-band ripple is rather large despite the fact that all components were well within ±½% tolerance. The reason is that the linear function is loading the quadratic factor following it to a small extent. To reduce the pass-band ripple to the necessary degree it was necessary to adjust the amplifier gains slightly. This was conveniently done by having tapping points available on the feedback resistors. As can be seen, the filter can tolerate large variations in temperature and supply line voltage, and the effect of varying the positive supply line would have been negligible.

8. Conclusions

It has been shown that active filters for use in relatively critical applications can be realized in microelectronic form using a combination of thin film and silicon integrated circuits and that the stability of the component values is sufficiently good for the filter specification to be satisfied over a reasonable temperature range and for long periods of time.

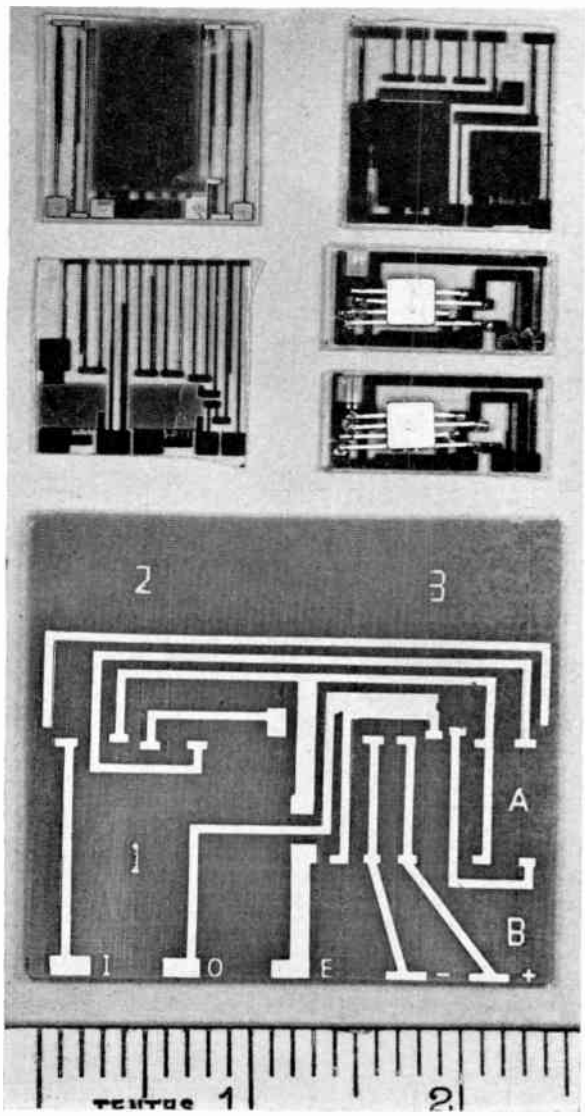


Fig. 7. Substrates for the filter.

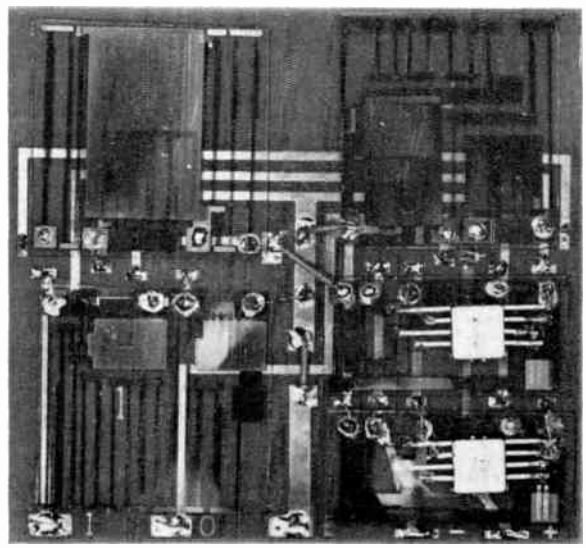


Fig. 8. The completed filter before encapsulation.

It should be emphasized that there are several possible approaches to active filter design and also that less critical filter requirements could be satisfied by much simpler circuits with wider tolerance components.

9. Acknowledgments

The authors wish to thank R. N. G. Piercey, R. Simkins and R. C. Foss for major contributions to this work and the Plessey Company for permission to publish this paper.

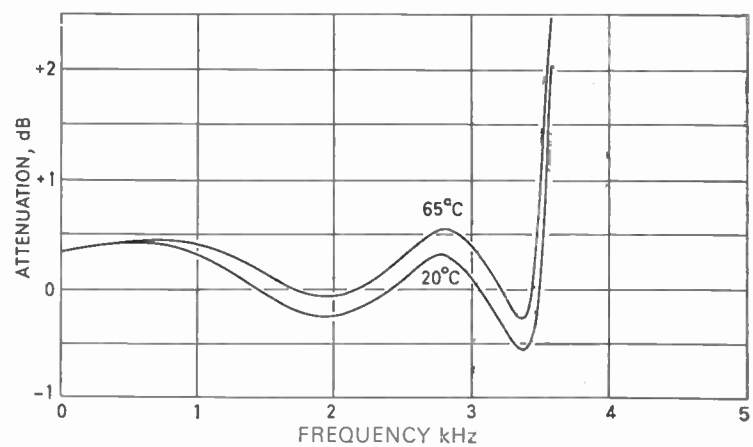


Fig. 9. Effect of temperature on thin film filter characteristic.

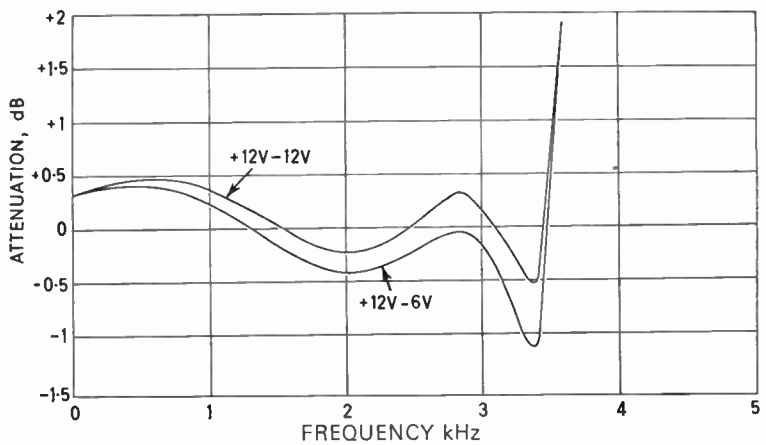


Fig. 10. Effect of supply line voltage on thin film filter characteristic.

10. References

1. J. G. Linvill, 'RC active filters', *Proc. Inst. Radio Engrs*, **42**, pp. 555–64, March 1954.
2. T. Yanagisawa, 'RC active networks using current inversion type negative impedance converters', *Trans. I.R.E. on Circuit Theory*, **CT 4**, No. 3, pp. 140–4, 1957.
3. D. J. Storey and W. J. Cullver, 'Network synthesis using negative impedance converters', *Proc. Instn Elect. Engrs*, **111**, No. 5, pp. 891–906, May 1964.
4. D. J. Storey and W. J. Cullver, 'Active low-pass linear phase filters for pulse transmission', *Proc. I.E.E.*, **112**, No. 4, pp. 661–8, April 1965.
5. R. P. Sallen and E. L. Key, 'A practical method of designing RC active filters', *Trans. I.R.E. on Circuit Theory*, **CT–2**, pp. 74–85, March 1955.
6. R. N. G. Piercey, 'Synthesis of active RC filter networks', *A.T.E. J.*, **21**, No. 2, pp. 61–75, April 1965.
7. R. H. Alderson and F. Ashworth, 'Vacuum deposited films of nickel chromium alloy', *Brit. J. Appl. Phys.*, **8**, pp. 205–10, June 1957
8. B. Hendry and D. S. Campbell, 'The effect of composition on the temperature coefficient of resistance of NiCr', *Brit. J. Appl. Phys.*, **16**, pp. 1719–25, November 1965.
9. E. O. Carr, 'Flip chips, easier to connect', *Electronics*, **36**, No. 42, pp. 82–4, 18th October 1963.

Manuscript received by the Institution on 28th April 1966. (Paper No. 1135.)

U.D.C.621.382

Galvanomagnetic Thin Film Devices

By

H. FRELLER, Ing.†

AND

K. G. GÜNTHER, Dr.rer.nat.†

Reprinted from the Proceedings of the Joint I.E.R.E.-I.E.E. Conference on 'Applications of Thin Films in Electronic Engineering', held in London on 11th–14th July 1966.

Summary: Thin films of InAs and InSb directly evaporated on to substrates, such as glass, ceramics or ferrite, have interesting properties which make them suitable for galvanomagnetic devices as Hall probes or magneto resistors. The reasons for this are as follows: the Hall-sensitivity increases with decreasing thickness of the semiconducting layer. The higher inner resistance of thin films facilitates matching of the device to a particular circuit. Directly evaporated layers have a better heat dissipation factor than thin slices cemented to a substrate.

Suitable methods of producing stoichiometric thin films of InAs and InSb with mobilities comparable to bulk material values are the flash-evaporation techniques and the three-temperature method. As the mobility shows a strong dependance on the crystallite size of the polycrystalline films, recrystallization techniques were subsequently applied to InSb films. This treatment results in larger crystallite sizes and mobilities of up to 30,000–40,000 cm²/Vs.

The properties of thin film devices are discussed with special regard to Hall probes. The average values of mobility and Hall coefficient versus temperature and the resulting characteristics such as Hall-sensitivity, inner resistance, control current and temperature coefficient, are discussed with reference to several typical devices.

List of Symbols

a	length of semiconducting film
α	temperature coefficient of resistance
B	magnetic induction
b	width of semiconducting film
β	temperature coefficient of Hall voltage
C_1, C_2	geometrical factors (varying with magnetic induction
d	thickness of semiconducting film
δ	thickness of a monolayer
$g_{1,2}$	geometrical factors varying with a and b
I_1	control current
K_0	Hall sensitivity
μ	carrier mobility
μ_H	Hall mobility
P_1	input power
P_2	output power
R_B	film resistance at induction B
R_H	Hall coefficient
R_0	film resistance at zero induction
ρ_B	resistivity at induction B
ρ_0	resistivity at zero induction
s	scattering factor
θ	Hall angle, with $\tan\theta = \mu B$
U_H	Hall voltage

1. Introduction

Since the discovery of semiconductors of high carrier mobility in particular indium antimonide and indium arsenide with $\mu = 70\,000$ and $30\,000$ cm²/Vs, respectively, a great deal of interest has been focused on the so-called galvanomagnetic devices.[1] These are electronic components whose output voltages and currents can be decisively influenced by external magnetic fields. The most important representatives of this group are the Hall generator[2] and the magneto-resistor, whose internal resistance varies appreciably in the presence of magnetic fields.[3] By using ground semiconductor platelets, film thicknesses of about 20 μm are obtained in the manufacture of such devices. However, it is desirable to work with films as thin as possible since, on the one hand, the voltage produced at the Hall generator is inversely proportional to the thickness of the film, and, on the other hand, the internal resistance of the order of several kilohms often desired for matching reasons can only be realized with film thicknesses of a few microns for both of the above-mentioned devices. Finally, eddy-

† Siemens A.G., Nuernberg, Germany.

current losses and the upper frequency limit depend on the film thickness and also call for semiconducting platelets which are as thin as possible.

Thin film techniques therefore suggest themselves for the manufacture of galvanomagnetic devices. This also makes it possible to deposit the semiconducting films directly on a relatively good heat-conducting substrate and thus obtain a good thermal contact between the semiconducting layer and the substrate. This also increases the load capacity of the device.

In the course of the last few years various attempts have been made to manufacture galvanomagnetic thin-film devices. The difficulties encountered are as follows:

Semiconducting properties of the required quality are only possible with a composition of the deposited compound films corresponding exactly to the stoichiometric composition. A high carrier mobility is also only possible with an ideal crystal structure with a minimum of lattice dislocations and grain boundaries. Consequently, only a few of the methods tried were successful, namely, thin film etching, flash evaporation[4, 5] and the so-called three-temperature method.[6] Annealing and recrystallization methods were also tried for subsequent improvement of the semiconducting properties.[7-9]

The present paper deals firstly with the principle of the techniques referred to above, followed by a discussion on the film properties obtained. Finally, discussion of the characteristics attained by the thin-film devices together with the range of scatter found in their properties is also included.

2. Techniques Employed in the Manufacture of Indium Antimonide and Indium Arsenide Thin Films

The thin-film technique offers a large number of possible preparation methods of which, however, only a few can be employed for the task in question.

Functional Hall generators and magnetoresistors are only possible if the Hall coefficient R_H and Hall mobility μ_H are at least comparable with the corresponding values of the bulk material (see Table 2). This results in the following prerequisites for the desired evaporated thin films:

(*a*) exact stoichiometry,

(*b*) avoidance of ionized impurities and

(*c*) avoidance of scattering centres (lattice dislocations, grain boundaries).

Since the evaporated thin films under discussion consist of compound semiconductors, condition (*a*), in particular, presents a number of problems.

Owing to the great difference in the vapour pressure of the individual constituents, vaporization of the compound in the usual manner without decomposition is not possible (Fig. 1). In order to attain exact stoichiometry, however, simultaneous vaporization of the constituents would call for an accuracy of control of the vapour source temperatures which is not possible in practice.

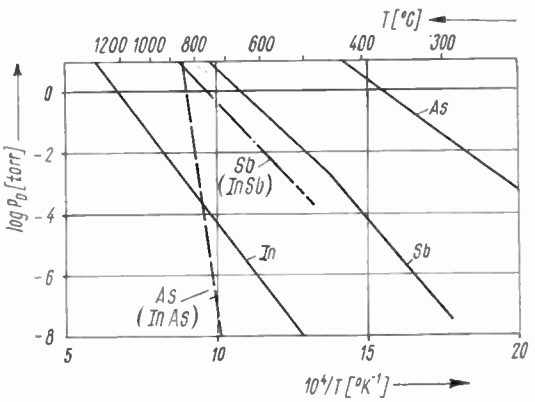


Fig. 1. Vapour pressure of the elements In, Sb, As and their compounds.

Of the many methods tried for evaporating InSb and InAs films, therefore, only two, namely, the flash evaporation method and the three-temperature method have had any practical success. (See Fig. 2.)

In the case of flash evaporation small particles of the compound InSb or InAs are dropped by a feeding assembly on to a vapour-source crucible whose temperature is high above those required for evaporating the constituents (nearly 1650°C and 1500°C, respectively, for the two compounds mentioned[10]).

The result is that both constituents evaporate flash-like and almost simultaneously. It is important in this connection to note that the grain size, V, of the dropped particles does not exceed a value prescribed by the distance D between the vapour-source and the substrate. The following equation shows how V and D are related:[10]

$$V \leq \pi \times D^2 \times \delta$$

This is the only way of ensuring that the condensing multilayers of the constituents are sufficiently thin and that they diffuse into each other and thus form the desired compound either during condensation or in subsequent heat annealing.

With the three-temperature method,[6] stoichiometric composition of the constituents in the vapour phase is deliberately foregone. The two crucibles for evaporating the constituents are heated to such a temperature that the more volatile constituent (Sb or As) in the vapour phase is in excess. The selection of the necessary vapour constituents is made during the condensation process.

For this purpose the substrate temperature is raised until condensation of the unsaturated, more volatile components can be excluded. Consequently, only the desired compound InSb or InAs condenses, while the excess of the more volatile component is re-evaporated into the vapour space.

This method can always be applied when the vapour pressure of the more volatile component above the compound is appreciably lower than that of the pure element. In this case there is always a limited region for the substrate temperature within which the selection outlined above of the impinging vapour particles is guaranteed. Table 1 applies, for example, for normal evaporation velocities.

Table 1

Crucible and substrate temperatures (T_1, T_2, T_3) in the three-temperature method

Compound	T_1 (In)	T_2 (As, Sb)	T_3 (substrate)
InSb	960°C	580°C	400–530°C
InAs	960°C	280°C	200–700°C

With both methods, the desired compound thin films with any desired thickness between $d = 0{\cdot}1$–10 μm can be reproducibly obtained on glass or mica substrates. The structure is naturally polycrystalline with mean crystallite sizes of up to a maximum of 20 μm.

