THE RADIO AND ELECTRONIC ENGINEER

The Journal of the Institution of Electronic and Radio Engineers

FOUNDED 1925 INCORPORATED BY ROYAL CHARTER 1961

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

VOLUME 28 OCTOBER 1964 NUMBER 4

THE PURPOSE OF A TECHNICAL PAPER

ONE of the most important objects of the Institution, to promote the advancement of knowledge on radio and electronic engineering, is effected primarily by publication of papers in its *Journal.* It is necessary, however, to make clear that *The Radio and Electronic Engineer*, in carrying out this task, differs from commercial technical periodicals, firstly by reason of its widely ranging coverage of the whole field of electronics, and, secondly and perhaps more important, in the varying types of paper which it publishes.

The paper on an original subject, announcing some 'break-through' in a particular corner of the field, is the type of paper most usually associated with the proceedings of any professional institution and the I.E.R.E. believes that the permanent and readily accessible recording of such work for reference by engineers throughout the world is of supreme importance. Frequently this class of paper is of *immediate* interest to only a relatively small group.

The broader interests of the main body of electronic and radio engineers are met by two further classes of paper: surveys or reviews and 'engineering development' papers. The former fulfil an important function in both fast developing and established fields and it would be a mistake to assume that their preparation calls for a degree of competence any less than the original paper. The survey of a fast developing field requires deep involvement in the forefront of research in order to assess the relative importance of different approaches, while the retrospective review of the development over the years of an important branch of engineering requires a length and breadth of experience which only the senior engineer can provide.

Engineering development papers are characterized by their valuable discussion of how several different branches of electronic engineering can be combined in a complete system. Possibly there may not be major innovations in any of these parts, but it is the final system which is unique. The preparation of this kind of paper is certainly within the compass of most project engineers, who indeed have often to prepare progress reports which, suitably modified, can form useful technical papers for the *Journal.* In order that the paper shall be of the greatest possible value to other engineers, it is essential that the depth of technical discussion is varied, that is, the novel features of a system, such as new circuits, are emphasized, if necessary to the extent of merely giving references to the literature on more conventional parts—such a paper must not be a mere 'handbook' on the system. By this approach, proper emphasis is given to the real contribution of the author, and—more important—the reader's time is saved since he need only follow up the points of particular interest, which should be referred to specifically in the abstract.

The 'engineering development' paper can be of almost any length, depending on the relative novelty or importance of the system from the point of view of the electronic engineer. In the limit, it might only describe one new circuit, mention of the application of which is made perhaps in a few sentences only. Often publication of this type of paper can have the added advantage of serving to establish priority in an invention which it may be impossible or undesirable to patent.

Underlying the purpose of all classes of paper published in the Institution's *Journal* is the basic injunction that the engineer should write a paper for the general good of his fellows. Whether or not personal benefit accrues, either directly by award of an Institution Premium or indirectly through the establishment of technical prestige, it is certain that progress in engineering can only be maintained by the open publication of advances.

F. W. S.

INSTITUTION NOTICES

The Radio Trades Examination Board

The Secretary of the Institution, Mr. G. D. Clifford, is retiring from the post of Secretary of the Radio Trades Examination Board. When Mr. Clifford joined the Institution in 1937 the Institution was one of several bodies running a Radio Servicing Examination and in 1942 he was instrumental in bringing together the interested parties who eventually formed the Radio Trades Examination Board. The Board's examinations have become the nationally recognized examinations for radio, television and electronics mechanics and technicians.

The Institution is one of four sponsoring bodies and has provided the entire administrative facilities for the Board; Mr. Clifford has been Secretary of the R.T.E.B. since it was formed.

The Radio Trades Examination Board has recently sponsored the formation of the Society of Electronic and Radio Technicians (see May issue of *The Radio and Electronic Engineer*) and Mr. A. J. Kenward is being released from his position as the Institution's Education Officer to take up the joint appointment of Secretary of R.T.E.B. and S.E.R.T. early in 1965.

Petition for Royal Charter by E.I.J.C.

The following extract from the *London Gazette* dated 4th September 1964 appeared under notices issued by the Privy Council Office on 2nd September:

> "Notice is hereby given that a Petition of the Chairman and Constituent Members of the Council of Engineering Institutions, praying for a grant of a CHARTER OF INCORPORATION, has been presented to Her Majesty in Council; and, Her Majesty having referred the said Petition to a Committee of the Lords of the Council, Notice is further given that all Petitions for or against such grant should be delivered at the Privy Council Office, on or before the 5th day of October next."

The signatories to the Petition included the President of the Institution of Electronic and Radio Engineers, Mr. J. L. Thompson.

Change of Address of the Indian Office

Members are asked to note that the address of the Administrative Office of the Institution's Indian Division is now :

7 Nandidrug Road,
Bangalore, India.

Correspondence regarding Indian Division activities as well as requests for regulations and application forms should be sent to this address.

Institution Dinner

An Institution Dinner will be held in the Lancaster Room of the Savoy Hotel on Thursday, 24th June, 1965.

Members are asked to make a note of this date; further details will be announced in a later issue of *The Radio and Electronic Engineer*.

International Conference on Microwave Behaviour of Ferri-magnetics and Plasmas

An International Conference on "The Microwave Behaviour of Ferri-magnetics and Plasmas" is to be held in London from 13th–17th September 1965. The Conference is being organized jointly by the I.E.R.E., the I.E.E., the Institute of Physics and the Physical Society, and the United Kingdom and Eire Section of the I.E.E.E.

The topics to be discussed will include:

Electro-magnetic wave propagation in ferri-magnetics and plasmas and in structures containing these media.

Microwave devices employing ferri-magnetics and plasmas.

Measurement of ferri-magnetic and plasma properties.

Non-linear phenomena in ferri-magnetics and plasmas.

Anti-ferri-magnetic phenomena and applications.

Behaviour of thin ferro-magnetic films.

Electron-beam interactions with ferri-magnetics and plasmas.

Microwave studies of solid-state plasmas.

Electromagnetic-acoustic phenomena in ferri-magnetics and plasmas.

Contributions not exceeding 2000 words are invited and synopses of approximately 500 words should be submitted to the Secretary of the Conference, c/o the I.E.E., Savoy Place, London, W.C.2, by 1st March 1965. Further information and registration forms will be available shortly and may be obtained from the Institution at 8–9 Bedford Square, London, W.C.1.

Correction

The following correction should be made to the paper "Multi-channel Open-wire Carrier Telephone System—Salisbury (Southern Rhodesia) to Kitwe (Northern Rhodesia)", published in the August issue of *The Radio and Electronic Engineer*.

Page 82, Fig. 11: The third item of the key *should read* "Special open wire pairs, 4-in spacing 0·775 dB/mile at 552 kc/s."

U.D.C. 537.56

Initiatory Cold Cathode Emission in Gas Discharges

By

Professor
F. LLEWELLYN JONES,
M.A., D.Phil., D.Sc.[†]

Reprinted, with additional material, from the Proceedings of the Symposium on "Cold Cathode Tubes and their Applications", held in Cambridge from 16th–19th March 1964.

Summary: The fundamental processes underlying the role of the cathode in the cold cathode glow discharge are discussed. Quantitative investigations of the cold field induced emission have been made and the mechanism investigated. This is the process which underlies the statistical time-lag in a cold cathode tube.

1. Introduction

The electron emission characteristics of cold metal surfaces in gases in a given electric field control a wide range of gas discharge phenomena.[1] These range from the initiation of vacuum breakdown, on the one hand, to the failure of high pressure gaseous insulation on the other. Undesired and premature breakdown in high voltage particle accelerators, triggered spark gaps, valves and switches are often related to the electron emission properties of the surfaces involved.