Without the addition of special doping impurities, the films have n-type conductivity and reveal an electron mobility which, above all, is strongly influenced by the mean crystallite size.[11, 12] (See Fig. 3). The scatter range shown in Fig. 3 is presumably

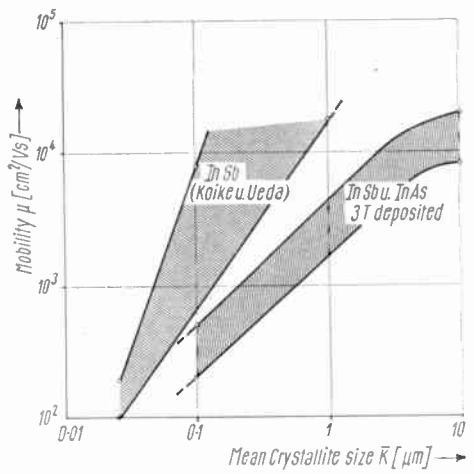
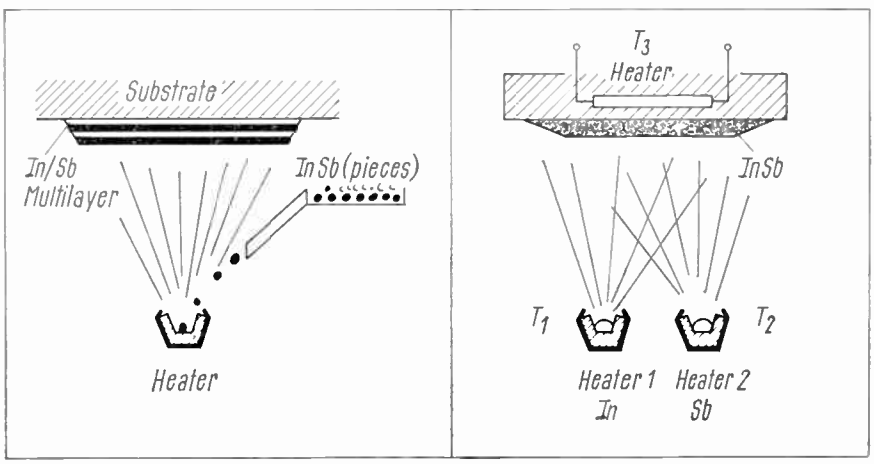


Fig. 3. Graph of Hall mobility against crystallite size for vacuum deposited III-V films.

caused by differing numbers of dislocations and other lattice imperfections in the crystallites. The upper curve of the mobility region therefore represents films with optimum crystallite quality.

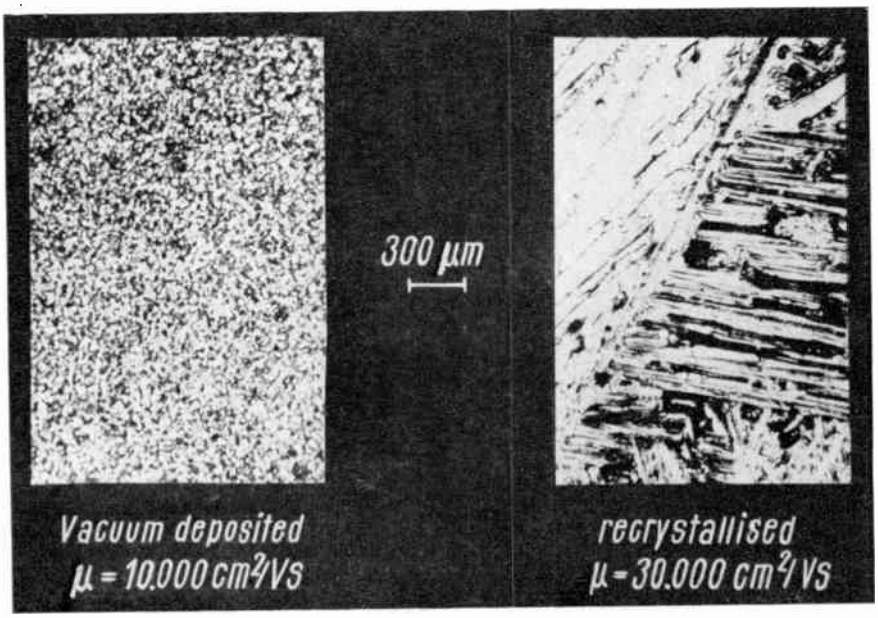
In order to approach the mobility values of the bulk material as closely as possible, it is necessary to produce evaporated thin films whose crystallites are as large as possible.

If this does not already take place during the condensation process, a further crystallite growth can be obtained with InSb by heat treatment. By heating InSb films over the melting point of T_s = 525°C and by applying a suitable thermal gradient on cooling, it is possible to obtain layers with dendritic



(a) Flash evaporation method. (b) Three-temperature method.

Fig. 2. Vacuum deposition of compound films consisting of two components.



(a) Vacuum deposited. (b) Recrystallized films.

Fig. 4. Micro-photographs of InSb films.

structure and needle crystals whose length reach several millimetres (see Fig. 4). It is necessary here, however, to provide the evaporated thin film with a protecting film or coating in order to prevent coalescence of the films, i.e. the formation of droplets on the surface. Foreign substances, such as MgF_2,[9] and also a subsequently evaporated In film, have been used for this purpose.[7] If this coating of indium is partly oxidized, its protective effect is improved. Once recrystallization has taken place, the desired homogeneity of the evaporated film can be re-established by selective etching. The same result is also obtained by surface oxidation of the InSb films normally produced according to the above-mentioned methods. In this case, subsequent etching is no longer required.

3. Properties of Evaporated InSb and InAs Films

For an appraisal of the evaporated films obtainable by the methods described above it will be useful, to begin with, to consider a few basic relations between the properties of the films and the characteristics of galvanomagnetic devices.

According to the equation:

$$U_H = \frac{R_H}{d} . J_1 . B . g_1\left(\frac{a}{b}, B\right) \qquad \text{.....(1)}$$

the open-circuit Hall-voltage U_H of a Hall-generator is first of all directly proportional to the Hall coefficient and inversely proportional to the film thickness d. For large side ratios the geometry function g_1 (a/b) approaches 1 and, for $a = 2b$, already reaches the value of 0·93.

Owing to the limited input power P_1, the control current I_1 cannot exceed a maximum value I_{1max} which, according to eqn. (2)

$$I_{1max} = \left[P_1 . \frac{\mu}{R_H} . \frac{b.d}{a}\right]^{\frac{1}{2}} \text{ for } \mu B \ll 1 \qquad \text{.....(2)}$$

depends both on the Hall coefficient and on the mobility μ of the charge carriers.

The maximum obtainable Hall voltage

$$U_{Hmax} = B . \left[\frac{R_H \mu}{d}\right]^{\frac{1}{2}} . \left[P_1 . \frac{b}{a}\right]^{\frac{1}{2}} \text{(for } \mu B \ll 1) \text{.....(3)}$$

is consequently prescribed by the three variables R_H, μ and d which are of interest for our purposes. The losses that can be dissipated $P_1 = ab.f(\lambda)$ depend decisively on the thermal conductivity of the zones bordering the semiconducting film and reach their maximum on evaporated films owing to the good heat dissipation to the substrate.

Another interesting factor in addition to the Hall voltage for most measuring and control tasks is the available output power, P_2. From eqns. (2) and (3) and, allowing for the internal resistance, P_2 has a maximum value of

$$P_{2max} = g_2 . P_1 . (\mu B)^2 \text{ for } \mu B \ll 1. \qquad \text{.....(4)}$$

Thus, P_2 depends to a particularly high degree on the mobility μ. (g_2 is mainly determined by the geometry

of the Hall contacts and, for normal designs, has a value of about 0·25).

The dependence of the internal resistance on a transverse magnetic field $R_B/R_0 = f(B)$ is explained on the one hand by an increase in the physical resistivity (as the result of a decrease in mobility in the magnetic field), and on the other hand, by the increase in the effective current paths (geometry effect).

For rectangular semiconductor layers, for instance, the following relationship is obtained.

$$\frac{R_B}{R_0} = \frac{\rho_B}{\rho_0}.(C_1 + C_2).(\mu B)^n \qquad \text{.....(5)}$$

in which C_1, and in turn C_2, are slightly dependent on the value of μB (e.g. $n = 2$ for $\mu B \ll 1$; $n = 1$ for $\mu B \gg 1$).

For small magnetic fields the physical resistivity effect also varies as the square of the mobility

$$\frac{\rho_B}{\rho_0} = 1 + s(\mu B)^2 \qquad \text{.....(6)}$$

Equations (1) to (6) show that, for the construction of thin film Hall generators, both quantities R_H and μ for a minimum possible film thickness d must exceed certain minimum rates, while in the case of magnetoresistors the most decisive factor is the mobility μ owing to the square dependence and may not be appreciably below the values of the bulk material. In this respect the film thickness d is interesting only for matching reasons.

Figure 5 shows the temperature dependence of the Hall coefficient obtained with various evaporated films and production techniques. A comparison of these curves with the figures for ambient temperature

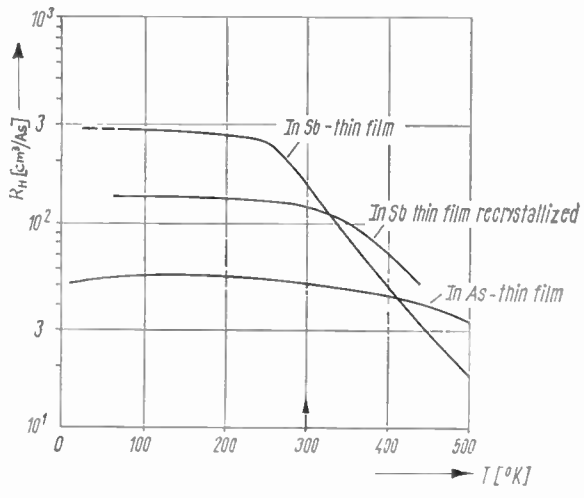


Fig. 5. Graph of Hall coefficient against temperature for vacuum deposited III-V films.

given in Table 2 shows that the Hall coefficients of the evaporated film are fully comparable with those of the bulk material. On the average the mobilities reached about 30% of the values of the bulk material.

Table 2

Hall coefficient and mobility of evaporated films compared with the bulk material (average values)

	InSb		InAs	
	μ [cm²/Vs]	R_H [cm³/C]	μ [cm²/Vs]	R_H [cm³/C]
Evaporated film	10 000	450	8 000	80
Evaporated film, recrystallized	25 000	80	—	—
Bulk material	60 000	500	24 000	120

The sudden change in the curves $R_H = f(T)$ at 270 and 400°K shows the transition to intrinsic conduction, the position at which this transition takes place being determined by carrier concentration and band gap. In the case of evaporated films the band gap coincides with that of the bulk material (0·17 eV for InSb; 0·36 eV for InAs).

The decrease in the Hall coefficient on recrystallization of the evaporated films might be explained by the segregation of p-doping impurities which have previously given rise to compensation. Depending on the substrate material chosen, the diffusion of impurities from the substrate is also naturally possible (e.g. alkali ions from glass substrates). Just as with the bulk material, the temperature coefficient of the Hall constant is minimized with increasing impurity concentration of the films and the change from impurity band conduction to intrinsic conduction is shifted in the direction of high temperatures.

The temperature dependence of mobility (Fig. 6) gives a maximum value at about 100°C, i.e. above the normal operating temperature, for films produced by the three-temperature method. The result is that, with such evaporated films, the temperature coefficient of resistance in the region of impurity band conduction is already negative and greater than the temperature coefficient of R_H. The effect of this behaviour on the temperature drift of the Hall-voltage will be discussed later.

Subsequent recrystallization results in a general increase in the mobility and the maximum value is displaced in the direction of low temperatures. So far, values of $\mu = 20\,000$ to $35\,000$ cm²/Vs have been attained; however, this is often only possible at the cost of a reduction in the Hall coefficient.

Measurements taken on the bulk material[13] show that the mobility drops severely with increasing impurity content. For example, if $R_H \simeq 500$ cm³/As

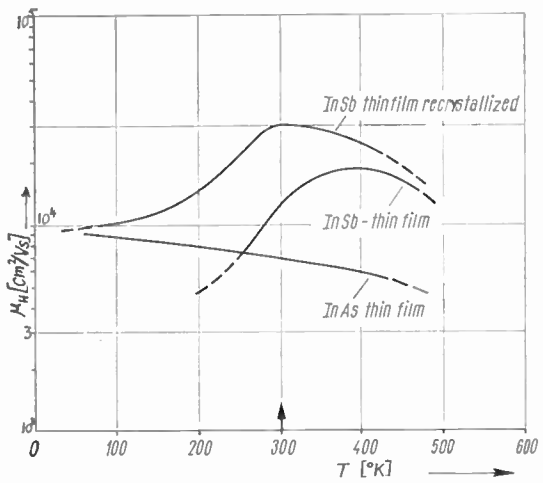


Fig. 6. Hall mobility against temperature for vacuum deposited III-V films.

($n = 1{\cdot}5 \times 10^{16}$), the mobility is approximately 70 000 cm^2/Vs; but if R_H is reduced to 70 cm^3/As, which corresponds to $n = 10^{17}\,cm^{-3}$, then the mobility falls to 35 000 cm^2/Vs. It must be concluded from this that the mobility of the recrystallized films at measured Hall coefficient $R_H = 70$ to 100 cm^3/As is mainly limited by impurity scattering and hardly any more by the crystallite sizes. A further increase in the mobility is therefore only possible with a greater purity of the evaporated films, i.e. smaller impurity and scattering centre concentrations.

With evaporated films obtained by the flash evaporation method, smaller values of carrier mobility are generally obtained[14] ($\mu \simeq 10^3$ to 5×10^3 cm^2/Vs). The increase in μ with temperature is similar to that shown in Fig. 5. The Hall coefficients of the films produced by this technique are also much smaller ($R_H = 80$ cm^3/As) without recrystallization.[15]

With a given scattering mechanism, the mobility also determines the change in the specific film resistance as a function of the magnetic induction B (eqn. (6)). In order to eliminate the geometry influence, the function $\rho_B/\rho_0 = f(B)$ is normally measured on elongated specimens with large values of a/b.

However, as Fig. 7 shows, the dependence of the film resistance on the magnetic field is, unlike in the case of bulk specimens, only influenced very slightly by the external dimensions a and b of the film. This behaviour is due to the polycrystalline structure whose grain boundaries and constrictions determine the effective geometry of the active elements and are not influenced by external dimensioning and bonding as long as a and b are appreciably larger than the mean crystallite size K.

Pre-requisites for the use of the evaporated film for devices are stability of the properties under long-time operation conditions and stability in operation at increased ambient temperature and under thermal shock conditions. Thermal shocks in the complete interval between 4°K and 500°K and tests extended over years at ambient temperature and over several months at 500°K revealed no irreversible changes in the evaporated films. Their applicability within the

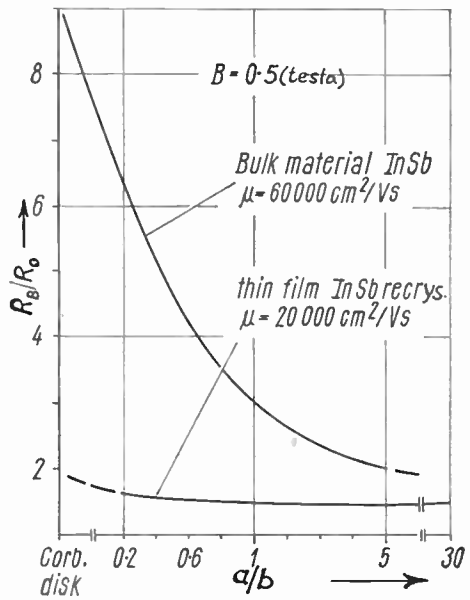


Fig. 7. Variation of magnetoresistance with probe geometry a/b for bulk material and evaporated films of InSb.

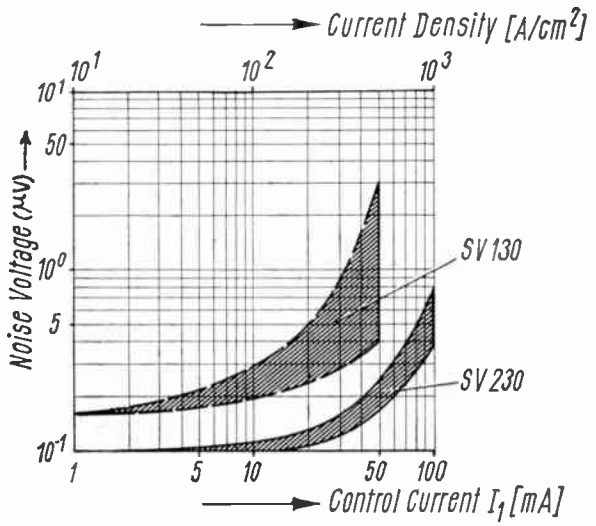


Fig. 8. Variation of noise-voltage of Hall probes with control current.

above-mentioned temperature interval would thus appear ensured.[16]

A certain amount of noise-voltage from the films due to their polycrystalline structure and the large surface-to-volume ratio is to be expected. Measurements show that the noise voltage is an inverse function of the frequency.[14] If the current density drops below values of about 10^2 A/cm^2, the noise signal is reduced to a limiting value independent of current which is in the range of the thermal noise level. The values given in Fig. 8 were obtained with a wide-band amplifier as summation measurements over a wide frequency range of 30 Hz–30 kHz.

4. Configuration and Characteristics of Complete Devices

For evaporated thin films of the type mentioned glass sheets, ceramic platelets and, in particular, ferrite slabs have proved successful. Ferrite substrates in connection with ferromagnetic concentrators are particularly interesting if it is intended to operate measuring or regulating devices at low magnetic inductions or weakly excited magnetic circuits. Since the films are deposited directly on good heat-conducting substrates, almost the entire surface of the substrate can be utilized for dissipation of the losses P_1. With a permissible temperature rise of $\Delta T = 10°C$ and operation of the semiconducting film in static air, the permissible specific input power is approximately given by

$$P_1/F = 200 \text{ mW/cm}^2 \qquad \ldots\ldots(7)$$

in which F is the area of the substrate in the case of ceramic and ferrite substrates.

The desired geometry of the evaporated film can be obtained with evaporation masks and by subsequent etching of a substrate whose whole surface is covered with the layer of material (see Fig. 9). Depending on the type of probe, the contacts or electrodes can also be produced with evaporation masks or deposited galvanically or purely chemically on the active semiconductor area. The ends of these contacts are widened out somewhat in order to be able to solder the connecting leads or foils—generally by the simultaneous method—or to attach them by a special welding process. Mechanical protection of the semiconductor and the connecting leads is provided by a resin coating whose consistency depends on the prospective operating temperature.

According to this technique, not only individual devices but—by analogy with the normal thin-film technique—also complete circuits can be deposited as required, for instance, for linearizing the characteristics of a Hall generator (as shown in Fig. 9).

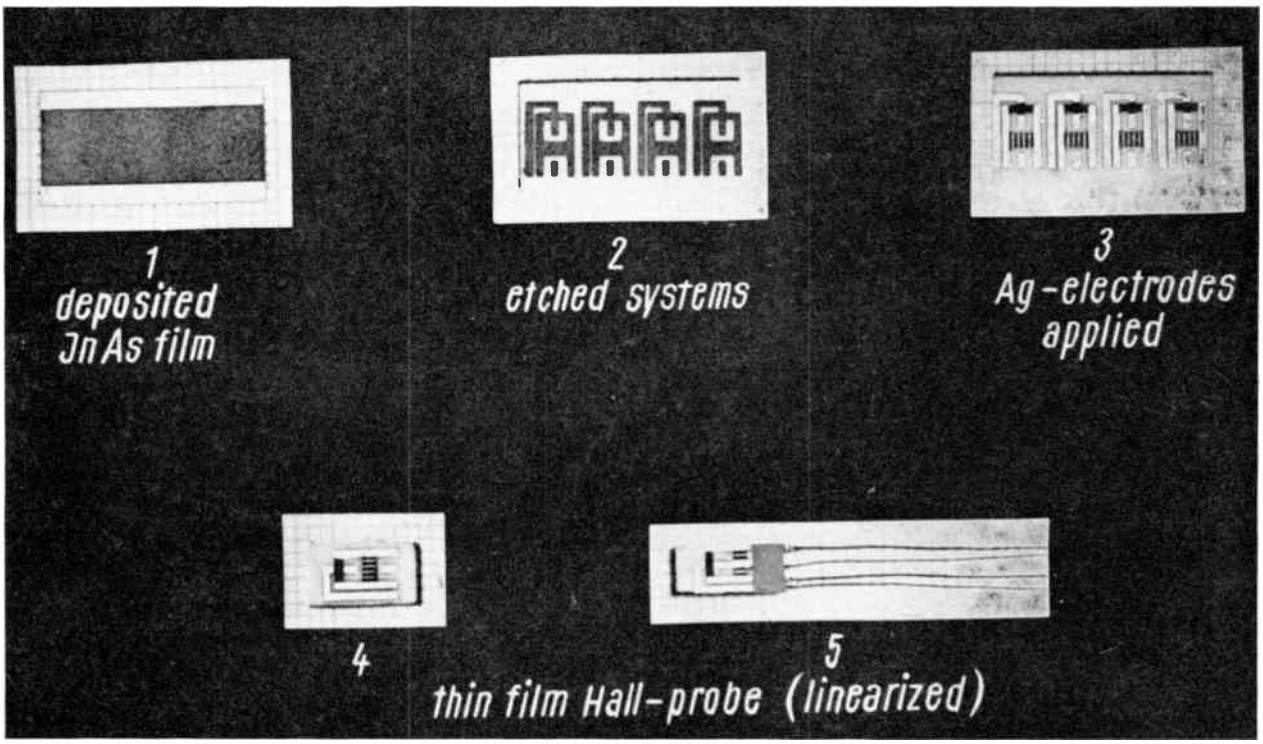


Fig. 9. Example for the preparation of several Hall probes from one vacuum-deposited layer.

So far, magnetoresistors, whose resistance characteristics are shown in Fig. 10, have been produced with evaporated raster electrodes or, in the case of recrystallized InSb films, indium inclusions will act as raster electrodes and therefore are not etched out after the recrystallization process.

Films with indium inclusions have so far revealed the most favourable characteristics since the effective raster electrodes located at distances from each other which are comparable with the crystallite width. The

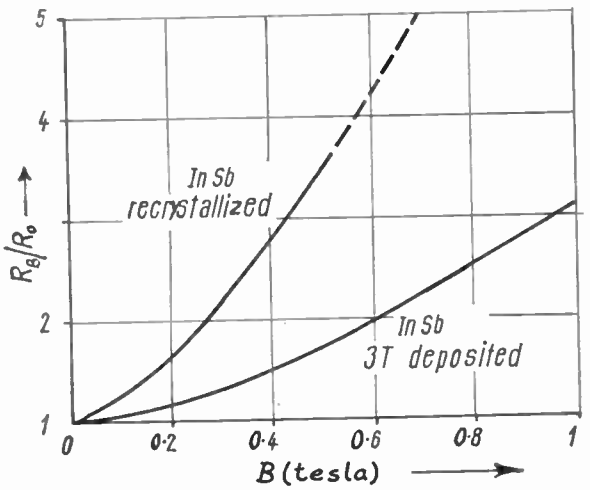


Fig. 10. Magnetoresistance of vacuum-deposited films with suitable electrodes.

(Raster electrodes or segregated In-enclosures respectively [after Wieder[10]].)

subsequent evaporation of raster electrodes with similarly small distances between them would be exceptionally difficult. Mobility values are still limited however, and the attainable change in resistance of evaporated films is considerably below that of bulk InSb specimens.

On the basis of the relationships implied in eqns. (3) and (4) only the Hall or output power varies as the square of the mobility, whereas the actual signal quantity, namely the Hall-voltage, varies as the root of the mobility. For this reason, the Hall effect can already be practically utilized at mobility values over $\mu = 5000\,\mathrm{cm^2/Vs}$ with the above-mentioned Hall coefficients. III-V evaporated layers in the form of thin-film Hall generators have therefore gained more significance, the authors restrict themselves in the following to a discussion of the attainable characteristics of evaporated Hall generators.

Experiments with various a, b dimensions have shown that the attainable Hall voltage as a function of the geometry corresponds exactly to the function g_1 given in eqn. (1), even in the case of thin-film generators. Layers in which $a/b = 1$ to 2 are therefore normally etched out of the original film, preference often being given to a cross shape.

The attainable Hall-voltage as a function of the control current is shown in Fig. 11, referred to a magnetic induction of $B = 10$ kilogauss (1 tesla) and on two typical designs by way of example. The continuous curves give the mean trace for both compounds InSb and InAs, and the hatched area shows the scatter range adhered to in mass production. The subsequent bend in the initially linear curve is explained by a temperature rise of the film caused by the Joule effect. The nominal value of the control current is such that deviations of 10% and 1% from the linear trace are not exceeded in the case of InSb and InAs, respectively. The hyperbolic curve 1 is a measure for the input power corresponding to the nominal control current ($P_1 = 300\,\mathrm{mW}$ with the types shown here with $F = 1{\cdot}6\,\mathrm{cm^2}$).

As a result of the increase already mentioned in the mobility with rising temperature and the good thermal contact between the evaporated film and the substrate, the control current can be increased far beyond its nominal value without destroying the films. A control current of double the nominal value is therefore permissible without endangering the device; this corresponds to a maximum permissible control power of about 1·2 W in the case of the types discussed here (see curve 2 in Fig. 11). The scatter range is the result of the unavoidable fluctuations in film thickness and imperfect concentration when evaporating large batches, but can be restricted to $\pm 30\%$ of the mean value by observing utmost cleanliness and taking due

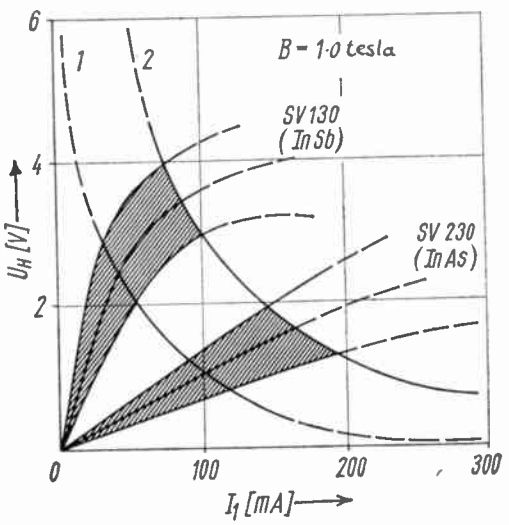


Fig. 11. Graph of Hall voltage against control-current at $B = 1$ tesla of vacuum deposited Hall-probes with ceramic substrates.

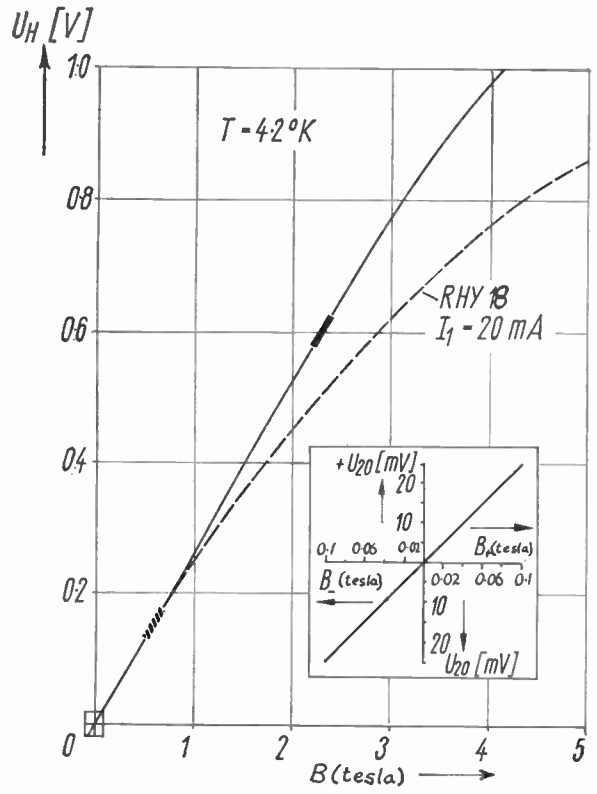


Fig. 12. Hall-voltage against magnetic induction of different vacuum deposited Hall probes.

wwwww upper limit of linearity.
– – – – usual Hall-probe $a/b = 2$.
——— Hall-probe with magnetoresistor in parallel (see Fig. 9).

care during the evaporation process and by adequate control of the critical temperatures of evaporator and substrate.

Figure 12 shows the Hall-voltage U_H as a function of the magnetic induction B for an arbitrary value of control current of $I_1 = 20$ mA, taking the low-temperature probe RHY 18 with InAs evaporated film as an example. The measuring temperature was 4·2°K so that it was possible to work with a superconducting coil up to inductions of 50 kilogauss (5 tesla). In its standard design, this probe has a linear range up to inductions of $B = 6$ to 7 kilogauss (0·6–0·7T). Above this value the characteristic $U_H = f(B)$ falls off monotonically (dashed curve). The enlarged section for small field strengths shows the exact antisymmetry on reversal of the field.

If this probe is connected in parallel with an InAs film with raster electrodes, the range of linearity can be extended to inductions of $B = 20$ to 25 kilogauss (2–2·5T). As already shown in Fig. 9 this combination is possible without having to take extra measures, i.e. only the geometry and the arrangement of contacts have to be changed.

The open-circuit sensitivity referred to the linearity range is

$$K_0 = U_H/B.I = R_H.g_1\left(\frac{a}{b};B\right)/d \qquad \text{.....(8)}$$

and reaches the mean values entered in Table 3 for the evaporated films produced according to the three-temperature method. Maximum values of $K_0 =$ 30 or 2 V/A kilogauss (300V/AT or 20 V/AT) for InSb or InAs, respectively, can be obtained for particularly thin films and by observing the utmost cleanliness and care in evaporating. The temperature coefficient β of Hall coefficient R_H and sensitivity K_0 also entered in Table 3 coincides with the corresponding values of the bulk material for both compounds.

Table 3

Mean parameters of evaporated Hall generators

Compound	K_0 [V/AT]	R_1 [Ω]	I_1 [mA]	α [% per degC]	β [% per degC]
InSb	75	200	10–50	−1·5	−1
InAs	10	30	60–100	+0·1	−0·1

The values given for K_0 are assigned to film thickness of $d \simeq 3$ μm. These are mean values since, depending on the nature of the substrate material employed (ceramics, ferrite), the peak-to-valley height of substrate and film may be of the same order of magnitude as the film thickness itself. For this reason in the normal case it is not possible to obtain thinner thickness of film.

Whereas the sensitivity is prescribed by the Hall coefficient and film thickness alone, the film resistance R_1 is also influenced by the mobility attained. With evaporated layers of the type obtained by the three-temperature method, film resistances of a few hundred to a maximum of 1 kilohm for InSb films, and of approximately 20 to 50 ohms for InAs films are obtained with the above-mentioned mobilities. These resistances are in a range which is often desired for matching reasons.

The temperature coefficient α of the film resistance at $-1{\cdot}5\%$ per degC is somewhat higher for InSb films than for the bulk material. This is a result of the temperature drift of the mobility already mentioned. In particular, the value of α is greater than that of β. This makes it possible to obtain an initial rise in the Hall voltage with increasing temperature and, by selecting a suitable series resistance, a particularly

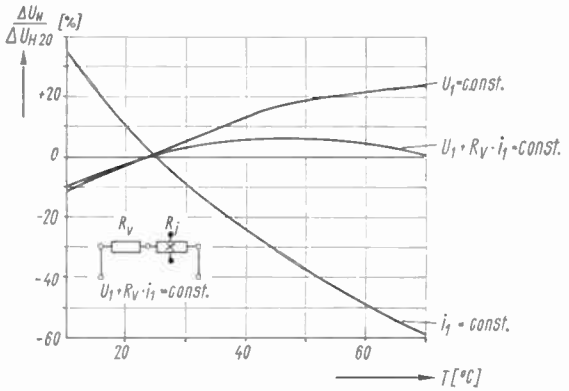


Fig. 13. Temperature dependence of Hall voltage for different operation conditions.

The standard designs can be used within the temperature interval of $-40°C$ to $+100°C$. For higher thermal loads thin film Hall generators with a temperature range of between -10 and $+200°C$ have been built (see Fig. 14).

The pre-requisite for the manufacture of such devices and to maintain narrow scatter ranges for the associated parameters is a definite composition of the evaporated film. This can be guaranteed by the evaporating techniques described at the beginning of this paper. Furthermore, an exact control and constancy of the other variables, such as gas pressure and composition, condition and structure of the condensation surface, and the critical evaporator and condensation temperatures must also be ensured.

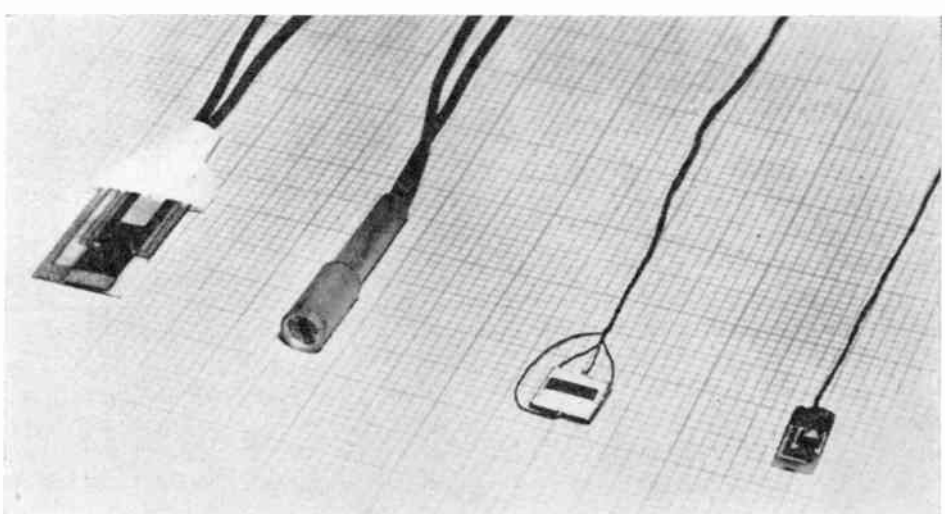


Fig. 14. Typical thin film Hall probes.

small temperature coefficient of the Hall-voltage when operating with constant control voltage (see Fig. 13). Values of about 0·1% per degC, i.e. a temperature drift of the Hall-voltage similar to that of InAs probes are possible.[17]

The characteristics and parameters mentioned have already been confirmed on a large number of evaporated Hall generators and are reproducible. The experience gained covers numerous variations with regard to construction and geometrical dimensions. These also include signal probes, field and axial-field probes and flux-sensitive ferrite probes whose active semiconducting areas vary from 8·4 mm square down to 1·4 × 0·7 mm, depending on the design. Evaporated probes with operating temperatures down to 4·2°K have been developed for the needs of cryogenics.

5. References

1. H. Welker, 'Neue Werkstoffe mit großen Hall-effekt und großer Widerstandsänderung im Magnetfeld', *Elektrotechn. Z., A*, 76, pp. 513–17, 1955.
2. F. Kuhrt, 'Eigenschaften und Anwendungen der Hall-generatoren', *VDE-Fachberichte*, 1956.
3. H. Weiss, 'Feldplatten—magnetisch steuerbare Widerstände', *Elektrotechn. Z.*, B, 17, pp. 289–93, 1965.
4. L. Harris and B. M. Siegel, 'A method for the evaporation of alloys', *J. Appl. Phys.*, 19, pp. 739–41, 1948.
5. R. K. Willardson and H. L. Goering, 'Compound semiconductors', p. 313. (Reinhold, New York, 1962).
6. K. G. Günther, 'Aufdampfschichten aus halbleitenden III-V-Verbindungen', *Z. Naturforschund*, 13a, pp. 1081–9, 1958.
7. H. H. Wieder and A. R. Clawson, 'Structure and galvanomagnetic properties of two-phase recrystallised InSb layers' *Solid-State Electronics*, 8, pp. 467–74, May 1965.