The cold emission characteristics are of the greatest importance in the initiation of the discharge as it is from the cathode that the necessary initiatory electrons are most likely to be produced. When no subsidiary ionization is deliberately introduced the application of an electric field to the cathode surface is the only process by which electrons can be generated.[2]

The cathode must also play an essential part in the maintenance of the discharge. The gas collisional processes active in the region known as the cathode fall constitute the agency which produces the continuous electron emission from the cathode, which is an essential part of the discharge current.

Some years ago the mechanism of regeneration was thought to be comparatively simple, resulting from the incidence of positive ions, even though some complexity was later introduced when the significance of the Auger processes was realized. Recent work, in which the particular collisional processes which lead to ionization growth have been investigated both experimentally and theoretically, has shown however that the whole process of electron emission is extremely complicated, particularly in the monatomic gases.[3]

In addition to the action of positive ions it is found that metastable atoms can play an important role not only in diffusing to the cathode but also in the generation of collision induced non-resonance radiation; this non-resonance radiation can penetrate the gas without absorption and produce extensive photoelectric emission. Furthermore, the action of scattered resonance radiation cannot be ignored.

The great complexity of the whole mechanism is due to the fact that some of these processes are pressure dependent and are also greatly dependent on the energy distribution, so that the relative importance of the various gas collisional processes which ultimately lead to cathode emission will depend upon the energy parameter E/p and the nature and state of the cathode surface.

Furthermore, the apparently simple positive ion action is itself found to be complicated. Not only will the nature of the actual ion incident be dependent on collisional processes such as charge exchange, but the action itself at the cathode is not simple.[4] There are two types of Auger interaction between the cathode and the approaching ion depending on the intensity of the electric field, which itself depends on the microgeometry of the electrode surface.

Another complexity is introduced by the presence of thin surface films. It seems likely that these can 'hold off' positive ions and allow them to accumulate and so set up a considerable surface field across a thin surface film. Such films are practically always present unless the greatest precautions are taken in the manufacture of cold cathode tubes, and it is likely that they are far from uniform and consequently can give rise to localized spots of high electron emission.[5] This leads to spots of high local current densities with characteristics akin to the abnormal cathode fall rather than the normal cathode fall. Such phenomena do not improve the stability of the glow discharge.

For reasons such as these it is of interest to be able to assess the electron emission properties of metal surfaces in gases, under moderate electric fields.

[†] Department of Physics, University College of Swansea.

2. Experimental Assessment of Electron Emission

A means of so doing is based on measurement of the statistical time lag t_s of spark initiation and a method of studying electron emission rates I in the range from a few electrons per second to about 10^6 electrons per second (10^{-13} A) has been developed at Swansea.[6]

When a voltage in excess of the static breakdown potential is applied to a spark gap, a discharge will only take place if there is a suitably placed electron present in the gas. This electron will be accelerated by the applied electric field, ionize the gas and initiate a sequence of electron avalanches. When the ionization becomes sufficiently great to discharge the capacitance formed by the plates the gap voltage collapses suddenly and the time interval between the application of the voltage impulse and the subsequent collapse of gap voltage is known as the total time lag t_T of sparking. When the formative time lag is negligible t_T may be taken to be the statistical time lag t_s. The mean $\bar{t}_s$ can be directly related to the average rate I of electron emission. The detailed analysis of the conditions necessary to make $\bar{t}_s = 1/I$ point to two methods of determining the rate of electron emission: (i) from a graph of $\ln N/n$ against t, and (ii) by determining the mean time lag t from the distribution of the N statistical time lags observed. n is the number of lags which have a duration longer than t. These two procedures are mutually independent and serve as a check on each other. Typical examples of $\ln N/n$ against t graphs (obtained in small gaps in which the volume generation of electrons is negligible) are found to be linear. This linearity provides confirmation of the initial assumptions made in the basic theory.

A fuller account of the basic theory and a discussion of the experimental conditions which must be fulfilled in the gap for application of time-lag techniques have been given previously[6] and further applications and developments have been described in later papers.[7,8] An instrument which operates on these principles has been designed for the investigation of the emission activity of metal surfaces and the design and construction has been described in detail.[9] The results obtained in this way have a bearing on many allied research problems such as those of glow to arc transitions which is greatly influenced by the state of the electrode surface; problems of electric contacts also involve the initiation of the micro-arc on closure and this also depends on the state of the electrode surface. The technique has also been used with semi-conducting materials as cathodes.[10]

In the absence of ambient gas, a field-dependent electron emission is known to occur from the cathode, and measurement of this requires a different technique. In work[11] using wide gaps the total emission was measured with an electrometer, while other work[12,13] has been carried out using Geiger-Müller counters. Both these methods for vacuum conditions have been used at Swansea, and the implications of all these measurements in the elucidation of the mechanism of the initiation of vacuum breakdown have been analysed.[14]

3. Experimental Results

3.1. *Effect of Nature of the Surface*

Electron emission from clean, smooth, cold cathodes of metals under applied electric field of the order of 10^4 V cm^{-1} was very small. However, the presence of microscopic irregularities on the cathode surface enhanced locally the applied electric field, which in turn increased the emission. The presence of a layer of fine dust or a layer of oxide produced very much the same effect causing enhanced emission at moderate field strengths. With increasing oxidation of tungsten, nickel, and iron it was found that the lags grew progressively shorter representing an increase in the electron emission which changed under these conditions from negligible amounts to values as high as 10^5–10^6 electrons per second. Removal of the metallic dust produced by the condensed vapour from the electrodes produced by this action of discharges from the oxidized surface by an air blast reduced the emission by a factor of 10^3.

An idea of the order of magnitude of the emission with different surface conditions is given by the data in the following Table, the results in which were obtained with a macroscopic applied field by 5×10^4 V cm^{-3}.

Table 1

Electron emission from different surfaces

Metal surface	Electrons emitted per second from:				
	Brass	Aluminium	Steel	Nickel	Molybdenum
Freshly turned	8×10^5	8×10^5	90×10^3	18×10^3	30×10^3
Tarnished oxide layer	5×10^5	7×10^5	30×10^3	8×10^3	7×10^3
Mechanically polished	2×10^5	7×10^5	10×10^3	6×10^3	8×10^3
Highly polished	1×10^5	7×10^5	5×10^3	6×10^3	2×10^3

It can be seen that metals are least active when they are clean and highly polished; after exposure to the atmosphere, polished metals show an increase in emission which can be an order of magnitude greater. Thus brass is more active than molybdenum which can be polished to a higher degree. These figures show that the actual state of the metal surface, rather than the metal itself, governs the electron emission activity, at least for metals of approximately the same physical properties. This result is important in the preparation of metal electrodes in attempts to prevent, for example, spark-over in high voltage equipment.

Investigations such as these enable a choice of suitable metal to be made when required for, say, electronic apparatus or high voltage equipment. The results also emphasize the care that is required in the preparation of such surfaces when they are to be subject even to moderate electric fields.

3.2. *Dependence of the Emission on Applied Electric Field*

By measuring the rate of electron emission for voltage pulses of different amplitudes a relation between the current and the applied electric field can be found. It appears that the emission mechanism can be very complicated and further experimental results are required before a complete picture of the process involved in the emission is obtained.

It has been established that there is a general correlation between the applied electric field E_m and the measured electron emission current I. An example of this field dependence is given in Fig. 1 for gold, nickel and molybdenum surfaces in nitrogen before and after outgassing treatment.