8. J. F. Spivak and J. A. Caroll, 'High-mobility InSb thin films by recrystallization', *J. Appl. Phys.*, **36**, pp. 2321–3, December 1965.
9. C. Juhasz and J. C. Anderson, 'Preparation of high mobility thin films of indium antimonide', 3rd International Vacuum Congress June 28th–July 2nd, Stuttgart, 1965.
10. J. L. Richards, P. B. Hart and E. K. Müller, 'Single Crystal Films', pp. 241–9, (Pergamon Press, London, 1964).
11. K. G. Günther and H. Freller, 'Eigenschaften aufgedampfter InSb-und InAs-Schichten', *Z. Naturforschung*, **16a**, pp. 279–81, March 1961.
12. R. Koike and R. Ueda, 'Crystallite sizes and carrier mobilities in evaporated indium antimonide films', *Jap. J. Appl. Physics*, 3, pp. 191–6, 1964.
13. C. Hilsum and A. C. Rose-Innes, 'Semiconducting III-V-Compounds', p. 128, (Pergamon Press, London, 1961).
14. M. Epstein, 'Current noise in evaporated films of InSb and InAs', *J. Appl. Phys.*, **36**, pp. 2590–91, December 1965.
15. C. Juhasz and J. C. Anderson, 'Electrical properties of flash evaporated epitaxial InSb films', *Physics Letters*, **12**, No. 3, pp. 163–4, 1964.
16. K. G. Günther and H. Freller, 'Hallgeneratoren für extreme Betriebstemperaturen', *Z. Instrumentenkunde*, **74**, No. 2, pp. 52–6, 1966.
17. J. A. Rose, 'Hall-generator-temperature coefficient nomograph', *Electrical Design News*, October 1963.

Manuscript first received by the Institution on 4th April 1966 and in final form on 29th November 1966. (Paper No. 1136.)

I.E.R.E. Graduateship Examination, May 1967—Pass Lists

The following candidates who sat the May 1967 examination at centres outside Great Britain and Ireland succeeded in the sections indicated. The examination, which was conducted at 74 centres throughout the world, attracted entries from 469 candidates. Of these, 219 sat the examination at centres in Great Britain and Ireland and 250 sat the examination at centres overseas. The names of successful candidates resident in Great Britain and Ireland will be published in the July–August issue of the *Proceedings* of the I.E.R.E.

	Candidates appearing	*Pass*	*Fail*	*Refer*
Section A: Great Britain	103	55	28	20
Overseas	147	39	91	17
Section B: Great Britain	116	35	57	24
Overseas	103	22	64	17

OVERSEAS

The following candidates have now completed the Graduateship Examination requirements and thus qualify for election or transfer to Graduate or corporate membership of the Institution.

AMARASENA, S. de S. *Ceylon.*
BEECH, A. L. *Wellington, New Zealand.*
CLIFF, W. N. *Holland.*
CONNOR, J. N. *Auckland, New Zealand.*
DAS, M. *West Bengal, India.*
GEORGE, K. V. *Agra, India.*
GUNAWARDANA, K. K. *Ceylon.*
KHOO, Y. C. *Kuala Lumpur, Malaysia.*
MARTIN, H. J. *Ohio, U.S.A.*
MOORTHY, B. K. *Lucknow, India.*
NADARAJAH, M. *Kuala Lumpur, Malaysia.*
PAKIANATHAN, T. C. *Malaysia.*
RANGACHAR, M. J. S. *Bangalore, India.*
RUPRAI, B. S. *Delhi, India.*
SHARMA, K. L. *Delhi, India.*
SINGH, K. B. *Punjab, India.*
STEVENS, J. E. R. *Wellington, New Zealand.*
TAY KAI MIANG *Singapore.*
VELLUPILLAI, J. *Malaysia.*
WIENET, E. *Israel.*
YOUNG, C. R. *Auckland, New Zealand.*
ZAIDI, A. H. *Lahore, Pakistan.*

The following candidates have now satisfied the requirements of Section A of the Graduateship Examination.

AGHAKHAN, A. S. *Kirkuk, Iraq.*
AZIZ, A. *Pakistan.*
BAINTON, D. J. *Geneva, Switzerland.*
BALASUBRAMANIAM, N. *Ceylon.*
BALASUNDARAM, A. *Ceylon.*
BUESNEL, H. J. *Kingston, Jamaica.*
CHAWATHE, A. K. *Bombay, India.*
COHEN, A. *Israel.*
EASAW, G. J. *West Malaysia.*
EDU, O. F. *Nigeria.*
EKANAYAKE, E. K. P. *Ceylon.*
FERNANDO, K. G. E. *Ceylon.*
FRANKLIN, L. R. *Singapore.*
GOONERATNE, N. P. *Ceylon.*
GRIFFITHS, R. R. *B.F.P.O.* 29.
HANDLEY, K. E. *Barbados.*
HO YING CHIU *Hong Kong.*
JAIN, M. P. *Germany.*
JUWE, S. *Nigeria.*
KOH KIA SWEE *Singapore*
LANGFORD, J. H. *New South Wales, Australia.*
LEV, I. *Israel.*
LOPIAH, C. P. *Ceylon.*
MAHENDRAN, K. *Ceylon.*
MILTON, R. A. J. *Canberra, Australia.*
NG SIOW FAN *Singapore.*
NWADIKE, D. M. *Enugu, Nigeria.*
ODUNDO, I. N. *Nairobi, Kenya.*
ODUSAMI, P. A. *Lagos, Nigeria.*
OYEKENU, J. B. A. *Lagos, Nigeria.*
PUGH, I. G. *Ontario, Canada.*
RAMACHANDRAN, R. *Madras, India.*
SHINTRE, A. K. *India.*
SKAE, F. R. *Salisbury, Rhodesia.*
STA. MARIA, J. A. *Kuala Lumpur, Malaysia.*
SUNDARA KRISHNAM, V. S. *Madras, India.*
WONG BAN KUAL *Singapore.*
ZARYWACZ, S. *Israel.*
ZUCH, E. *Israel.*

French Awards to British Electronic Engineers

At a ceremony held at Cercle des Armées, Paris, on 7th June 1967, five members of the Institution of Electronic and Radio Engineers were honoured by the Société d'Encouragement pour la Recherche et l'Invention, by being awarded medals of 'l'Ordre du Mérite pour la Recherche et l'Invention'.

These awards are made to persons who have attained particular distinction by their activities or by the contribution they have made to the development of research and invention. The Société was founded in 1955 and this was the first time it had awarded medals to engineers from outside France.

The presentations were made by Professor L. Escande, Chancelier General de l'Ordre du Mérite pour la Recherche et l'Invention. The citations were in the following terms:

Grande Médaille d'Or to Admiral of the Fleet the Earl Mountbatten of Burma, K.G., O.M., F.R.S. (Charter President of the I.E.R.E.), who, during a brilliant military and civilian career has constantly exerted a most beneficial influence on the development of technique, the support of research and the development of specialized education: who played a major role in creating the National Electronics Research Council (now the National Electronics Council) and who actively participates in the evolution of techniques on a national and international level, by his association with numerous learned societies and institutions, both British and foreign.

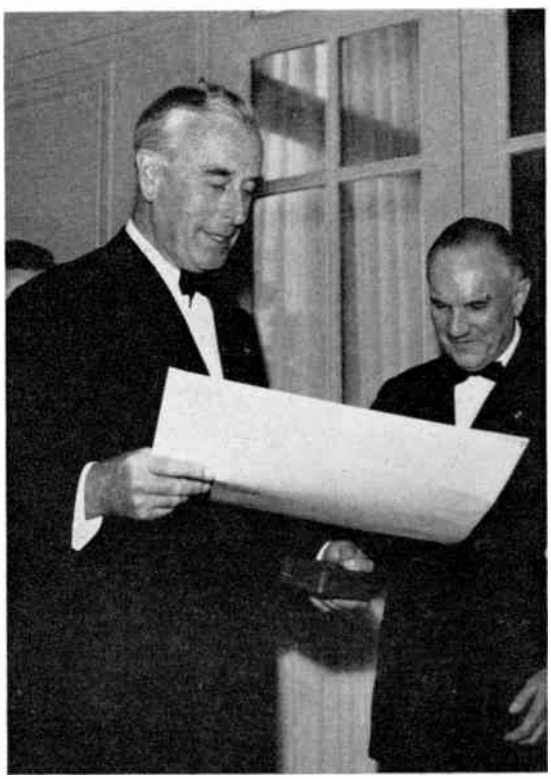
Médaille d'Or to Leslie H. Bedford, C.B.E. (Past President of the I.E.R.E.), who, as a specialist of international reputation, is one of the foremost British research engineers and who is known, among other achievements, for his original research work in television, radar, image orthicon cameras and, more recently, guided weapons.

Médaille de Vermeil to Graham D. Clifford, C.M.G. (Secretary of the I.E.R.E.), who, through his persistent action on a national and international level, has exerted an important influence on, and has made an efficient contribution in, the fields of radio, television and electronics, particularly with regard to the furtherance of education and to organization of symposia and conventions and international co-operation.

Médaille d'Argent to Wing Commander Gerald E. Trevains, R.A.F. (member of the Committee of the Institution's French Section), who, through his functions at the British Embassy in Paris, is in daily contact with British and French engineers and who has devoted himself to the promotion of British-French co-operation in the fields of research and development, particularly in radar and electronics and who has been active in fostering closer relationships between British and French learned societies.

Médaille de Bronze to Paul J. C. Prevost (Secretary of the Institution's French Section), who, during the war contributed to the improvement of radar and within various European organizations has contributed to remote control systems and tests for guided missiles and who is actively contributing to the design and development of the *Europa I* rocket to be used for European satellite launching.

In expressing the appreciation of himself and the other four recipients of the medals, Lord Mountbatten recalled the excellent co-operation which had existed for many years between French and British engineers. He referred particularly to the Institution Convention at Bournemouth shortly after the War, on the possibilities of the use of radar in time of peace for the navigation of shipping and aircraft, when the French Air Arm collaborated by sending French planes to a nearby airport to demonstrate radar navigation.



Lord Mountbatten with Professor L. Escande.

As a further instance of scientific and technical collaboration between the French and English, Lord Mountbatten pointed to the Eurovision system which during the past sixteen years had 'blazed a trail of international visual communication'.

Lord Mountbatten concluded by expressing hopes for future scientific and technical collaboration and pleasure at being associated with the French in all that is new in scientific thought.

U.D.C. 621.314.5

Voltage Stabilized Sinusoidal Inverters Using Transistors

By

C. RIDGERS,
C.Eng., A.M.I.E.R.E.†

Presented at a meeting of the Southern Section held in Bournemouth on 15th November 1966.

Summary: A voltage stabilized, sinusoidal inverter system and an analysis of the operation is described. Design equations are derived for output voltage, distortion, stabilization and regulation, for loads of varying power factor. The particularly important transistor parameters are emphasized, and an inverter design example together with photographs of all relevant waveforms is included.

List of Principle Symbols

A area of stack (in^2)
B peak flux density (kilogauss)
$\cos\theta$ power factor
E_N lowest d.c. input voltage
E'_N, E'_X effective values of E_N and E_X
E_X highest d.c. input voltage
f_0 inverter drive frequency
f_R filter resonant frequency
g_l length of air gap (in)
g_r gap ratio $= \frac{g_l}{l_i}$
I_N d.c. input current on full load, full pulse width at E_N volts
I_{NL} no load input current, full pulse width at approximately E'_N volts
I_X input current pulse, full load at E_X volts at reduced pulse width
L_1 inductance of T1 primary
L_2 inductance of series inductor
L_D inductance of load
L_T external, setting-up inductance
l_i magnetic path length (in)
M $= \frac{R_o}{R_D}$
N number of turns
N_R secondary to primary turns ratio
P_D load power
P_L loss power
P_o total inverter power output
P_R reverse power
P_{in} input power
Q_w working Q-factor
R_D resistance of load
R_{eff} ratio of input voltage to input current at full pulse width at full load conditions
R_o $= \frac{V_o}{P_o}$ = total resistive load on inverter including losses
R_X $= \frac{E'_N}{I_{NL}}$
t_x $= \frac{1}{2f_0}$ = maximum value of τ
V_{ceo} maximum rated collector voltage (with base open-circuit)
V_D, V_o load (i.e. output) voltage
V_{cx} peak collector voltage
V_{pk} general symbol for peak value of voltage
$V_{pk/tr}$ transient peak voltage
V_{pri} primary voltage of T1 (a.c.)
VA product of load current and voltage
VAr volt-ampere reactive $= VA\sqrt{1-\cos^2\phi}$
X $= (1 \pm x) = \frac{f_R}{f_0}$
x $= \frac{\delta f}{f_0}$
Z_R harmonic impedance ratio
$\delta\mu$ a.c. or incremental permeability (usually evaluated at B in use)
$\delta\mu'$ permeability of core and gap
δf $= (f_0 \sim f_R)$
τ base-drive pulse width
η forward efficiency
ω $2\pi \times$ frequency

† Elliott-Automation Space and Advanced Military Systems Ltd., Camberley, Surrey.

1. Introduction

Static inverters using thyratrons or mercury arc rectifiers have been in use since the 1930's. Since the advent of transistors and thyristors, inverters can be made smaller and more efficient, thus widening their field of use.

Early papers[1,2] on inverters do not lay emphasis on sinusoidal operation or output voltage stabilization, although references 1 and 3 demonstrate conditions for sinusoidal generation.

Later papers by Yarrow[4] and Baxandall[5] refer to Class 'D' (switching) inverters of the self-oscillatory type.

This paper is a development of the Class D type of operation in a driven mode and including output voltage stabilization.

2. Load and ½VAr Tuning

Since the inverter includes a tuned circuit, this will be detuned by a reactive load and the extent of this detuning must be limited.

Many loads are inductive. Gyros and synchros exhibit a low power factor, of the order of 0·2 to 0·7, especially on starting up. Any circuitry involving transformers will show an inductive reactance (but occasionally capacitive by transformation) and the order of power factor is 0·85 to 0·9.

Only lagging power factors, i.e. inductive loads, are considered in this paper.

Load parameters are specified in terms of watts, VA, power factor (cos ϕ) and VAr. Any standard electrical text book[9] will give the derivation of these, and a full understanding is necessary before investigating inverters.

2.1. *Load Resistance*

The load power, in watts, is VA cos ϕ, the load resistance will vary from V_D^2/VA to $V_D^2/\text{VA}\cos\phi$, and the Q-factor of the load and overall system will be proportional to the load resistance. With a lagging power the filter detuning is such that the harmonic distortion increases with decreasing power factor since this means an increasing value of X (see Table 1 in Appendix 4 and Fig. 6).

However the increase in Q tends to compensate for this. It is usually necessary to design an inverter to meet a certain specified distortion content in the output voltage and to be safe it is recommended that this be calculated using the minimum value of external load resistance which is likely to be applied, i.e.

$$R_D = \frac{V_D^2}{\text{VA}} \qquad \ldots\ldots(1)$$

The actual inverter load resistance is, in practice, less than R_D because of losses and any deliberate 'bleed'. It is designated R_o and is by definition equal to $M.R_D$.

$$R_o = \frac{MV_D^2}{\text{VA}} \qquad \ldots\ldots(2)$$

The other important parameter is the value of the load inductance and this is given by

$$L_D = \frac{V_D^2}{\omega_0 \text{VA}\sqrt{1-\cos^2\phi}} \qquad \ldots\ldots(3)$$

2.2. *Reactive Loading*

If the inverter filter is tuned exactly to the drive frequency then inductive loads will cause the filter resonant frequency (f_R) to rise (i.e. increasing X) whilst a capacitive load will do the reverse.

Provided the amount of detuning, $\pm x$, is limited and symmetrical, such an arrangement can accomodate loads of leading and lagging power factors.

When the load power factor is always of one sign, say lagging, a different system is necessary in order to accommodate the full range of cos ϕ.

For unity power factor, the load inductance is infinity, whilst at cos ϕ it is L_D. If it be arranged that at unity power factor the detuning is at the maximum agreed amount negatively, i.e. $-x$, and that at cos ϕ, when load inductance is L_D, the detuning is at the maximum agreed amount, positively, i.e. $+x$, then at some intermediate value of the load inductance, the detuning is zero, and the maximum range of power factor variation has been catered for.

There will be, therefore, two limit frequencies and one centre frequency to which the filter, with an external inductance, may resonate and these will be $f_0+\delta f$, $f_0-\delta f$ and f_0. When initially setting up the filter, it must be tuned, with the appropriate external inductance to meet the three frequency conditions above, and this is conveniently done at f_0, the drive frequency, using an external setting up inductance, L_T, the value of which is nearly $2L_D$ (see Appendix 1).

This procedure is termed ½VAr tuning as a consequence of the value found for L_T. When set up the inverter may supply a load of zero power factor equal to the original VAr, i.e. VA $\sqrt{1-\cos^2\phi}$.

3. Inverter Circuitry

The inverter circuit of Fig. 1 is a driven type where the output filter is tuned to the base drive frequency, within the limits of $\pm\delta f$. The drive is obtained from a separate stable source and the bases are square-wave driven. The output frequency is constant, but the effects of detuning the tuned circuit must be investigated.

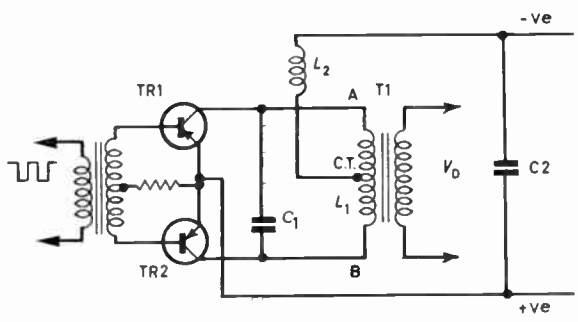


Fig. 1. Basic inverter circuit.

In order to provide voltage control, the square-wave base drive is width modulated, thus the transistors are inactive for a portion of each half cycle. The transient effects of this action will be discussed under Section 6.

From Fig. 1, the action is that TR1 is switched on to a saturated condition, and TR2 is off. Current flows through L_2, T1 half-primary and TR1 collector. The voltage across TR1, (c to e) is not zero, but is finite, small and constant. After a period of one half cycle of the drive waveform, conditions change over, TR1 is switched off and TR2 driven into saturation. Current flows through the other half of T1 from L_2, for one half cycle, when the transistors are again switched. If the inductance L_2 was absent, the d.c. input voltage would appear across the T1 half primary, whilst across the whole would appear twice this, and these voltages would be 'clamped', thus a square wave must appear across L_1 (= T1 primary).

The presence of L_2 'unclamps' the centre tap, C.T., and this point and point B (TR2 collector) are free to assume any voltage. Two other parameters now decide the shape of the waveform and its amplitude. Firstly, if the circuit L_1C_1 is tuned to the drive frequency, then T1 primary voltage is sinusoidal, and secondly the voltage at C.T. must be a double frequency, half sinusoid, with a mean value equal to the d.c. input voltage.

The action of the inverter is often likened to an inductor-input filter, bi-phase rectifier system 'running backwards', and indeed the circuit equations are identical except that these can only be applied to Fig. 1 when the circuit is tuned exactly to the drive frequency and the base drive is full width, i.e. 180° per transistor.

The voltage at point C.T. at exact resonance is a series of half sine waves similar to Fig. 11. At a higher or lower frequency the half sine-wave appears as in Figs. 7(b) or (c) (but at double frequency). Since the mean voltage at this point must be equal to the d.c. (presuming no volts drop in L_2), then, as shown in Appendix 2,

$$V_{pk} = \frac{\pi V_{d.c.}\sqrt{X^4+Q^2(X^2-1)^2}}{2X^2} \qquad \text{.....(4)}$$

When $X = 1$ then

$$V_{pk} = \frac{\pi V_{d.c.}}{2} \qquad \text{.....(4a)}$$

Since this voltage appears across half the primary, the whole primary voltage is

$$V_{pri} = \pi V_{d.c.} \quad \text{peak}$$
$$= 2{\cdot}22V_{d.c.} \quad \text{r.m.s.} \qquad \text{.....(5)}$$

and

$$V_{cx} = \pi V_{d.c.} \qquad \text{.....(6)}$$

Equation (6) is the peak voltage applied to the 'off' transistor and eqns. (5) and (6) are identical to those obtained for the bi-phase rectifier analogy.[6] The ratio of eqns. (4) and (4a) is plotted in Fig. 2 for two values of Q (2 and 10) with X as a variable. An experimental curve is included which tends to follow the theoretical value for low values of X. As X increases the voltage across the tuned circuit rises and the losses increase, thus the Q drops, and the shape of the curve alters accordingly.

In order to obtain this curve, the applied d.c. voltage was kept very low (nearly 2V) to prevent destruction of the transistors.

The conclusions drawn from Fig. 2 and eqn. (4) are that voltage control is advisable and that when first setting up an inverter, which may be greatly out of tune, extreme caution and a low input voltage must be used.

The decision to fit voltage stabilization depends on the application. Gyros and synchros, having widely changing power factors (inductive) and reasonably high Q's will cause the inverter to operate as illustrated in Fig. 2.

For low Q loads, even with low power factors, or purely resistive loads, voltage control may be unnecessary, and every case must be considered individually.

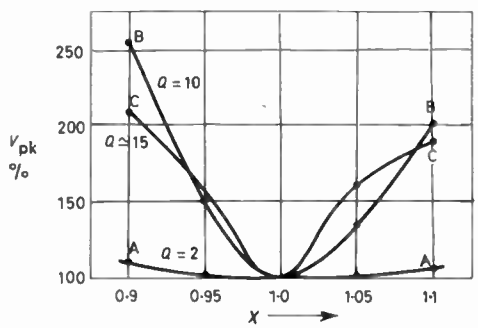


Fig. 2. Variation of output voltage with filter detuning. Curves A and B: calculated values, Curve C: experimental.

Another result of the detuning is illustrated in Fig. 7 where because of the phase difference between the drive waveform and the output, the transistors do not now switch at zero volts, but at some value, say V_R volts. The transistors now switch power and they must therefore be switched rapidly. Also, because of the presence of C_1, V_R volts are commutated, with a change of polarity, from the now 'on' transistor to the now 'off' transistor. The 'off' transistor will however conduct with an opposite polarity potential on its collector and thus a low resistance conducting path now exists through C_1, TR1, TR2, back to C_1. A large current can pass, and one or both transistors will be burnt out.

The reverse current flow can be prevented by placing rectifiers in each collector circuit as shown in Fig. 4. Two diodes known as 'anti-commutation diodes' are fitted, partly for symmetry and partly for added safety against stray capacitance effects. This latter can be seen in Fig. 8 where the charged collector capacitance, isolated by the diodes, falls slower than the transformer half sine-wave. The effect is more pronounced with silicon transistors, and if the voltage remains high, then upon switch-on the transistor will discharge the collector capacitance, again with a high peak current and possible chance of failure. Shunting the transistor with a resistor (say 3·3 kΩ) discharges the capacitor and obviates the trouble (see Figs. 4 and 5).

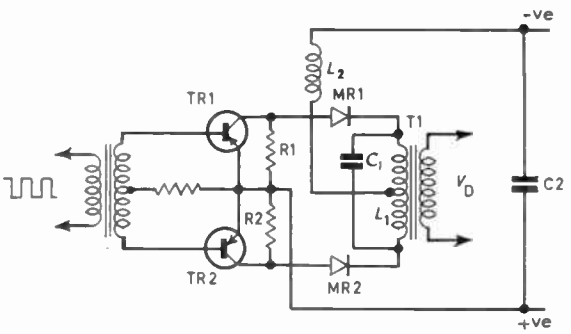


Fig. 4. Inverter with anti-commutation diodes.

4. A.C. Conditions

4.1. *Current Paths*

The inverter circuit may be redrawn as in Fig. 3, which shows the a.c. paths. Note that L_2 appears in parallel with half of L_1, and that a second harmonic current must flow through L_2 and C_2. Therefore the Q of L_2 and C_2 may be important, and the impedance of C_2 at the second harmonic must be low. Usually C_2 is an electrolytic type (1000–10 000 μF) and, in general, the highest possible value, with the greatest ripple rating, is preferred.

From the rectifier analogy, the critical value of L_2 may be calculated from one of the 'standard' formulae[6,7] which involve R_X, the highest d.c. effective resistance as seen by the d.c. power supply.

In order to determine R_X, it is necessary to know the inverter input current, off load, at an input voltage such that full pulse width of the base drive still obtains, at the correct a.c. output voltage. Due to changes in the IR drops in the system, approximations are inherent, but the error may be neglected. The effective value of E_N is E'_N and this is equal to ($E_N - IR$ drop) at full load. Therefore the inverter will supply the correct output voltage at minimum load, but at full pulse width at this value of input voltage, E'_N.

The input current, equal to I_N at full load and full pulse width will become, at no load,

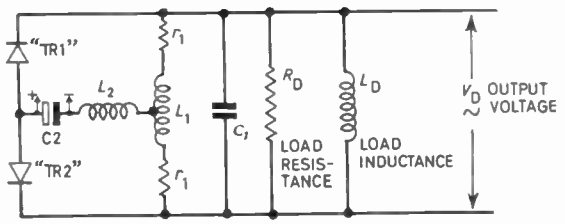


Fig. 3. Inverter a.c. circuit.

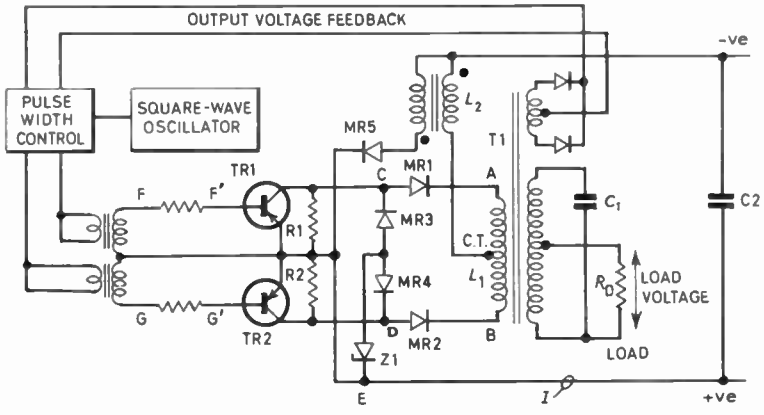


Fig. 5. Final inverter circuit.

$$I_{NL} = I_N \cdot \frac{P_L}{P_o}$$

Therefore

$$R_x = \frac{E'_N}{I_{NL}} = \frac{E'_N P_o}{I_N P_L} \qquad \text{.....(7)}$$

A representative value for $L_{2\,crit}$ is[8]:

$$L_{2crit} = \frac{R_x}{19f}$$

It is preferable that L_2 be about twice the critical inductance and the suggested value is given in eqn. (8)

$$L_2 = \frac{R_x}{10f} \qquad \text{.....(8)}$$

The effect of too small a value for L_2 can be seen in Fig. 9.

4.2. *Minimum Load*

It is advisable always to maintain some load on an inverter. This is also true of inductor-input filter systems, and it is recommended that about 10% of the load power be made a permanent 'bleed' load. Actually the losses of the tuned circuit, transformer etc., contribute to the permanent load and if it can be arranged that the inverter minimum load be some 10% of the output then L_2 is calculated on this basis. The inverter loading (no load to full load) is therefore 10% to 110%. L_2 is nearly twice the critical value for the 10% loading and 22 times the critical value for the full load case. Under these conditions the current drive is a square-wave for full load but at minimum load will appear as a square-wave with a superimposed second harmonic. This latter waveform has an increased peak to mean ratio, but this is of no consequence at the lower power throughput, and consequent lower peak current.

4.3. *Filter Detuning*

The inverter output circuit is not necessarily tuned to the drive frequency and in general will be subject to a detuning of $\pm\delta f$; thus the behaviour of the tuned filter to harmonic currents is of paramount interest. The actual harmonic voltage which appears across the tuned circuit is a function of the harmonic impedance ratio and the harmonic content of the input waveform.

The impedance ratio is given by eqn. (36) in Appendix 4, and numerical values are given in Table 1 and Fig. 6, derived from eqn. (37). In general, a minimum overall Q of 3 and detuning not greater than 5% is indicated.

When the pulse width is reduced for voltage control, then the harmonics reduce and become zero sequentially (Appendix 4). This fact may be used when low distortion is required but then other means must be used for voltage control. Since under voltage control (pulse width variation) conditions, the base drive width may vary from full towards 90°, depending on loading, input volts etc., the harmonic content may vary considerably but will never be worse than the designed conditions for full pulse width, and the reduction in harmonics can only be regarded as a bonus, when it occurs.

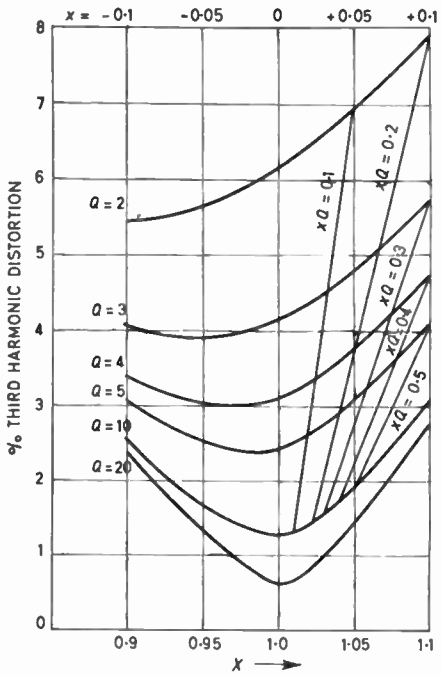


Fig. 6. Graph of percentage 3rd-harmonic distortion against $X(=f_R/f_0)$, for different values of Q.

4.4. *Filter Parameters*

In order to decide the value of L_1 and C_1, consideration must be made of the Q and detuning $\pm\delta f$ in relation to the required distortion. It is possible to obtain complex expressions relating these various factors, but the labour involved is considerable and uneconomic in relation to the accuracy of the final answer, and a simple approach is recommended here.

The filter is correctly tuned when an external inductance L_T is placed in parallel with L_1. The effective inductance has to resonate with C_1 at frequency, f_0, and as shown in Appendix 1 (eqns. (25) and (26)), this effective inductance is equal to $4xL_D$.

Now

$$Q_w = \omega_0 C_1 . R_o = \frac{R_o}{\omega_0 (L_{eff})} \qquad \text{.....(9)}$$

Therefore

$$C_1 = \frac{Q_w}{\omega_0 R_o} \qquad \text{.....(10)}$$

and

$$x.Q_w = \frac{R_o}{\omega_0 4L_D} \qquad \text{.....(11)}$$

Substituting the values of R_o and L_D from eqns. (2) and (3)

$$x.Q_w = \frac{M.\sqrt{1-\cos^2\phi}}{4} \qquad \text{.....(12)}$$

The value of 'xQ' found from eqn. (12) can now be used in eqn. (37) to find the value of x. Alternatively, x can be found from Fig. 6, interpolating if necessary, using the value of xQ found from eqn. (12). In either case, the distortion is that laid down in the original specification of the inverter.

The value of L_1 may now be found from eqn. (25) and C_1 from eqn. (10). Note that the value of the capacitor is that required *at the load voltage.*

4.5. *Filter Components*

Since capacitors are made in discrete steps, and a higher voltage than V_o may be employed, considerable compromise and manipulation of the various parameters is usually necessary. The design example, given later in the paper, illustrates the method.

Evaluation at the load voltage will usually result in an uneconomically large value for C_1. Operation at a higher voltage is possible, by an additional (higher voltage) winding on the transformer and the usual impedance transformation ratios apply, i.e. the capacitor value is reduced by the square of the voltage ratio.

The capacitor current is fixed at $V_c.\omega C$ irrespective of the load on the inverter. If the transformer winding supplying the capacitor has a resistance, R_c, then the power dissipated is:

$$\text{power} = I_c^2.R_c = V_c^2.\omega^2.C^2R_c \qquad \text{.....(13)}$$

It is necessary to compromise on the capacitor voltage since raising this must raise R_c, unless the transformer can be increased in size (giving increased window area and decreased turns per volt). Increasing the transformer size will need more iron and consequently greater losses, hence it is on the selection of the capacitor voltage that most of the 'design compromise' is needed.

It is also important to note that this capacitor current will heat the capacitor, since it too will have losses given by

$$\text{losses} = I_c^2.r_c = V_c^2.\omega^2.C^2r_c \qquad \text{.....(14)}$$

Careful choice of the capacitor is vital, a unit designed for d.c. blocking is not satisfactory. An a.c. capacitor is wound from interleaved paper and foil and has several 'lead-outs' from the length of the foil which are then paralleled. Impregnation is essential, the actual filling depending on temperature rating. Examples of true a.c. capacitors for higher frequencies (400–1200 Hz) are the sealed tropical range of capacitors manufactured by Wego Capacitors Ltd.

5. Voltage Stabilization

The need for stabilization has been highlighted earlier in the discussion on detuning. The control circuitry must hold the output voltage when changes in load, input voltage and load power factor variations, cause V_D to deviate.

There are four basic methods of stabilizing the output voltage:

(1) An inverter, as previously discussed, preceded by a dissipative d.c. voltage stabilizer, with (d.c.) feedback from the inverter output to the stabilizer.

(2) As (1) but using a non-dissipative (switching) stabilizer.

(3) Two inverters, but of approximately half power, with the outputs in series. The 'master' is directly driven whilst the 'slave' drive is delayed in time. The phase shift is made inversely proportional to output voltage.

(4) An inverter where the power drive is less than full cycle, and the 'on' time is controlled by the output voltage.

Method 1 is of use only for the smallest of inverters.

Method 2 is very satisfactory but does require an additional 'black box' which is usually as big as the inverter itself.

Method 3 is slightly more complex, but is a very satisfactory system. It has the disadvantage that the inverters must be designed to cope with the maximum input voltage, that 'double everything' is required, and that when the two series outputs are not exactly in phase (i.e. when the input voltage is high) then at certain angles there can be an increase in the fifth harmonic in the output voltage.

Method 4 which is described in this paper is an elegant and sophisticated system, although it has disadvantages.

By reducing the width of the square-wave base-drive, the tuned circuit is supplied with (square-wave) power for less time. Due to the 'flywheel' effect of the tuned circuit, the voltage across it must remain sinusoidal (provided the Q is not too low) and since the power input is now reduced, then the output voltage must drop also.