Over a limited range of electric field intensity 10^4–10^5 V cm^{-1} electron emission from the outgassed surfaces obeys a relation of the form

$$I = AE^2 \, e^{-D/E} \qquad \ldots\ldots(1)$$

This is analytically the same as the Fowler-Nordheim field emission equation from which an effective work function ϕ of the order of 0·2 eV and emitting areas of the order of 10^{-14} cm^2 are obtained. These values of ϕ and S are consistent with the view that the emission takes place from sites of low work function on the metal surface.

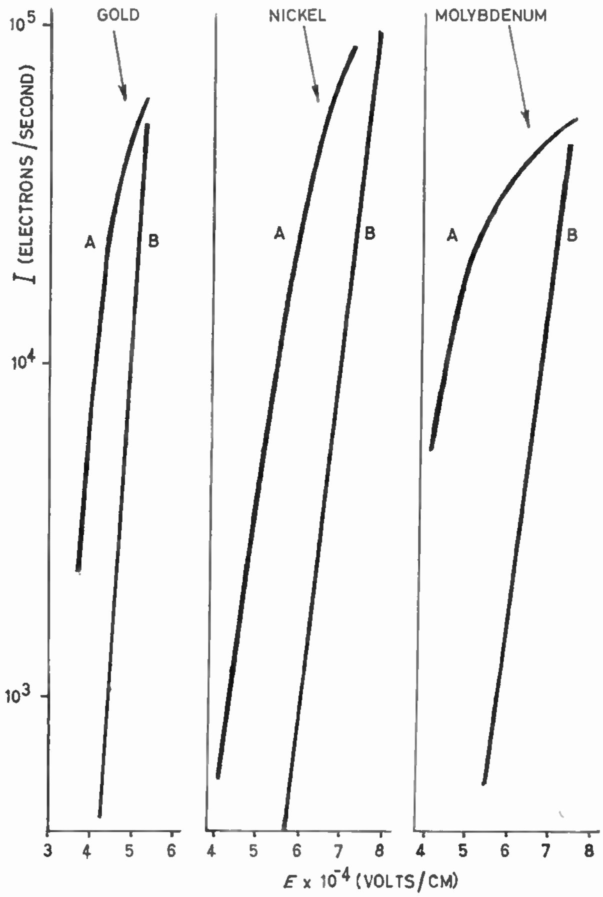
On the other hand, over the limited range of field the emission can also be satisfied by a relation of the form

$$I = B \exp (CE^{1/2}) \text{ (at constant temperature)} \ \ldots\ldots(2)$$

which is of the same analytic form as the field-assisted thermionic equation of Schottky. Comparison of this equation with experimental data yields values of $\phi = 2$ or 3 eV and $S = 10^{-8}$ cm^2. These values of ϕ and S appear at first sight to be more reasonable than those obtained using the Fowler-Nordheim theory. However, Schottky emission is strongly temperature dependent and this fact enables a distinction to be made between these two apparently possible emission mechanisms.

3.3. *Influence of Electrode Temperature*

Experiments were carried out using gold cathodes at temperatures between 80° K and 600° K. Now if the Schottky emission mechanism were responsible for the observed emission, then as the temperature was raised over this range a very large increase in I would be expected. However, only a small ($<$ 10-fold) increase was observed. It thus seems that field-assisted thermionic emission described by the Schottky equation (2) is not the process by which the measured electron emission was produced. It thus appears that a cold field process is the mechanism of the electron extraction. However, the nature of the temperature dependence requires further experimental investigation over wider temperature ranges in order to establish the precise relationship and to assess this in terms of known emission mechanisms.



A Unoutgassed B Outgassed

Fig. 1. Graphs showing variation in emission with applied electric field in nitrogen (pressure 550 mm Hg).

3.4. *Influence of Gas Atmosphere*

The influence of the ambient gas on cold field dependent cathode emission is complicated as the atmosphere can effect the processes of electron production in various ways. Firstly, the presence of water vapour or indeed any electron-attaching molecules can lead to the production of free electrons in detachment processes in an electric field. Such electrons, in whatever part of the gap they are released, generate avalanches in their movement to the anode and positive ions so produced produce cathode emission which then initiates breakdown. On the other hand, electron-attaching molecules can lead to the formation of negative ions from free electrons and in this way rapid lateral diffusion of free electrons is reduced or eliminated.

Gases also affect the electron emission as the result of their physical adsorption on the cathode surface. The more complex the gas molecules the more significant is this effect; and the inert gases for example only affect the emission in this way to a very small extent. The influence of water vapour in this way has been examined in some detail by various workers at Swansea[8, 9] who have examined emission for damp and dry air, nitrogen and hydrogen. One outstanding influence with which all experimental data agree is that in the presence of traces of water vapour the first time-lag of emission on application of the electric field is small and of the same order as that of subsequent lags. This is in marked contrast to experience with dry non-attaching gases. The time-lag before the first electron is emitted after the application of the field can be very long ($> 10^3$ times the value of subsequent time-lags).

Another process by which the nature of the gas atmosphere influences cold emission is due to the nature of the positive ions and their energy on approaching the surface. Over the range of gases which have been investigated, namely dry and damp air, dry argon, hydrogen and nitrogen, no great dependence of the order of magnitude of the cold field dependent emission has been found on the nature of the gas when pure, the dependence on the value of the electric field being over-riding.

However, it should be realized that the action of positive ions on approaching the surface will depend on the local electric field at the surface rather than on the macroscopic applied field. Thus the action of the positive ions in extracting electrons will depend to some extent on the micro-geometry of the surface. There is evidence to show that positive ions, produced by the passage of a measuring spark can influence the generation of subsequent initiatory electrons. By using reverse fields and changing pulse repetition rates it is possible to control the number of these post-breakdown positive ions, thereby changing the concentration of ions at the cathode. Experiments along these lines by means of this technique enable the influence of positive ions on the emission of electrons to be studied and the mechanism of the interaction of the positive ions with the electrode surface to be examined.

4. References

1. F. Llewellyn Jones, "Ionization and Breakdown in Gases" (Methuen, London, 1957).
2. F. Llewellyn Jones, "Electrical activity of metal surfaces", *Electrical Review*, **171**, p. 125, 27th July 1962.
3. D. K. Davies, F. Llewellyn Jones and C. G. Morgan, "Primary ionization coefficient of helium", *Proc. Phys. Soc.*, **80**, No. 516, pp. 898–908, October 1962.

   "Ionization growth times and secondary processes in helium", *ibid.*, **81**, Pt. 4, pp. 677–81, April 1963.