The dependence of the output voltage on the pulse width is dealt with in Appendix 5. It is important to arrange that on switching off the bases, they are driven to a positive voltage (for p-n-p) to ensure that holes are swept out of the base rapidly and a fast switch-off

time obtained. Thus a waveform similar to Fig. 10 is required. Such waveforms are usually generated from a square-wave by delaying the 'firing point' of a square-loop saturable reactor (magnetic amplifier). (See Appendix 6.) This means that as the pulse width is decreased, its centre moves to the right. Since the sine-wave across the tuned circuit will set itself symmetrically about the power input pulse, this is tantamount to a delay in time, i.e. a phase delay between the output waveform at full pulse width and at a reduced pulse width. In single-phase inverters, this is usually tolerable, but in three-phase systems, correction for this effect is necessary.

To complete the voltage stabilization system, a direct voltage, proportional to the output, is fed back to the pulse width modulating reactor and the system becomes 'closed-loop'.

6. Reverse Power Circuit

Consider Fig. 1: If the base-drive is reduced, then both transistors will be 'off' together at some period. Previous to this, current flowed through L_2 and charged its magnetic field. This current will now cease and a high voltage ($-L\,\mathrm{d}i/\mathrm{d}t$) will appear across the inductor, which may destroy the transistors. This trouble can be obviated by placing an auxiliary winding on L_2 as in the final circuit arrangement shown in Fig. 5. (Since L_2 with its auxiliary or secondary winding does not operate as a true transformer it is referrred to as inductor of L_2 henry with a secondary winding having a turns ratio N_R.)

The secondary winding placed on L_2 is connected in anti-phase via a diode between the positive and negative rails.

The action is as follows:

When a transistor is 'on', current flows and charges the magnetic field of L_2 ($\frac{1}{2}L_2I^2$). If $\tau/t_x < 1$ (see Appendix 5) then when neither transistor conducts, the primary voltage of L_2 rises. This appears as a secondary voltage of reversed polarity and when this is equal to that of the supply, diode MR5 (the reverse power diode) conducts and the L_2 energy discharges back into the power supply. The power supply must accept the returned power. The inverter auxiliary circuits (i.e. oscillator, control circuits, etc.) will absorb some or all of the current and C_2 will assist by preventing any transient rise if it is sufficiently large. With large inverters, the returned power may be an embarrassment, unless the supply is derived from storage batteries (which, of course, will accept a reverse current).

Because power is returned, then the forward power must be greater by the same amount. This aspect reflects the major disadvantage of this method of control.

Consider the situation at the highest input voltage, E_x. The base-drive pulse width, calculated on the basis of power throughput only, may be, say, 50% wide. The peak current is set by the impedances in the circuit, and these are determined by the initial design carried out at the minimum input voltage. Thus if the input current at E_N is I_N, the peak current at E_X will be:

$$I_X = \frac{E_X}{E_N} \cdot I_N$$

(see Appendix 5).

Now the input power must be sufficient to supply the load and the returned power, and since the peak current cannot increase, then the pulse width of the base drive must do so. A considerable amount of power is then merely 'shunted' backwards and forwards through L_2. The matter is dealt with in detail in Appendix 5.

The turns ratio, N_R, of L_2 must be such that under all normal conditions of operation, at full pulse width drive, the secondary voltage is never great enough to operate the reverse power diode. When voltage control is used, then the primary voltage of T1 is constant and when the input voltage is such that pulse width is maximum (i.e. input voltage is E_N) then the transformer primary voltage is found from eqn. (5). Should X change from unity the primary voltage of T1 increases, but the feedback reduces the pulse width, thus T1 and L_2 never operate with higher voltages than those given by eqns. (4) and (5). Thus the turns ratio, N_R, of L_2, can be calculated for the 'in-tune' case at full pulse width, using an input voltage of E_N (for full load) or E_N' (for no-load).

From the expressions for the peak and mean voltages at the point C.T. (eqns. (4) and (4a)) the excursion above the zero (mean d.c.) line is

$$V_{up} = 0{\cdot}571 \qquad \text{.....(15)}$$

and also

$$V_{down} = V_{d.c.}$$

When current through L_2 is interrupted, the voltage at C.T. rises, and it is this rise which must be 'caught' by MR5. As this is in the same direction as the normal peak (V_{up}) excursion, a phase reversal from primary to secondary is needed and the following equation must be satisfied;

$$N_R \cdot V_{up} < V_{d.c.}$$

and replacing with known data;

$$N_R \cdot 0{\cdot}571 V_{d.c.} < V_{d.c.}$$

or

$$N_R \cdot 0{\cdot}571 < 1$$

The limiting value for N_R is 1·75 and allowing a safety factor, a value from 1·5 to 1·0 is practical. A ratio of 1·5 has given excellent results and is recommended.

No allowance has been made for the volts drop across the diode MR5 or IR drops in the secondary. These both add to the safety factor.

Since L_2 secondary voltage rises to, and cannot rise above $V_{d.c.}$, then the 'upwards' transient on the primary must be $V_{d.c.}/N_R$ above the input voltage, $V_{d.c.}$. Therefore the voltage at C.T. and the collectors must rise to a peak transient voltage of

$$V_{pk/tr} = V_{d.c.} + \frac{V_{d.c.}}{N_R} = V_{d.c.}\left[\frac{1+N_R}{N_R}\right] \quad \text{.....(16)}$$

$V_{d.c.}$ must be taken as the highest input voltage (i.e. E_x) and this transient voltage must be within the peak voltage rating of the transistors.

7. Protection

The inverter needs protection against both voltage and current transients and overloads.

Voltage protection is incorporated in the circuit of Fig. 5, which shows the complete inverter system. The diodes, MR1 and MR2 are the anti-commutation diodes, and MR5 is the reverse power diode associated with the secondary of L_2.

The tuning capacitor, C_1, is placed across a winding of higher voltage than the output load (R_D). The power supply rail is decoupled by C_2 and the transistor drive, with variable width, is supplied from an oscillator via the pulse width control.

Diodes MR3 and MR4 and the Zener diode Z1, protect both transistors against power supply transients and against occasional mishandling during test. Z1 avalanches at a voltage some 10% lower than the agreed safe upper limit collector voltage, and will also protect against internally generated 'spikes' from the iron-cored components (T1 and L_2) on switch-off, and load transients.

Current protection is essential but a more difficult exercise. Although the load conditions can be laid down and therefore the peak current of the transistors calculated, the inverter is likely to be subjected to a short circuit and the nature of the short circuit current waveform is difficult to ascertain. In addition upon switch on under normal conditions, there is an 'in-rush' current through the system which must be catered for. This 'in-rush' depends on the state of the iron-cored components, e.g. whether they have been recently shorted (which leaves the iron in a state of remanence), and on phasing and load conditions. In general, at least twice the maximum full load peak current (at E_X) must be allowed for.

Under short-circuit conditions, the peak current which flows, is limited only by (a) leakage inductance (b) d.c. resistances (c) transistor base drive.

The short-circuit current will flow for a time depending on the type of overload protection system used, and it should be arranged, by a combination of protective devices, that the collector currents cannot reach a figure higher than the agreed maximum value. When a short-circuit appears and the particular protective device operates it is possible that the transistor may have to switch, or be forced to switch, a power of the order of the product of the maximum short circuit current and the maximum d.c. input voltage. Only the short-circuit current is under the control of the designer and by judicious selection of the resistances of L_2 and T1 primary the current drawn from the supply may be limited. There is considerable advantage in fitting a small amount of series (i.e. leakage) inductance in the a.c. output leads, but the merits depend on the inverter application.

8. Transistor Ratings

The power transistor heating is due to current flow and voltage drops, during the following periods:

(1) switch on,

(2) saturation of collector and base,

(3) switch off, and

(4) leakage.

During period (3), the transistor switches considerable power, although the time concerned may be short. The actual power is the product of the peak transient voltage given by eqn. (16) and the peak current from eqn. (41):

$$\text{peak switched power} = E_X\frac{(1+N_R)}{N_R} \cdot E_X\frac{I_N}{E_N}$$

$$= \frac{E_X^2}{E_N} \cdot \frac{I_N}{N_R} \cdot (1+N_R) \quad \text{.....(17)}$$

This switched power is very much greater than the 'static' dissipation, and it is in this mode that most transistor failures occur due to second breakdown.[10,14,15,20] Obviously the transistor chosen must be able to handle this power, and it is worth while noting that manufacturers do not always supply data regarding the transient peak power switching capabilities of their devices. The usually quoted figures for $I_{c(max)}$ and V_{ceo} are no guide, since peak power switching is related to the time of switching. When the designer considers switching times, he must take into account the actual base drive waveform 'fall-time', and the possible deterioration of this in service. Choice of a transistor with an f_T of the order of 1 MHz is advisable for 400 Hz inverters.

During period 2, the dissipation is:

$$P_2 = [V_{ce(sat)} \cdot I_{c(peak)} + V_{eb} \cdot I_b]\frac{\tau}{2t_x} \quad \text{.....(18)}$$

where $I_{c\,(peak)}$ is given by eqn. (41) for the worst case, and V_{cb} and I_b are the required values to obtain the saturated condition at a collector current of $I_{c\,(peak)}$. Usually it is advisable to supply of the order of 50% more base current than is obtained from the sum, I_c/h_{fe}.

The leakage dissipation (period 4) will vary over the half cycle due to the varying collector voltage, and only an estimate can be derived. Therefore taking a maximum figure for the collector voltage:

$$P_4 \simeq [V_{pk/tr} . I_{co}] . \left[\frac{2t_x - \tau}{2t_x}\right] \qquad \text{.....(19)}$$

In the 'switch-off' period, the losses are:

$$P_3 = \int_0^{t_{sw}} P_{sp}\,dt$$

where the peak switched power is that found in eqn. (17).

The switch-on losses are lower but less amenable to calculation since the actual switch-on voltage is not immediately available. From the point of view of safety factor, it is advisable to consider both switching losses to be equal. Therefore, the total switching dissipation is, assuming a linear current decay,

$$P_{sw} = P_1 + P_3 = \frac{t_{sw}}{2t_x} . P_{sp} \qquad \text{.....(20)}$$

where t_{sw} is the time taken to switch off the transistor.

The total dissipation is the sum of eqns. (18), (19) and (20). In practice it may be found that the leakage loss can be ignored.

From the notes above, it is apparent that the most important transistor parameter is the peak power switching ability. This is particularly so if short-circuit-proof inverters are being considered. In general, the designer must ensure that the short-circuit current is limited to the permitted maximum, that the peak power switching ability is adequate, and that when the over-current protection device operates, it will interrupt the current within a millisecond or so of the actual application of the short-circuit. This precludes the use of fuses or magnetic devices, and transistors or thyristor systems are called for.

9. Iron-cored Components

In general, standard practice applies here.[11,12] The inductor L_2 carries d.c. and is designed accordingly, i.e. the flux density is kept as low as possible (say 4 to 8 kilogauss), and the sum of the a.c. and d.c. flux densities kept well below the 'knee' of the B/H curve.

Transformer T1, designed to have an inductance of L_1 across the section feeding C_1, must meet certain conditions. These are:

(a) The inductance of an inductor with an air-gap and not carrying out-of-balance d.c. (i.e. is symmetrical or 'push-pull') is:

$$L = \frac{3{\cdot}2\delta u' . N^2 . A}{l_i . 10^8} \text{ henry}$$

where

$$\delta u' = \frac{\delta u}{1 + g_r \delta u}$$

If δu is high, or g_r, is large, or both, then

$$\delta u \simeq \frac{1}{g_r} = \frac{l_i}{g_1}$$

therefore

$$L = \frac{3{\cdot}2N^2 A}{g_1 . 10^8} \text{ henrys}$$

The inductance is varied by the gap, using a brass clamp.

(b) The number of turns on a winding, relative to the voltage across that winding must satisfy the following relationship:

$$E = 2{\cdot}9\ BANf \times 10^{-3} \qquad \text{.....(21)}$$

The flux density B, must be kept as low as possible to minimize losses, say of order of 4 to 7 kg at 400 Hz.

The standard shape laminations are satisfactory, either 'T' and 'U' or 'E' and 'I'. For both L_2 and T1, brass clamps are essential, and that used for the latter must allow fine adjustment of the inductance. When adjustments are complete, the T1 clamp must allow the whole assembly to be locked with an epoxy resin to prevent change of the inductance, and to be impregnated to decrease acoustic noise.

Either silicon steel or Radiometal can be used, but in general, for 50 Hz inverters, good quality silicon laminations should be used, whilst at 400 Hz and above, use of Radiometal will result in smaller losses.

10. Waveforms

Figures 7 to 17 show typical inverter waveforms. The base-drive is either 180° or 75° wide, and the frequency is always 400 Hz. The filter is tuned exactly to the drive frequency, i.e. ½VAr tuning is not used. Where detuning due to a reactive load has occurred, this is stated in the text. All other cases are for resistive loading.

For the purposes of the photography, the polarity of the c.r.t. plates was reversed, so that, although p-n-p transistors were used, the waveforms are 'standard' for n-p-n or p-n-p, i.e. 'up' is increasing voltage (towards the h.t. line or above) whether positive or negative.

For Fig. 16, an additional phase inversion was applied to obtain the two waveforms in phase. In all cases, the synchronizing waveform was the square-wave base-drive signal before the pulse-width-modulating

stage, i.e. taken from the square-wave oscillator in Fig. 5.

Figs. 7(a), (b) and (c). Fig. 7(a) is the 'standard' collector waveform at exact tune.

Fig. 7(b) is again the collector waveform, but from each side of the diode MR1. The filter is detuned due to the capacitive load and the commutated reverse transient is clearly visible below the upper trace, but is non-existent on the lower trace due to the diode action.

Fig. 7(c) is similar but with an inductive load. Once again the commutated transient is removed by the diode.

An additional effect, apparent on the lower trace, is that the collector voltage tends to follow an exponential decay because of the charged collector capacitance.

Fig. 8, upper trace, shows this more clearly, whilst the lower trace is sinusoidal due to shunting the transistor with 10 kΩ.

Fig. 9 demonstrates the effect of too low a value for L_2.

Fig. 10 shows the out-of-phase base drive. Note particularly that the bases are switched off to a reverse bias.

Fig. 11 is similar to the filter waveform of an inductor-input system rectifier unit, although additional transient effects are visible here.

Fig. 12 shows output voltage against the narrowed base drive. The two waveforms are in phase.

Fig. 13 is the same, but with full pulse width. Again the base and output are in phase, but the output phase has shifted 'left' as compared to Fig. 12. There is also an increase in distortion due to the increased harmonic content of the collector current.

Fig. 14 gives the complex collector and transformer waveform for narrowed pulse width of the base drive. The reverse-power transient can be clearly seen above the sine-wave.

Fig. 15 shows the centre-tap waveform at 75° base-drive. The reverse-power transient is clearly evident.

Fig. 16 is the waveform of the current taken by breaking the line at I (in Fig. 5) and inserting a low resistance across which the c.r.o. is connected. Forward current is the upward 'square' pulse whilst the returned current below the base line, shows the linear run-down mentioned in the text. Two points of interest are that L_2 discharges completely and that the forward pulse shows an exponential rise.

Fig. 17, upper trace, was taken with an input voltage of 2 V, and with the reverse power diode disconnected. The transient peaks rise to approximately 24 V. The lower trace has the diode reconnected and demonstrates its efficacy.

11. Design Example

Consider the design of an inverter to supply a mixed load of:

(a) three 115 V, 400 Hz synchro control chains, each chain requiring 7·5 W (Muirhead size 18).

(b) resistive load of 25 to 60 W.

and used under the following conditions:

(i) synchros applied, with stators connected, with an additional (parallel) load of 25 W,

(ii) no synchros applied, but resistive load of 60 W,

(iii) synchros, only, applied, with rotors open-circuit.

When the synchro chains are open-circuited, the inverter will see the rotor impedance alone, and this is 110–j1100 Ω.† The open-circuit load, for the three parallel chains, is therefore 147 mH (the resistance term is ignored), and the total loading on the inverter under the above three conditions is:

(i) 36 VAr (= 147 mH at 115 V) + 22·5 W + 25 W. (Total ≃ 48 W.)

(ii) 60 W.

(iii) 36 VAr.

However, the inverter specification is likely to be given in 'standard form' similar to that described in Section 11·1, and therefore the latter is used as the design starting point.

11.1. *Power and Load Requirements*

VA	75
load	inductive
cos ϕ	0·8 to 1·0
voltage	115 V
frequency	400 Hz
distortion	5% total
Output voltage stabilization and regulation, not greater than 1%	against input voltage change and load change of zero to full load
Input voltage	24 to 28 V d.c.

Inverter to be capable of driving reactive load of full VAr.

11.2. *Design*

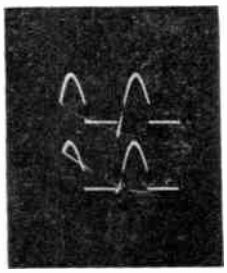
Conditions at E_N (from eqns. (1), (2) and (3)):

Load conditions	VA	W	VAr	cos ϕ	R_D
1	60	48	36	0·8	276 Ω
2	60	60	0	1·0	220 Ω
3	36	0	36	0	

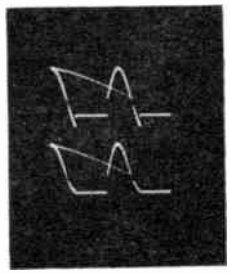
† Data from Muirhead Ltd.

Figs. 7 to 17. Inverter waveforms,. Connections refer to Fig. 5. For explanations see Section 10.