   "The contribution of positive ions and non-resonance radiation to secondary ionization processes in helium", *ibid.*, **83**, Pt. 1, pp. 137–44, January 1964.
4. E. Jones and F. Llewellyn Jones, "Theory of secondary emission of electrons in hydrogen", *Proc. Phys. Soc.*, **80**, Pt. 2, pp. 450–64, August 1962.
5. F. Llewellyn Jones and C. G. Morgan, "Surface films and field emission of electrons", *Proc. Roy. Soc.*, **A218**, pp. 88–103, 9th June 1953.
6. F. Llewellyn Jones and E. T. de la Perrelle, "Field emission of electrons in discharges", *Proc. Roy. Soc.*, **A216**, pp. 267–79, 22nd January 1953.
7. C. G. Morgan and D. Harcombe, "Fundamental processes of the initiation of electrical discharges", *Proc. Phys. Soc.*, **B66**, pp. 665–79, August 1963.
8. C. G. Morgan, Proc. Vth Int. Conf. on Ionization Phenomena in Gases, Venice, 1957, Vol. I, p. 434.
9. F. Llewellyn Jones and D. J. Nicholas, "The theory and design of an analyser for investigating the electron emission characteristics of surfaces in gases", *Brit. J. Appl. Phys.*, **13**, pp. 514–20, October 1962.
10. W. F. Gunn, P. G. Davies and W. MacDonald, "Field emission from semiconductors", Proc. VIth Int. Conf. on Ionization Phenomena in Gases, Paris, 1963, Vol. II, p. 383–5.
11. W. J. R. Calvert, "Pre-breakdown current and vacuum break-down", *Proc. Phys. Soc.*, **69**, Pt. 6, pp. 651–60, June 1956.
12. K. Kerner, "On electron emission from cold metallic surfaces under moderate field strengths ($\sim 10^4$ V/cm)", *Z. angew. Phys.*, **8**, No. 1, pp. 1–8, 1956.
13. C. Watts, "Investigations on the electrical activity of metal surfaces, with special reference to electron emission in applied electric fields", Ph.D. Thesis, University of Wales, 1961.
14. F. Llewellyn Jones and W. D. Owen, "Cathode electron emission and vacuum breakdown", Proc. VIth Int. Conf. on Ionization Phenomena in Gases, Paris, 1963, Vol. II, pp. 105–10.

    "Initiatory electron emission and vacuum breakdown," *Proc. Phys. Soc.*, **83**, pp. 283–91, February 1964.

Manuscript first received by the Institution on 3rd March 1964 and in final form on 20th August 1964. (Contribution No. 81.)

U.D.C. 537.56

Temporal Growth of Ionization in Hydrogen

By

C. G. MORGAN, M.Sc., Ph.D.†

AND

W. T. WILLIAMS, B.Sc.†

Reprinted from the Proceedings of the Symposium on "Cold Cathode Tubes and their Applications", held in Cambridge from 16th–19th March 1964.

Summary: This paper is an analysis of the collisional processes involving positive ions and photons, as well as primary electron ionization, which are involved in the rapid growth of ionization. This analysis is particularly relevant to the action of trigger tubes, spark gaps, thyratrons and other cold cathode devices.

1. Introduction

When a voltage V in excess of that required to cause breakdown (V_s) is applied between electrodes immersed in a gas, breakdown does not occur instantaneously. There is a time delay, t_{tot}, which may be as short as a nanosecond or as long as several minutes. This delay comprises two distinct and separable components. The first is the statistical or initiatory lag t_s which is spent awaiting the arrival of the electron which initiates the breakdown process. The second is the formative lag t_f and is the time taken for the discharge current, once initiated, to grow to some arbitrary value, for example, to become sufficiently large to give rise to space-charge distortion and cause the collapse of the applied voltage to glow or arc discharge maintenance values.

The present paper is concerned with the formative growth times t_f of ionization currents in uniform electric fields in hydrogen between plane parallel gold electrodes. The object of the investigation is to assess quantitatively, from simultaneous measurements of the temporal variation of ionization current and applied voltage, the nature and relative magnitudes of the various possible atomic, ionic and photon collisional processes which are responsible for ionization growth in gases. This knowledge may be obtained from such measurements with the aid of an appropriate theory of ionization growth.

For any electrical discharge in a gas it is possible to write a generalized continuity equation for the rate of production and loss of active particles. For example,

$$\frac{\partial n}{\partial t} = q - Rn^2 - an + D_{-}\nabla^2 n - \operatorname{div} nW_{-} \quad \ldots(1)$$

represents the rate of change of electron density in a discharge with time when the electrons are produced at a rate q and are lost by drifting with an average velocity W_{-}, by diffusing with a diffusion coefficient D_{-}, by suffering recombination with positive ions R and by attachment to electro-negative gas molecules (a). Similar continuity equations may be written for other active particles, e.g. positive ions, metastable atoms, photons, etc.[1,2,3]

1.1. *Steady-state Spatial Ionization Growth*

The continuity equation (1) may be solved subject to boundary conditions which are prescribed by the experimental conditions. In the simplest possible case, steady-state growth in a non-attaching gas at a pressure p between large plane parallel electrodes distance d apart to which a steady voltage $V \leqslant V_s$ is applied, the solution of (1), when $E/p = V/pd$ is constant, is

$$I = I_0 e^{\alpha d}/[1 - \omega/\alpha(e^{\alpha d} - 1)] \quad \ldots\ldots(2)$$

I represents the small ($\leqslant 10^{-7}$ A) steady ionization current flowing between the electrodes, I_0 is the constant externally generated initiatory electron current at the cathode, α is the Townsend primary ionization coefficient, i.e. the ionization probability per electron per unit path length parallel to the applied field, and (ω/α) is the generalized secondary ionization coefficient.[4] It represents approximately the linear sum of all secondary ionization processes simultaneously active in the particular experimental conditions considered.

When d is made sufficiently large so that the denominator of equation (2) approaches, and in the limiting value of $d = d_s$ (the sparking distance), attains the value zero, then the condition

$$1 - \omega/\alpha(e^{\alpha d} - 1) = 0 \quad \ldots\ldots(3)$$

specifies the onset of breakdown. This equation, the Townsend breakdown criterion, represents a replacement condition in which every electron regenerates itself by means of the secondary processes it initiates. The corresponding voltage, the static sparking potential V_s, is given by

$$V_s = (E/p)(p/\alpha)\ln(1 + \alpha/\omega) \quad \ldots\ldots(4)$$

When the criterion (3) is satisfied with $V = V_s$ and

† Department of Physics, University College of Swansea.

$d = d_s$ a current will flow even after I_0 is reduced to zero but will eventually die out due to statistical fluctuations in the ionization processes. If I_0 is maintained constant the current I will increase in time at a steady rate.

The values of the primary (α) and generalized secondary (ω/α) coefficients may be determined from measurements of the spatial dependence of I upon d, when $V \leqslant V_s$, keeping E/p constant. However, it is difficult to ascertain precisely the physical nature and relative importance of the many individual secondary processes which comprise (ω/α) by this means. For this purpose it is necessary to examine ionization growth in the non-steady state when $V \geqslant V$. Nevertheless, steady state measurements form a necessary precursor to such work.[5,6,7]

1.2. *Non-steady-state Temporal Growth of Ionization*

When $V \geqslant V_s$ steady-state conditions no longer prevail; the replacement condition (3) is more than satisfied so that every electron liberated at the cathode leads, by the secondary processes it initiates, to the liberation on the average of more than one cathode-emitted electron. The current in the gas therefore increases in time as well as spatially in the discharge gap and will eventually attain a value which is large enough to discharge the capacitance formed by the gap and its associated circuit elements: the gap voltage collapses. Once started the collapse can become extremely rapid. This sudden discontinuity in the gap voltage can be used to define the formative growth time t_f.

The relevant continuity equations for electrons and positive ions may be written in the form

$$\partial\left(\frac{I_-(x,t)}{W_-}\right)\Big/\partial t = -\partial I_-(x,t)/\partial x + \alpha I_-(x,t) \quad \ldots\ldots(5)$$

$$\partial\left(\frac{I_+(x,t)}{W_+}\right)\Big/\partial t = \partial I_+(x,t)/\partial x + \alpha I_-(x,t) \quad \ldots\ldots(6)$$

where $I_-(x,t)$ and $I_+(x,t)$ are the electron and positive ion currents at a distance x from the cathode at time t after growth is initiated. A complete solution with correct boundary conditions has been given by Davidson.[1,2,3] When there is no initial charge distribution in the gap, i.e. both I_0 and V applied simultaneously at $t = 0$, Davidson has shown that the exact solution of these continuity equations is given by the Laplace contour integral:

$$I_-(0,t)/I_0 = \frac{1}{2\pi i}\int_c \frac{e^{pt}\,dp}{pF(p)} \quad \ldots\ldots(7)$$

Integration is carried out around the infinite semicircular arc to the right of the imaginary axis and $F(p)$ is given by

$$F(p) = 1 - \frac{\alpha\gamma[e^{(\alpha-p/W)d}-1]}{(\alpha-p/W)} - \frac{\delta[e^{(\alpha-p/W_-)d}-1]}{\alpha-p/W_-} \quad \ldots\ldots(8)$$

with

$$(1/W) = (1/W_-)+(1/W_+)$$

when the dominant secondary processes are cathode secondary emission due to the incidence at the cathode of positive ions (γ-effect) and unscattered photons (δ/α-effect), i.e.

$$(\omega/\alpha) = (\delta/\alpha)+\gamma \quad \ldots\ldots(9)$$

Other algebraic forms of the exact solution (7) may readily be obtained from the Laplace contour integral by expanding it in a suitable way convenient for the desired purpose. The most convenient form for the present case is

$$I_-(0,t)/I_0 = \frac{1}{F(0)} + \sum_\lambda \frac{e^{\lambda t}}{\lambda[\partial F(p)/\partial p]_{p=\lambda}} \quad \ldots\ldots(10)$$

which may further be simplified by neglecting (justifiably at large t) the complex λ's and so lead to the approximate solution

$$I_-(0,t)/I_0 = 1/F(0)+B'\,e^{\lambda t}$$

where B' is the coefficient of the term involving the real value of λ in equation (10). For small overvoltages $B' \simeq 1/F(0)$ so that

$$I_-(0,t)/I_0 = \frac{1-e^{\lambda t}}{1-(\omega/\alpha)(e^{\alpha d}-1)} \quad \ldots\ldots(11)$$

and this gives the electron current at the cathode when $t \geqslant (2d/W_+)$.

This expression may be used to calculate the gap current $\bar{I}(t)$ in terms of the electron $I_-(x,t)$ and positive ion $I_+(x,t)$ using the following relations:

$$\bar{I}(t) = \frac{1}{d}\int_0^d [I_-(x,t)+I_+(x,t)]\,dx \quad \ldots\ldots(12)$$

$$I_-(x,t)\,e^{-\alpha x} = \frac{I_0(1-e^{\lambda(t-x/W_-)})}{1-(\omega/\alpha)(e^{\alpha d}-1)} \quad \ldots\ldots(13)$$

and

$$I_+(x,t) = \int_x^d \alpha I_-(0, t+x/W_+-x'/W)\,e^{\alpha x'}\,dx' \quad \ldots\ldots(14)$$

It is important to note that the gap current at any time t depends exponentially on the growth constant λ and this depends, in the present work, on the nature and magnitude of the individual secondary coefficients γ and δ/α. It is for this reason that measurements of ionization growth times can be used to provide knowledge of the individual values of γ and δ/α (and any other active secondary process in more complex cases). However, in order to do so it is necessary to relate $\bar{I}(t)$ to measurable quantities.

Two such quantities are $V(t)$, the applied voltage, and $I_e(t)$, the current in the circuit external to the gap.

The variation of gap voltage $V(t)$ with time can be computed from the relationship

$$V_0 - V(t) = \frac{1}{C_0}\int_0^t [\bar{I}(t) - I_e(t)]\,\mathrm{d}t \qquad \text{......(15)}$$

where C_0 is the gap capacitance, V_0 is the unit function gap voltage applied at $t = 0$, and

$$I_e(t) = \left[\frac{\mathrm{e}^{-B_0 t}}{RC_0}\right]\int_0^t \bar{I}(t)\,\mathrm{e}^{(B_0 t)}\,\mathrm{d}t \qquad \text{......(16)}$$

$$B_0 = \frac{1}{R}\left[\frac{1}{C} + \frac{1}{C_0}\right] \qquad \text{......(17)}$$

when the equivalent external circuit comprises a resistance R in series with a capacitance C, the whole being in parallel with the discharge gap.

Both $V(t)$ and $I_e(t)$ may be observed by suitable oscillographic or other electronic means and comparison of their measured values with those calculated using the above relations gives information about the individual secondary processes. We have previously used observations[6-9] of $V(t)$ in order to examine the role of ions and unscattered photons in hydrogen and the role of ions, metastable atoms, resonance and non-resonance photons in helium. In the present paper we extend our investigation to make simultaneous observations of $I_e(t)$ and $V(t)$.

2. Apparatus and Experimental Techniques

2.1. *Ionization Chamber, Vacuum and Gas Purification Systems*

Plane parallel gold electrodes 3·6 cm in diameter, profiled at the edges to prevent field distortion[10] were mounted centrally on quartz supports in an 8 cm diameter borosilicate glass envelope. Gold was chosen because its work function is known to remain constant (to within 0·01 eV) in hydrogen.[9] The electrodes could be separated up to a maximum distance of 1·2 cm by means of a screw mechanism driven by an external magnet. The separation was measured by means of a travelling microscope and could be set to within ±1%.

The envelope of the ionization chamber was connected to a gas purifying and ultra-high vacuum system. Initial leak testing in this was carried out using a high sensitivity mass spectrometer (A.E.I. type LD1) which when set to accept ions of mass number 4 was sensitive to 1 helium atom in 10^4 air molecules. Leak rates as low as 2×10^{-11} mmHg/s at a pressure of 10^{-5} mmHg could be detected. After faulty seals and other components, detected by this means, were remedied, a base pressure of 5×10^{-9} mmHg was achieved in the system. The pumps used were a 'Vacsorb' cryogenic sorption pump and a 'Vacion' pump in conjunction with a Bayard-Alpert ionization gauge. The usual procedures of bake-out of 250°C and outgassing for 24 hours were followed. At a pressure of 5×10^{-9} mmHg the observed rate of rise of pressure was 9×10^{-10} mmHg/min corresponding to a specific influx of 3×10^{-13} litres mmHg/s/cm^2. This is satisfactory when the large volume (5 litres) and large area (2000 cm^2) of exposed metal surface of the chamber are borne in mind and ensures a high purity of gas used.

Spectroscopically pure hydrogen supplied by the British Oxygen Company was further purified by passage over heated titanium hydride and was admitted to the ionization chamber via a palladium thimble and two vapour traps immersed in liquid air. The gas pressure was measured by using a stainless steel bellows micromanometer. In this way the pressure could be set and measured to within ±0·5%.

2.2. *Static Current and Voltage Measurements*

The experimental procedures for the measurement of static ionization currents under steady ($V < V_s$) voltages have been described in earlier papers[5-8] and were again used. The currents could be measured to ±1% in the range 10^{-13} Å to 10^{-7} Å. The initial photoelectric current I_0 was produced by a high pressure mercury vapour lamp placed external to the ionization chamber. The ultra-violet light was admitted to it through a quartz window in the envelope. The intensity of illumination could be varied by an iris. All static voltage supplies were electronically stabilized and continuously monitored against a standard cell. The total overvoltage impulse applied to the gap was constant to better than 0·1% for times up to 10^{-4} second in the absence of ionization growth. The overvoltage impulse rise time was 2×10^{-8} second. The absolute accuracy of the voltage applied to the gap was 1%, being limited by the manufacturer's tolerance of the resistances in the measuring circuit.