Fig. 7. (a)
Upper trace: A to E
Lower trace: F to E
$X = 1$



(b)
A to E
C to E
$X = 0{\cdot}9$



(c)
A to E
C to E
$X = 1{\cdot}1$

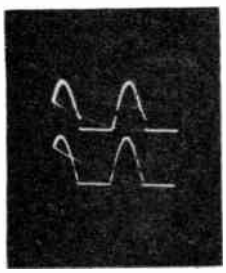


Fig. 8.
Upper trace: C to E
Lower trace: C to E

Fig. 9.
A to E
—

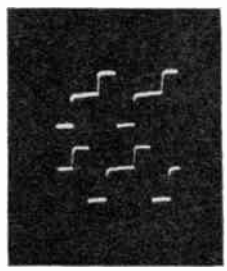


Fig. 10.
C to E
F to E

Fig. 11.
Upper trace: CT to E
Lower trace: —

Fig. 12.
F to E
Load voltage

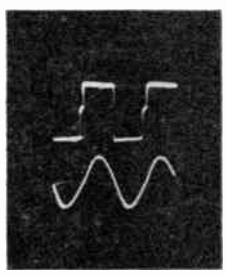


Fig. 13.
F to E
Load voltage

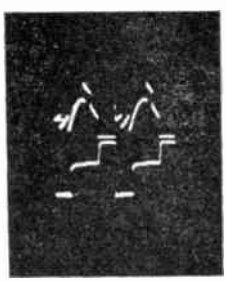


Fig. 14.
Upper trace: A to E
Lower trace: F to E

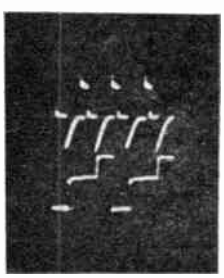


Fig. 15.
CT to E
F to E

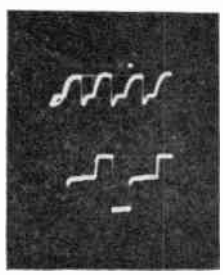


Fig. 16.
E to + ve line
F to E

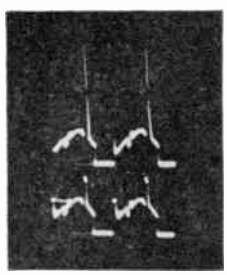


Fig. 17.
A to E
A to E

The tuning will be $\frac{1}{2}$VAr which will satisfy condition (3), and conditions (1) and (2) are met by designing for 60 W with a safety factor of, say, 10%. The inverter bleed will be of the order of 10% and the tuned circuit should be designed for losses of the order of 10% of the load.

Therefore, load + safety factor	66 W
bleed	4 W
losses	10 W
Total	80 W

Now VAr = 36 and $\frac{1}{2}$VAr = 18.

So $L_D = 147$ mH (eqn. (3))

and

$L_T = 294$ mH (eqn. (26)).

From eqns. (1) and (2):

$$R_D = 220, \qquad R_o = \frac{115^2}{80} = 165\,\Omega$$

and $M = 0{\cdot}75$

Allow 1 V drop for transistors, 0·5 V each for L_2 and T1 primary, and 1 V for series diodes. Therefore, design for full power at full pulse width at 21 V input.

Therefore

$$E'_N = 21 \text{ V} \qquad \text{and} \qquad E'_X = 25 \text{ V}$$

Then

$$V_{cx} = 66 \text{ V (eqn. (6))}$$

and

$$V_{pk/tr} = 46{\cdot}7 \text{ V (eqn. (16))}$$
$$(V_{dc} = 28 \text{ V and } N_R = 1{\cdot}5)$$
$$V_{pri} = 26{\cdot}6 \text{ V r.m.s. (eqn. (5))}$$

and turns ratio, output winding to whole primary, is:

$$\frac{115}{46{\cdot}6} = 2{\cdot}47$$

Some allowance should be made for any secondary IR drop.

Input current at

$$E_N = \frac{\text{total power}}{V_{input}} = \frac{80}{21} = 3{\cdot}82 \text{ A}$$

Input current at E_X will be

$$\frac{28}{24} \times 3{\cdot}82 = 4{\cdot}45 \text{ A}$$

Power input $= 24 \times 3{\cdot}82 = 92$ W

Efficiency may be calculated in various ways and, representatively,

$$\eta = \frac{70}{92} = 76\%$$

At E_N and off load, input current is

$$I_{NL} = \frac{\text{minimum power}}{E'_N} = \frac{14}{21} = 0{\cdot}67 \text{ A}$$

Therefore

$$R_x = \frac{E_N}{I_{NL}} = \frac{21}{0{\cdot}67} = 31{\cdot}5\,\Omega \text{ (eqn. (7))}$$

and

$$L_2 = 8 \text{ mH (eqn. (8))}$$

This inductor must have a volts drop of not greater than 0·5 V at 4·45 A and have a secondary of 1·5 times the primary turns, with a resistance yet to be decided.

From eqn. (12), $xQ = 0{\cdot}1125$ and using this in Fig. 6, for a distortion of 4·5%, gives a value of $x = 0{\cdot}04$.

Consideration of the stability of the tuned circuit filter under environmental conditions, particularly temperature changes, suggests that selection of $x = 0{\cdot}025$ and consequent $Q_w = 4{\cdot}5$ is preferable.

From eqn. (25) $L_1 = 15{\cdot}5$ mH

From eqn. (10) $C_1 = 10{\cdot}9\,\mu$F (at 115 V)

It is advisable here, in a practical design, to check the values of x and Q using the equations given in Appendix 1 for the exact case. A check gives $x = \pm 0{\cdot}025$. By operating the capacitor at 240 V the value is reduced to 2·5 μF and capacitor current is:

$$I_c = V_c \omega C = 1{\cdot}5 \text{ A}$$

This is a more economical arrangement but one must note that I_c flows through a 240 V winding on T1 and if the IR losses must not exceed, say, 7 watts, then the winding resistance must not be greater than about 3 Ω (eqn. (14)). This single, simple, fact virtually settles the dimensions of T1.

In ordinary transformer practice, winding resistances of 3 Ω at coil voltages of 240 V using a 1:1 transformer having 5% regulation correspond to a 480 VA transformer. This inverter is 80 VA and the ratio 480/80 = 6:1 is a useful rule-of-thumb comparison. For less demanding applications the ratio is 3:1, but it cannot be reduced much below this.

Waveform distortion is nearly inversely proportional to transformer size.

From eqn. (45) for full load conditions:

$$\frac{\tau}{t_x} = 0{\cdot}842 \qquad \text{and} \qquad P_R = 13{\cdot}3 \text{ W (eqn. (46))}$$

$$I_R = \frac{\text{max. forward current}}{\text{turns ratio}} = \frac{4{\cdot}45}{1{\cdot}5} = 2{\cdot}96 \text{ A}$$

Confirmation:

$$\text{forward power} = E_x . I_{F\,(max)} \frac{\tau}{t_x}$$
$$= 28 \times 4{\cdot}45 \times 0{\cdot}842 = 105 \text{ W}$$

This must equal sum of power required by inverter and returned power:

$$\text{Inverter power} = 92\text{ W}$$
$$\text{Returned power} = 13{\cdot}3\text{ W}$$
$$\text{Total} = 105{\cdot}3\text{ W}$$

When the inverter is unloaded it will operate at the same degree of mistuning at E_N and at E_X and therefore from eqn. (48),

$$\frac{\tau}{t_x} = \left(\frac{21}{28}\right)^2 = 0{\cdot}5625$$

If the feedback voltage is 24 V and is rectified and 'backed-off' against a Zener voltage, then ±1% change on output voltage corresponds to

$$2\% \times 24\text{ V} = 0{\cdot}48\text{ V},$$

and the control system must be capable of reducing pulse width from full to 0·5625 (or 0·5 with safety margin) with a control voltage of change of 0·48 V.

11.3. *Choice of Transistor*

During switching the transistor handles considerable power, albeit for a short period. Nevertheless, it is in this mode that most breakdowns occur, usually due to second breakdown.[10,14–16,20] It is important therefore, to ensure that the device is operating inside the power rating curve for the switching period considered, and if this information is not available, then within the static power rating characteristic.

For this application the peak switched power is

$$V_{pk/tr} \times I_x = 46{\cdot}7 \times 4{\cdot}45 = 207\text{ W}$$

Immediately the inverter is switched on the pulse width is maximum and the feedback voltage zero, therefore there is a possibility that for one cycle, the peak collector voltage may be

$$V_{cx} = 28\pi = 87{\cdot}5\text{ V}$$

Again on switch-on, a surge current flows to 'charge' the inductances and this surge may be greater if the output has previously been shorted. Any overload system needs an overcurrent to operate and the transistor must be able to supply this current. An overload ability of at least 100% is recommended and for this unit therefore, a 10 A rating is necessary.

The choice of a transistor is obviously limited to a silicon high voltage type and the following characteristics are reasonable:

$$V_{ce} = 100\text{ V} \qquad P_{sw} = 250\text{ W} \qquad I_c = 10\text{–}15\text{ A}$$

The 2N4348[16] meets this specification and has ratings of V_{ceo} of 120 V, I_c of 10 A and static power dissipation of 120 W at 25°C case temperature; the dissipation is approximately 20 times this for peak switching power, with switching speeds of nearly 15 μs.

Although the 2N4348 can be switched in 15 μs or less, such a figure may only be met in regenerative circuitry. Considering that a pulse-width modulating system of base drive is to be used, it is unlikely that the transistor will be switched faster than say 25 μs.

Obviously more circuitry may be used to ensure more rapid switching but economics must be studied as well as 'in-service' deterioration. Therefore, a '2%' switching-time is allowed for this design (i.e. 25 μs), and a check must be made that the semiconductor will handle the peak power for this period, i.e. 207 W for 25 μs.

From eqns. (17), (18) and (19), the transistor dissipation can be found. The peak collector current is 4·5 A, but to ensure that the transistor always remains saturated, it should be 'forced' by applying more base-drive than is strictly necessary. Thus, choosing a collector current of 10 A, the base drive is (from the data sheet) 0·8 A at 1·4 V. Therefore

$$P_2 = (1 \times 4{\cdot}45 + 0{\cdot}8 \times 1{\cdot}4)\,0{\cdot}842 = 4{\cdot}77\text{ W (eqn. (17))}$$

$$P_4 = \text{negligible}$$

$$P_{sw} = \frac{25}{10^6} \times \frac{10^3}{1{\cdot}25} \times 4{\cdot}45 \times 47{\cdot}4 = 4{\cdot}23\text{W (eqn. (19))}$$

Total dissipation = 9 W.

12. Conclusion

The operation and analysis of a controlled transistor inverter are shown to be relatively simple if load parameters and transistor characteristics are known.

It is shown that the size of the output transformer is a function of capacitor current, and that the choice of transistor is dependent on its peak power switching capabilities.

The usual compromises required in all electronic equipment are highlighted and satisfactory values suggested.

13. Acknowledgments

Full acknowledgment is made to Messrs. C. J. Yarrow and P. J. Baxandall for their original papers which stimulated the author's interest in this study.

14. References

1. C. F. Wagner, 'Parallel inverter with inductive load', *Elect. Engng*, **55**, pp. 970–80, September 1936.
2. C. F. Wagner, 'Parallel inverter with resistance load', *Elect. Engng*, **54**, pp. 1227–35, November 1935.
3. F. N. Tompkins, 'Operation of self-excited inverter', *Electronics*, **13**, pp. 36–9, and 81, September 1940.
4. C. J. Yarrow, 'Transistor convertors for the generation of high-voltage low-current d.c. supplies', *Proc. Instn Elect. Engrs*, **106B**, pp. 1320–4, May 1959. (I.E.E. Paper No. 2929E.)

5. P. J. Baxandall, 'Transistor sine-wave LC oscillators', *Proc. I.E.E.*, **106B** Suppl. No. 16, pp. 748–58, 1959. (I.E.E. Paper No. 2978E); also 'The elimination of residual even-harmonic distortion in transistor oscillators', *Electronic Engng*, **36**, pp. 97–9, February 1964.
6. F. E. Terman, 'Radio Engineers Handbook', Section 8, p. 589 (McGraw-Hill, New York, 1943).
7. R. W. Landee, D. C. Davis and A. P. Allbrecht, 'Electronic Designers Handbook', Section 15 (McGraw-Hill, New York, 1957).
8. F. E. Terman, 'Radio Engineers Handbook', p. 601 (McGraw-Hill, New York, 1943).
9. 'Standard Handbook for Electrical Engineers', Ed. A. E. Knowlton, Section 2–163 (McGraw-Hill, New York, 1957).
10. P. Schiff, 'Preventing second breakdown in transistor circuits', *Electronics*, **37**, No. 18, pp. 66–74, 15th June 1964.
11. R. W. Landee, D. C. Davis and A. P. Allbrecht, 'Electronic Designers Handbook', Section 14 (McGraw-Hill, New York, 1957).
12. V. G. Welsby, 'Theory and Design of Inductance Coils', 2nd edition, Chapter V, p. 54 (Macdonald, London, 1960).
13. F. E. Terman, 'Radio Engineers Handbook', Section 3, pp. 145–9 (McGraw-Hill, New York, 1943).
14. 'Characterization of Second Breakdown in Silicon Power Transistors', RCA Application Note, No. SMA–30.
15. 'Second Breakdown in Transistors under conditions of cut-off', RCA Application Note, No. SMA–21.
16. Data Sheets for: RCA Types 2N4348, 2N3773, 2N4240; Motorola Types 2N3448, 2N3487.
17. 'Control Engineers Handbook', Ed. J. G. Truxall, Section 7 (McGraw-Hill, New York, 1958).
18. G. M. Ettinger, 'Magnetic Amplifiers' (Methuen, London, 1953).
19. H. F. Storm, 'Magnetic Amplifiers' (Wiley, London, 1955).
20. 'Power Transistor Handbook', 3rd printing, Motorola Semiconductors Products Division Inc.

15. Appendix 1

The three filter resonant frequencies due to the (external) addition of shunt inductances are:

$$\omega_a = \frac{1}{\sqrt{L_a C_1}}$$

where

$$L_a = \frac{L_1 L_D}{L_1 + L_D}$$

and

$$_a = f_0(1+x)$$

$$\omega_b = \frac{1}{\sqrt{L_b C_1}}$$

where

$$L_b = \frac{L_1 L_T}{L_1 + L_T}$$

and

$$_b = f_0$$

$$\omega_c = \frac{1}{\sqrt{L_1 C_1}}$$

where

$$f_c = f_0(1-x)$$

From

$$\frac{\omega_b}{\omega_c} = \sqrt{\frac{L_1}{L_b}}$$

and substituting:

$$L_T = \frac{(1-x)^2 L_1}{x(2-x)} \qquad \text{.....(22)}$$

Ignoring x^2 terms

$$L_T \simeq \frac{(1-2x)L_1}{2x} \qquad \text{.....(23)}$$

From

$$\frac{\omega_a}{\omega_b} = \sqrt{\frac{L_b}{L_a}}$$

and substituting,

$$L_T = \frac{(1+x)^2 L_D L_1}{L_1 - L_D . x . (2+x)} \qquad \text{.....(24)}$$

From eqns. (22) and (24),

$$L_1 = \frac{x(2-x)(1+x)^2 + (1-x)^2 x(2+x)L_D}{(1-x)^2}$$

Simplifying by ignoring third and higher power of x

$$L_1 \simeq \frac{4xL_D}{(1-2x)} \qquad \text{.....(25)}$$

Replacing in eqn. (23),

$$L_T \simeq 2L_D \qquad \text{.....(26)}$$

If $x = 0{\cdot}1$, then from eqn. (25),

$$L_1 = \frac{L_D}{2} \qquad \text{.....(27)}$$

If $x = 0{\cdot}05$,

$$L_1 = \frac{L_D}{4{\cdot}5} \qquad \text{.....(28)}$$

It must be remembered that eqns. (24) to (28) are all approximate and in approximating the value of L_1 eqn. (25) should be used.

16. Appendix 2

Phase angle of a parallel tuned circuit

$\omega_0 = 2\pi f_0$ (f_0 = operating frequency)

$\omega_R = X . \omega_0 = 2\pi f_R$ (f_R = resonant frequency)

$$Y = \frac{1}{j\omega_0 L} + \frac{\omega_0 C}{j} + \frac{1}{R}$$

$$= \frac{1 - \frac{\omega_R^2 LC}{X^2} + \frac{j\omega_0 L}{R}}{\frac{j\omega_R L}{X}}$$

Now

$$\omega_R^2 LC = 1$$

and

$$\frac{\omega_0 L}{R} = \frac{1}{Q}$$

Therefore

$$Y = \frac{1 - jQ\left(1 - \frac{1}{X^2}\right)}{Q\omega_0 L} \qquad \text{.....(29)}$$

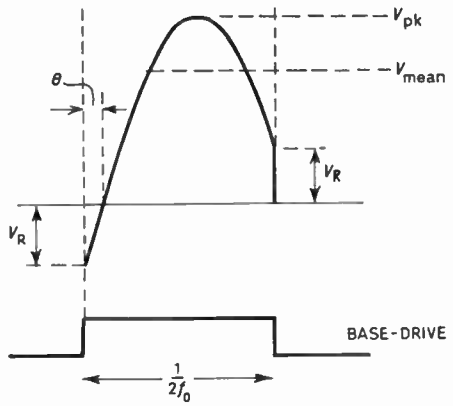


Fig. 18. Collector waveform.