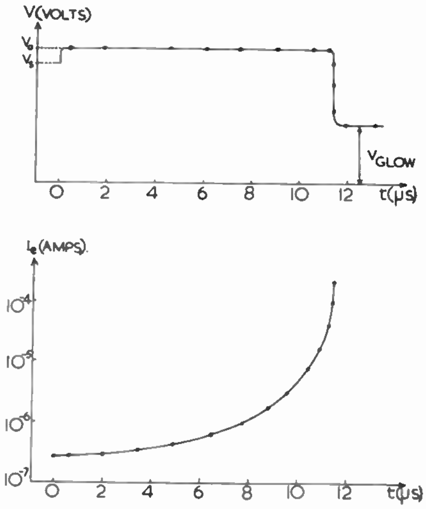
2.3. *Measurement of t_f and I_e*

The method used for the measurement of t_f was that described in references 5–8 and only the method used for measurement of I_e will be briefly described here.

A non-inductive cracked-carbon resistance was connected between the cathode of the discharge gap and earth and the voltage impulse developed across this when ionization growth occurred, was fed to a cathode follower amplifier, a travelling-wave amplifier, having a gain of 10, and finally to the probe unit of a Type 581 Tektronix oscilloscope. The time-base of this oscilloscope was triggered in synchronism with that of another high-speed oscilloscope used to display the gap voltage[8] so that both $V(t)$ and $I_e(t)$ could be recorded together. Photography was carried out using cameras incorporating Wray F/1 lenses and Ilford 5G91 film. Time-base calibration was made using crystal-controlled oscillators at frequent intervals during each experiment.

3. Results and Conclusions

Figure 1 shows a typical graph of $V(t)$ and $I_e(t)$ derived from oscillograms obtained in a preliminary experiment in which the electrode separation and gas pressure were adjusted so that the value of the primary ionization coefficient was a maximum. This condition was used in order to increase the signal/noise ratio involved in the measurement of $I_e(t)$. In this particular case $V_s/p_0 d$ was 113·5 V/cm.mmHg and $p_0 d$ = 3·51 mmHg (corrected to 0°C). The overvoltage, defined by $(V_0 - V_s)/V_s$ 100%, was 3·8%.



p_0 = 3·01 mmHg $\quad$ d = 1·167 cm

V_s = 398·5 V $\quad$ ΔV = 3·8%

Fig. 1. Temporal variation of gap voltage $V(t)$ and current in external circuit $I_e(t)$.

The applied voltage was observed to collapse rapidly 11 µs after its application and this agrees quite well with the observation that the current increases with considerable rapidity after 10·7 µs. During the initial period of current growth the rate of growth is constant and the applied voltage too remains sensibly constant, collapsing only when the sharp increase in current occurs. The discontinuities occurring at approximately the same instant in these curves indicate the onset of a space-charge-controlled glow discharge. It follows that the early (and slowest) stages of ionization growth in the present work occurs in the absence of significant space charge distortion. Even so the initial rate of growth of current is large.

From the initial slope of such $I_e(t)$, t curves the growth constant λ may be evaluated and, with the above analysis, used to estimate the relative proportion of secondary emission due to the γ and δ/α processes. Table 1 gives data obtained for two values of the energy parameter E_0/p_0 showing the trend of variation of γ and δ/α. Figure 2 shows the dependence of λ on overvoltage.

Table 1

Gas	Surface	E_0/p_0 (V/cm. mmHg)	$P_0 d$ mmHg × cm	ΔV%	%γ	%δ/α	λ seconds^{-1}
H_2	Au	110·5	4·04	0·61	51·7	48·3	$1·5 \times 10^5$
H_2	Au	110·9	4·04	1·33	49·6	50·4	$3·5 \times 10^5$

The observed constant initial rate of growth in an undistorted field is in agreement with the theory and shows that the initial stages of the observed temporal

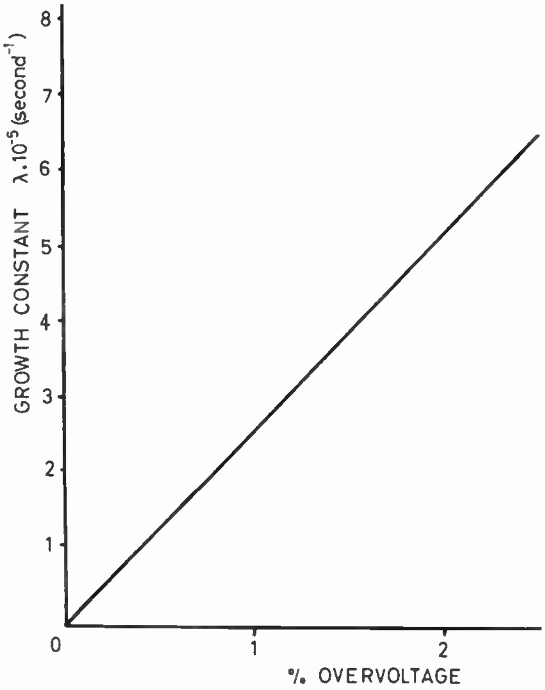


Fig. 2. Dependence of λ on overvoltage using hydrogen and gold electrodes. E_s/p_0 = 110 V/cm.mmHg.

growth of ionization in the non-steady state when $V > V_s$ is produced by the same collisional processes which operate in the steady state when $V < V_s$.

When the rate of growth (and hence the growth constant itself) becomes a function of time, due to the onset of space-charge which leads to a spatio-temporal dependence of (α) and (ω/α) the analysis given above requires modification since implicit in it is the assumption of ionization growth in uniform fields. The way in which this modification can be carried out by allowing for space-charge distortion in order to assess quantitatively the later stages of growth has been given by Davies, Evans and Llewellyn Jones.[11]

4. Acknowledgments

We wish to thank Professor F. Llewellyn Jones for his helpful advice and encouragement during this work. We also wish to thank Mr. M. H. Davies for his help in carrying out design and constructional work and for undertaking some of the experiments and computations involved.

5. References

1. P. M. Davidson, "Formative time lags in the electrical breakdown of gases" (mathematical appendix), *Brit. J. Appl. Phys.*, 4, p. 170, June 1963.
2. P. M. Davidson, "Theory of the temporal growth of ionization in gases, involving the action of metastable atoms and trapped radiation", *Proc. Roy. Soc.*, **A249**, pp. 237–47, 1st January 1959.
3. P. M. Davidson, "Theory of temporal growth of ionization between parallel plates in the inert gases", *Proc. Phys. Soc.*, **80**, Pt. 1, pp. 143–50, July 1962.
4. F. Llewellyn Jones, "Ionization and Breakdown in Gases" (Methuen, London, 1957).
5. D. K. Davies, F. Llewellyn Jones and C. G. Morgan, "Primary ionization coefficient of helium", *Proc. Phys. Soc.*, **80**, pp. 898–908, October 1962.
6. D. K. Davies, F. Llewellyn Jones and C. G. Morgan, "Ionization growth times and secondary processes in helium", *Proc. Phys. Soc.*, **81**, Pt. 4, pp. 677–81, April 1963.
7. D. K. Davies, F. Llewellyn Jones and C. G. Morgan, "The contribution of positive ions and non-resonance radiation to secondary ionization processes in helium", *Proc. Phys. Soc.*, **83**, Pt. 1, pp. 137–44, January 1964.
8. C. G. Morgan, "Temporal growth of ionization in gases", *Phys. Rev.*, **104**, No. 3, pp. 566–71, 1st November 1956.
9. F. Llewellyn Jones and E. Jones, "Experimental determination of the individual secondary ionization coefficients in hydrogen and the dependence on cathode work function", *Proc. Phys. Soc.*, **75**, Pt. 5, pp. 765–71, May 1960.
10. F. M. Bruce, "Calibration of uniform-field spark-gaps for high-voltage measurement at power frequencies", *J. Instn Elect. Engrs*, **94**, Pt. II, pp. 138–48, April 1947.
11. A. J. Davies, C. J. Evans and F. Llewellyn Jones, "Electrical breakdown: the spatio-temporal growth of ionization in fields distorted by space charge", *Proc. Roy. Soc.* **A281, pp.** 164–83, 1964.

Manuscript first received by the Institution on 3rd March 1964, and in final form on 31st July 1964. (Paper No. 932.)

DISCUSSION

Mr. G. F. Weston: Glow discharge tubes are normally filled with inert gas at pressures above 20 torr, and up to now it has been assumed that at these pressures the secondary Townsend mechanism, ω/α, was mainly due to the ejection of electrons from the cathode as a result of bombardment by positive ions and/or metastable atoms. The authors' results showing that non-resonant radiation from the metastable atoms accounts for over 80% of the ω/α effect at these pressures is therefore of great interest.

In studying the secondary mechanism of discharges with composite cathodes such as Au–Cs–O, Hall and Stocker† have found measurable field emission at 10^4 V/cm which is dependent on surface charge. They postulate that there is a semiconducting layer on their substrate and that the behaviour of the cathode in a glow discharge is due to field emission and field enhanced secondary emission due to surface charging by the positive ions. This latter process might be expected to produce a rather longer time constant than a classic γ effect due to positive ions.

Have the authors considered such a possibility? It would, of course, only apply if the cathode were covered or had patches of a monolayer or more of impurities.

Dr. Morgan (*in reply*): Some years ago we examined in detail enhanced electron emission from cold cathodes in gases (Llewellyn Jones and Morgan,‡ Morgan and Harcombe§). This work showed the important role played

† R. F. Hall and B. J. Stocker, "An Inert-gas Glow Discharge with Low Maintaining Potential (22–35 V)", Proceedings of the 6th International Conference on Ionization Phenomena in Gases, Paris 1963. Also *The Radio and Electronic Engineer*, 28, No. 5, November 1964 (to be published).

‡ F. Llewellyn Jones and C. G. Morgan, "Failure of Paschen's law and spark mechanism at high pressure", *Phys. Rev.*, **82**, p. 970, 1951 (Letter).

F. Llewellyn Jones and C. G. Morgan, "Surface films and field emission of electrons", *Proc. Roy. Soc.*, **A218**, p. 88, 1953.

§ C. G. Morgan and D. Harcombe, "Fundamental processes of the initiation of electrical discharges", *Proc. Phys. Soc.*, **B66**, p. 665, 1953.

by positive ions and surface contamination in the electron ejection mechanism. The work was carried out with a variety of gases and with cathodes of various metals and semiconductors in a number of surface states. The study disclosed that enhanced electron emission ($\sim 10^3$ to 10^6 electrons/second) can readily be obtained with macroscopic applied electric fields as low as 10^4 V/cm and could act as a γ process.

In the investigation described in the present paper, however, the experiments were carried out using ultra-high-vacuum techniques and carefully purified gases. Gold and silver cathodes were employed in systems which were initially evacuated to 10^{-9} torr and which had specific influxes of only 10^{-13} litre torr/second/cm^2. It is therefore unlikely that there was significant gas or surface contamination present.

Mr. W. J. Saysell: Has measurement been made of the additional time delay for the transition from glow to arc discharge?

Dr. Morgan (*in reply*): In the present work the current was deliberately restricted in order to avoid damage to the cathode. It was never permitted to increase beyond 1 mA. However, the oscillograms of interelectrode voltage and current in the external circuit show that it is the early stages of current growth, from 10^{-10} to 10^{-5} A, say, which occur in undistorted fields which are the slowest; the glow develops rapidly once the current density is large enough to cause significant space-charge distortion; voltage collapse then occurs in a time of less than 1 μsec. If permitted, the current would develop to arc values in times of this order. On general grounds the transition time would be expected to be a function of the ratio of electric field E to gas pressure p, (E/p), occurring most quickly when p is large, since this parameter controls the current density.

STANDARD FREQUENCY TRANSMISSIONS

(Communication from the National Physical Laboratory)

Deviations, in parts in 10^{10}, from nominal frequency for **September 1964**

September 1964	GBR 16kc/s 24-hour mean centred on 0300 U.T.	MSF 60 kc/s 1430–1530 U.T.	Droitwich 200 kc/s 1000–1100 U.T.	September 1964	GBR 16 kc/s 24-hour mean centred on 0300 U.T.	MSF 60 kc/s 1430–1530 U.T.	Droitwich 200 kc/s 1000–1100 U.T.
1	− 150·9	− 151·6	− 8	16	− 149·4	− 150·4	− 2
2	− 149·6	− 149·7	− 8	17	− 150·0	− 149·9	− 2
3	− 149·8	—	− 5	18	− 150·5	− 150·2	0
4	− 150·4	− 150·0	− 4	19	− 150·2	—	+ 1
5	− 149·8	− 149·6	− 5	20	− 150·3	− 149·8	+ 1
6	− 150·1	− 150·6	− 11	21	− 150·6	—	+ 1
7	− 150·2	− 151·8	− 12	22	− 150·3	− 151·2	+ 3
8	− 151·1	− 150·3	− 8	23	− 151·1	− 151·5	+ 3
9	− 149·9	− 150·4	− 8	24	− 151·4	− 151·4	+ 2
10	− 150·6	− 151·7	− 8	25	− 151·4	− 151·5	+ 3
11	—	− 151·1	− 8	26	− 150·3	− 149·5	+ 3
12	− 150·3	− 149·1	− 7	27	− 150·1	− 150·1	+ 5
13	− 148·9	− 148·4	− 5	28	− 150·2	− 150·7	+ 5
14	− 149·5	− 150·3	− 4	29	− 149·9	− 150·3	+ 6
15	− 150·9	− 149·7	− 3	30	− 150·3	− 150·3	+ 7

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

Note: the phase of the GBR/MSF time signals was retarded by 95·5 milliseconds at 0000 UT on 1st September, 1964.

U.D.C. 621.387.3

Statistical and Formative Time Lags in Cold Cathode Tubes

By

D. W. E. FULLER, B.Sc.†

Reprinted from the Proceedings of the Symposium on "Cold Cathode Tubes and their Applications" held in Cambridge from 16th–19th March 1964.

Summary: The influence of pulse amplitude, pulse rise-time, pulse repetition rate, tube geometry, cathode material and gas filling are discussed. Methods of priming are compared.

1. Introduction

Gaseous breakdown is a wide ranging subject and cold cathode tubes designed to operate in the glow discharge region constitute a relatively narrow field. This paper discusses those aspects which are relevant to the statistical and formative time lags of cold cathode tubes.

The mechanisms involved in gaseous breakdown have been discussed in numerous papers and an appropriate text book[1–4] should be consulted for detailed discussion and references.

2. Definitions

2.1. *Formative Time Lag T_f*

For breakdown to occur a suitable electron must appear which initiates an unbroken series of avalanches. We consider that breakdown has occurred when the current has built up to some arbitrary value, e.g. 10^{-7} A cm^{-2}. The time between the appearance of this electron and the current reaching this value is called the formative time lag. For any individual tube this lag has a constant value, for given circuit conditions.

2.2. *Statistical Time Lag T_s*

The time between the application of the potential difference to the tube and the appearance of the electron which initiates breakdown is known as the statistical lag.