From Fig. 18, θ is the delay (or advance) due to the phase angle of the tuned filter, and from eqn. (29):

$$\theta = \tan^{-1}\frac{Q(X^2-1)}{X^2}$$

and

$$\cos\theta = \frac{X^2}{\sqrt{X^4 + Q^2(X^2-1)^2}}$$

Now mean value of Fig. 18 must equal the direct voltage applied, therefore

$$V_{mean} = V_{d.c.} = \frac{1}{\pi}\int_{-\theta}^{\pi-\theta} V_{pk}\sin\omega_0 t \; d(\omega_0 t)$$

which reduces to

$$V_{d.c.} = \frac{2V_{pk}}{\pi}\cos\theta$$

Therefore

$$V_{pk} = \frac{\pi V_{d.c.}\sqrt{X^4 + Q^2(X^2-1)^2}}{2X^2} \qquad \text{.....(30)}$$

and when $X = 1$,

$$V_{pk} = \frac{\pi V_{d.c.}}{2} \qquad \text{.....(31)}$$

17. Appendix 3

The current drive waveform to the inverter tuned circuit is shown in Fig. 19. This can be analysed by using a Fourier series:

$$I = c + a_1\sin\omega t + a_2\sin 2\omega t + a_3\sin 3\omega t \ldots + b_1\cos\omega t + b_2\cos 2\omega t + b_3\cos 3\omega t \ldots \quad (32)$$

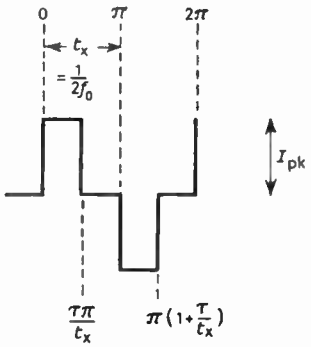


Fig. 19. Waveform of current drive to filter.

From inspection of Fig. 19, c is zero and only odd harmonics are present.

To find 'a_n':

$$a_n = \frac{1}{\pi}\int_0^{2\pi} I_{pk}\sin n\omega t \; d(\omega t)$$

$$= \frac{1}{\pi}\int_0^{\frac{\tau\pi}{t_x}} I_{pk}\sin n\omega t \,.\, d(\omega t) - \frac{1}{\pi}\int_{\pi}^{\pi(1+\frac{\tau}{t_x})} I_{pk}\sin n\omega t \; d(\omega t)$$

This reduces to

$$a_n = \frac{2I_{pk}}{n\,.\,\pi}\left[1 - \cos\frac{\tau}{t_x}n\pi\right] \qquad \text{.....(33)}$$

Solving this for values of n and τ/t_x will give the coefficients of the sine terms in eqn. (32).

Similarly, for b_n:

$$b_n = \frac{1}{\pi}\int_0^{2\pi} I_{pk}\cos n\omega t \; d(\omega t)$$

This reduces to $b_n = 0$, thus all cosine terms vanish. Therefore if $\tau/t_x = 1$, i.e. full pulse width, eqn. (32) becomes (from eqn. (33)),

$$I = \frac{4I_{pk}}{\pi}[\sin\omega t + \tfrac{1}{3}\sin 3\omega t + \tfrac{1}{5}\sin 5\omega t \ldots] \qquad \text{.....(34)}$$

Individual harmonics become zero, sequentially, with pulse width reduction, and, for example, if $\tau/t_x = 2/3$ ($= 120°$) then eqn. (32) becomes

$$I = \frac{4I_{pk}}{\pi}[0{\cdot}933\sin\omega t + 0{\cdot}15\sin 5\omega t \ldots] \quad \ldots\ldots(35)$$

18. Appendix 4

If the drive frequency is ....f_0

and the nth harmonic ....nf_0

The filter resonant frequency ....$f_R = Xf_0$

then the nth harmonic is ....nf_R/X

Now the impedance at operating frequency is ....Z_1

and at nth harmonic is ....Z_n

Therefore impedance ratio, Z_R is Z_n/Z_1 and is equal to Y_1/Y_n.

Therefore

$$Z_R = \frac{\dfrac{1}{j\omega_0 L} - \dfrac{\omega_0 C}{j} + \dfrac{1}{R}}{\dfrac{1}{jn\omega_0 L} - \dfrac{n\omega_0 C}{j} + \dfrac{1}{R}}$$

and

$$Z_R = \frac{1 - \dfrac{\omega_R^2}{X^2}LC + \dfrac{j\omega_0 L}{R}}{\dfrac{1}{n} - \dfrac{n\omega_R^2 LC}{X^2} + \dfrac{j\omega_0 L}{R}}$$

Now

$$\omega_R^2 LC = 1$$

and

$$\frac{\omega_0 L}{R} = \frac{1}{Q}$$

Therefore

$$Z_R = \frac{Q\left(\dfrac{X^2-1}{X^2}\right) + j}{Q\left(\dfrac{X^2-n^2}{nX^2}\right) + j} \quad \ldots\ldots(36)$$

where $n = 3, 5, 7$ etc.

From Appendix 3, eqn. (34), the harmonic content of the current drive waveform is obtained. Therefore the 3rd to fundamental voltage ratio is

$$V_{\cdot(3rd/1st)} = \tfrac{1}{3}Z_R = \frac{1}{3}\left|\frac{Q\left(\dfrac{X^2-1}{X^2}\right) + j}{Q\left(\dfrac{X^2-9}{3X^2}\right) + j}\right| \quad \ldots(37)$$

This expression has been calculated for various values of X and Q, and the results are given in Table 1 and Fig. 6. It is interesting to note that, since extreme stability of the tuned filter is unlikely in practice, i.e. that x is never zero, and since high-Q's are impossible at low frequencies (400 Hz), then an inverter cannot be designed for 1% distortion without additional output filtering.

An estimate of the total harmonic content can be obtained thus:

For harmonics above the 3rd, the filter impedance is approximately inversely proportional to the frequency, therefore the harmonic voltages are (for 3rd, 5th and 7th)

$$V_H = \frac{4}{\pi}I_{pk}[\tfrac{1}{3}Z_3\sin 3\omega t + \tfrac{1}{5}\tfrac{3}{5}Z_3\sin 5\omega t + \tfrac{1}{7}\tfrac{3}{7}Z_3\sin 7\omega t]$$

$$V_H = \frac{4}{\pi}Z_3 I_{pk}[\tfrac{1}{3}\sin 3\omega t + \tfrac{3}{25}\sin 5\omega t + \tfrac{3}{49}\sin 7\omega t] \ldots(38)$$

and r.m.s. is

$$V_{H\,(r.m.s.)} = \frac{4}{\pi}Z_{(3)}I_{pk}\sqrt{\tfrac{1}{9}\tfrac{1}{2} + 6\tfrac{9}{25}\tfrac{1}{2} + \tfrac{9}{2400}\tfrac{1}{2}}$$

$$= \frac{I_{pk}Z_{(3)}0{\cdot}254}{\pi} \quad \ldots\ldots(39)$$

where Z is the filter impedance to the 3rd harmonic.

Now the fundamental r.m.s. voltage is

$$V_{f\,(r.m.s.)} = \frac{4I_{pk}}{\pi}Z_f 0{\cdot}707$$

The ratio is therefore

$$\frac{V_{H(r.m.s.)}}{V_{f\,(r.m.s.)}} = \frac{Z_R \times 1{\cdot}08}{3}$$

where $$Z_R = \frac{Z_{(3)}}{Z_f}$$

From eqn. (37), the total harmonic voltage content becomes

$$1{\cdot}08 \times \%\,3\text{rd harmonic} \quad \ldots\ldots(40)$$

where the 3rd harmonic is given in Table 1.

Table 1

Percentage 3rd harmonic

$X =$	0·9	0·95	1·0	1·05	1·1	
$x =$	−0·1	−0·05	0	+0·05	+0·1	$(= \delta f/f_0$
Q						
2	5·44	5·63	6·13	6·9	7·9	
3	4·08	3·9	4·13	4·77	5·7	
4	3·58	3·03	3·1	3·71	4·7	
5	3·08	2·54	2·4	3·07	4·06	
10	2·52	1·66	1·25	1·88	3·07	
15	2·44	1·43	0·83	1·57	2·86	
20	2·38	1·34	0·625	1·45	2·77	

Total harmonic distortion (r.m.s.) is approximately 10% greater than above figures.

In the text and the design example, the convenience of the composite factor, 'xQ_w' is illustrated.

Constant xQ curves are drawn across the distortion curves of Fig. 6, for values from 0·1 to 0·5. These curves are straight lines between the limits of $Q = 10$ and $Q = 2$.

It will be noted that in the previous expressions, Q is given as equal to $\omega_0 L/R$ and not $\omega_R L/R$. This is because only the constants of the former are known or are immediately available. It will also be noted, from Fig. 6, that at the lower Q values, the actual value of X corresponding to minimum distortion is not unity. This follows, of course, well-known theory on low-Q tuned circuits[13] and there is some advantage in deliberately introducing an offset factor from 1·0 to say 0·95 or 0·9. The advantages resulting from this can only be determined by practical tests since many factors will affect the decision. For example, the tuned filter tuning will drift as the ambient and its own temperature change, and if the drift is towards $X > 1$ then initial tuning of X less than unity is advantageous. Also, it must be noted that Fig. 6 is calculated for constant Q, and since the losses and hence the final Q will vary with X, the curves of Fig. 6 will be modified in practice.

19. Appendix 5

Considering Figs. 20 and 21, during time t_R, current flows from L_2 secondary back into the supply. The equation $E = -L\,\mathrm{d}i/\mathrm{d}t$ must be satisfied, therefore the discharge current is linear. The maximum current, I_R, must be I_p/N_R, operating from an inductance (on the secondary) of $L_2 N_R^2$. If L_2 is high, $\mathrm{d}i/\mathrm{d}t$ is small and the returned power may be regarded as constant over the time t_R.

The value of the input current at E_N is I_N and will be as eqn. (41) at E_X.

$$I_X = \frac{E_X}{E_N} \cdot I_N \qquad \text{.....(41)}$$

More accurately, the effective values of E_N and E_X should be used here (i.e. E_N and E_X minus the circuit voltage drops) but since these can be only estimates, and there is little difference in the two ratios, the known value is preferred.

During period τ, the forward power is

$$P_F = P_O + P_{(\text{losses})} + P_{(IR\,\text{drops})} + P_R = E_N . I_N + P_R \qquad \text{.....(42)}$$

where P_R is the returned power during period t_R.

An approximation to P_R is

$$P_R = E_X \frac{I_p}{N_R} \frac{t_R}{t_x}$$

where I_p = primary current of $L_2 = E_X I_N / E_N$.

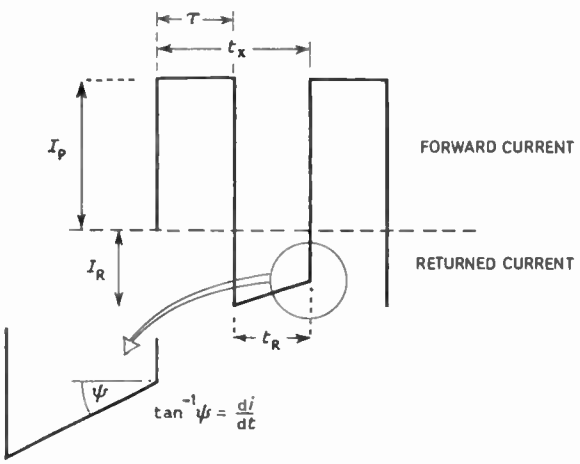


Fig. 20. Input current waveform.

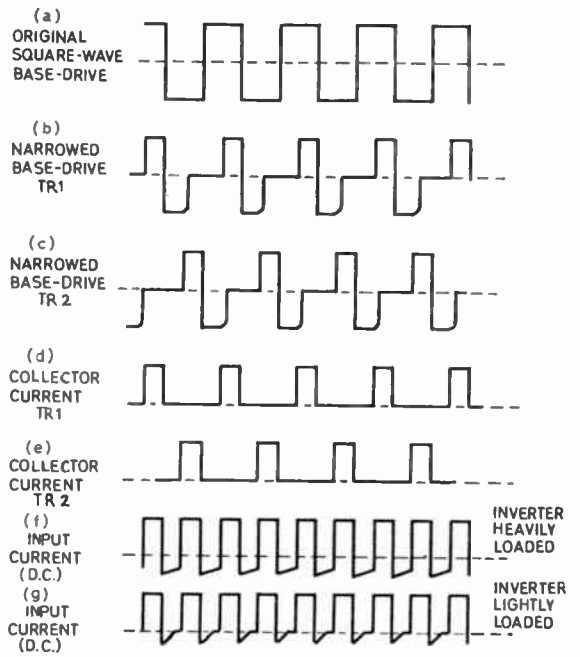


Fig. 21. Inverter current and drive waveforms.

Therefore

$$P_R = \frac{E_X^2}{E_N} \frac{I_N}{N_R} \frac{(t_x - \tau)}{t_x} \qquad \text{.....(43)}$$

Now

$$P_F = E_X I_p \frac{\tau}{t_x} = \frac{E_X^2}{E_N} I_N \frac{\tau}{t_x} \qquad \text{.....(44)}$$

From eqns. (42), (43) and (44)

$$\frac{\tau}{t_x} = \frac{E_N^2 N_R + E_X^2}{E_X^2 (N_R + 1)} \qquad \text{.....(45)}$$

and

$$P_R = \frac{I_N(E_X^2 - E_N^2)}{E_N(N_R + 1)} \qquad \text{.....(46)}$$

Equations (45) and (46) should be used only in the full power case, because of the assumptions made regarding the constancy of the returned current.

Under no-load conditions, interest centres only on τ/t_x, since the need is to know the range of control of the pulse width.

It is postulated that at E_N input voltage, the output voltage is correct, at full load and full pulse width. A value E'_N ($= E_N - IR$ drops) will therefore give the correct output voltage at no load. The ratio of the output voltages, with no voltage stabilization, will be the same as the ratio of E_X to E'_N, and in order to reduce this output voltage ratio to unity, the pulse width must be decreased.

Now since the no-load current at an input voltage E'_N is I_{NL}, then the input power is,

$$P_{in} = E'_N \cdot I_{NL}$$

At the input voltage of E_X, the no-load input current increases to:

$$I = \frac{E_X}{E'_N} I_{NL} \qquad \text{.....(47)}$$

and the input power is:

$$P_{in} = \frac{E_X^2 I_{NL}}{E'_N} \frac{\tau}{t_x}$$

These two powers must be equal (neglecting IR drops due to the increased current) and therefore

$$\frac{\tau}{t_x} = \left(\frac{E'_N}{E_X}\right)^2 \qquad \text{.....(48)}$$

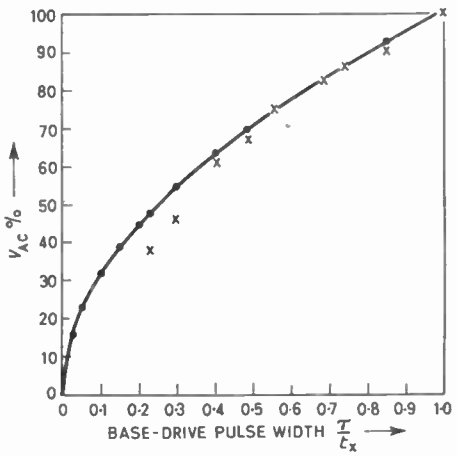


Fig. 22. Graph of output voltage against base-drive pulse width.

Equation (48) is plotted in Fig. 22 with normalized parameters together with an experimental curve. This will not match since eqn. (48) ignores the returned power, and in the experimental case P_R, although small, is finite.

20. Appendix 6

The pulse width modulator, using a self-saturating reactor, is standard circuitry, but is usually 'tailor-made' for any specific purpose.

A circuit for inverter use is given in Fig. 23, and a brief and simplified description of operation is given below, but it is recommended that the literature on the subject be studied for a detailed explanation.[17-19]

The reactor is a three-limb core, the two a.c. coils being wound on the outer limbs and the control and

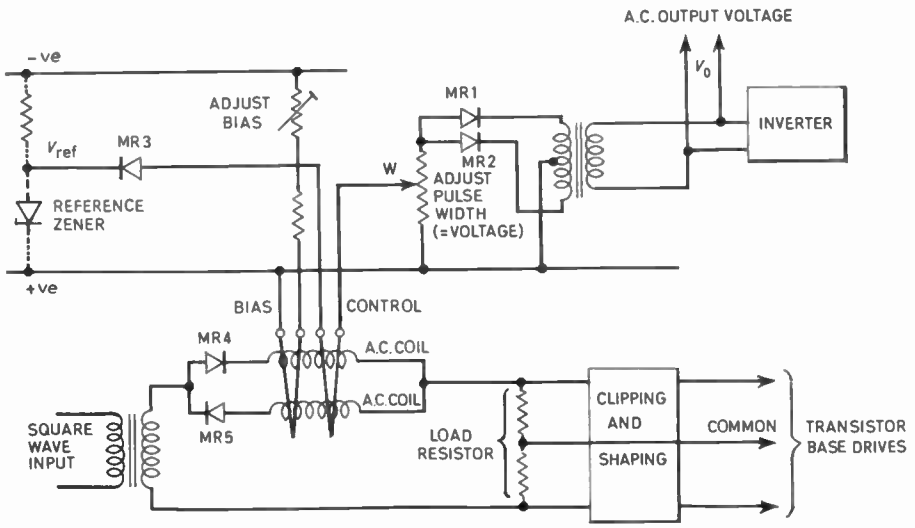


Fig. 23. HCR reactor pulse width control.