Not every electron is successful in producing breakdown, unless the applied potential difference is considerably higher than the static breakdown potential V_s. Also the electrons will appear at random intervals. For any given tube and circuit this lag will have a random scatter.

2.3. *Static Breakdown Potential V_s*

A tube will break down in a finite time if the applied potential V is greater than the static breakdown potential V_s. It will not ever break down at a potential below V_s. The static breakdown potential is thus the limiting potential at which breakdown will occur after an infinitely long time.

3. Theory

3.1. *Statistical Time Lag T_s*

The electron multiplication and secondary emission processes leading to a breakdown are statistical in nature and the problem must be discussed in terms of probabilities.

The probability P of an electron producing breakdown will depend upon the position at which it is liberated, the gas and the potential difference applied to the tube.

For tubes with plane parallel or coaxial electrodes P can be calculated from the relation derived by Wijsmann[5],

$$P = \frac{1 - \frac{1}{q}}{1 - \frac{1}{q} + \frac{1}{q_x}}$$

where q is the average number of secondary electrons liberated at the cathode per initial electron starting from the cathode, and q_x is the corresponding quantity where the initial electron is liberated at a point x in the gas.

In the case where all electrons are emitted from the cathode

$$P = 1 - \frac{1}{q}$$

P is clearly larger in this case and these electrons are therefore the most efficient in producing breakdown.

In the case of an ideal diode with large plane parallel electrodes

$$q = \gamma[\exp(\eta V) - 1]$$

where γ is the probability that a secondary electron is liberated per positive ion produced in the gas,
η is the ionization co-efficient per unit potential difference,
V is the potential difference applied to the tube.

† Hivac Ltd., South Ruislip, Middlesex.

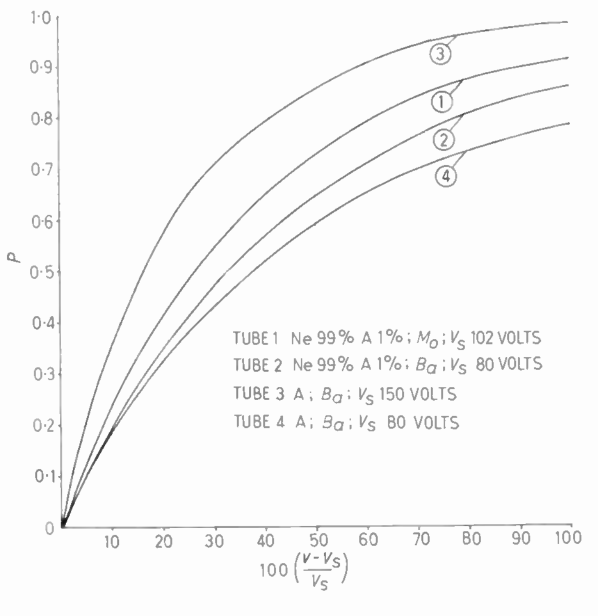


Fig. 1. Probability P of an electron producing breakdown against percentage overvoltage $100\left(\frac{V-V_s}{V_s}\right)$.

Thus $$P = 1 - \frac{1}{\gamma[\exp(\eta V)-1]}$$

γ can be calculated with sufficient accuracy by putting q equal to one when V equals the static breakdown potential of the tube V_s.

All graphs of P as a function of the percentage overvoltage $100(V-V_s)/V_s$ will start at the origin and converge asymptotically to the line $P = 1$.

Various tubes will give curves that are members of this family, differing in shape from each other. It will be seen that the shape of the curve is determined by the rate of change of η with respect to the reduced field E/p over the relevant range.

Figure 1 gives P against $100(V-V_s)/V_s$ for typical tubes.

Tubes 1 and 2 are designed to have their static breakdown potential V_s at its minimum value $V_{s(\min)}$. They are filled with a Penning mixture (neon 99%, argon 1%) which has a very slow decrease of η with increasing field in this region. The resulting curves are effectively those for the case where η is a constant.

Tube 3 is designed to have V_s greater than $V_{s(\min)}$ and is filled with a pure rare gas (argon). In this case η increases with overvoltage and a steeper curve is obtained.

Tube 4 is designed to have V_s at its minimum value but is filled with a pure rare gas (argon). This has a well defined maximum value of η. A rapid decrease of η with overvoltage occurs and a more gradual increase of P is therefore obtained.

If a tube is to operate at the minimum breakdown $V_{s(\min)}$ then a Penning mixture will give a smaller change of η than a pure gas. The situation is improved if the tube can operate above $V_{s(\min)}$ using a pure gas, thus taking advantage of the rapid increase of η with E/p at low values of E/p.

If N primary electrons leave the cathode per second then the probability of a statistical time lag exceeding t is $\exp(-PNt)$ and the mean statistical time lag $\bar{T}_s$ is $1/PN$. It is important to note that this expression applies only to the case where no ions are already present in the gap.[6]

Where the overvoltage is sufficiently high, $P \rightarrow 1$ and $\bar{T}_s \rightarrow 1/N$ giving the mean statistical lag of the tube when used under optimum conditions. If we denote this lag by $\bar{T}_{s(\min)}$ then

$$\bar{T}_s = \bar{T}_{s(\min)}.(1/P)$$

The factor $\bar{T}_{s(\min)}$ is determined by the design of the tube while $(1/P)$ may be regarded as an overvoltage coefficient of mean statistical lag and is a function of the circuit design.

Curves of $1/P$ against overvoltage are given in Fig. 2. It will be seen that the use of small overvoltages may require $\bar{T}_{s(\min)}$ to be multiplied by factors of 2 to 10.

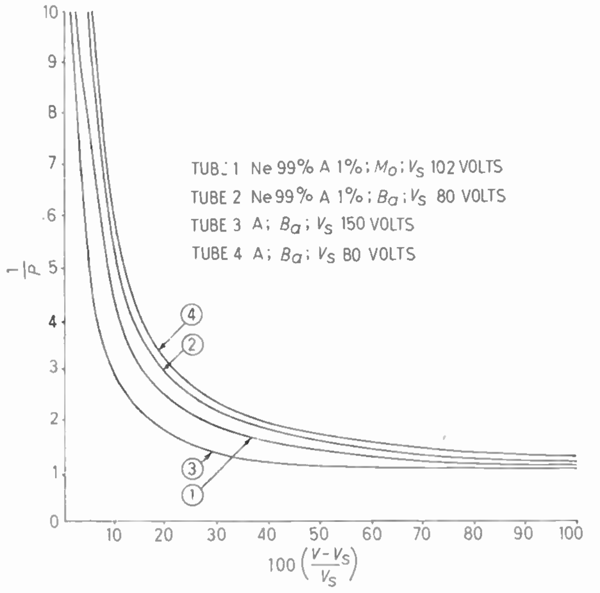


Fig. 2. Overvoltage coefficient of statistical lag, $1/P$, against percentage overvoltage $100\left(\frac{V-V_s}{V_s}\right)$.

In Fig. 3 line A shows the relation between the probability of failure to strike and the duration of the applied pulse for a (hypothetical) tube with no formative lag.

In order to obtain a probability of failure of about 1 in 10^4 it is necessary to use a pulse width of $10\bar{T}_s$. Operation at an overvoltage such that $P = 0{\cdot}5$ would give a probability of failure of about 1 in 100.

Let $\tau_{i(s)}$ be the value of τ_i at the reduced field corresponding to static breakdown.

Then $$\tau_i = \tau_{i(s)} \cdot \frac{V_s}{V}$$

and $$T_f = \tau_{i(s)} \cdot \frac{V_s}{V} \cdot \frac{1}{\varepsilon} \ln(1+\varepsilon i_e/i_0)$$

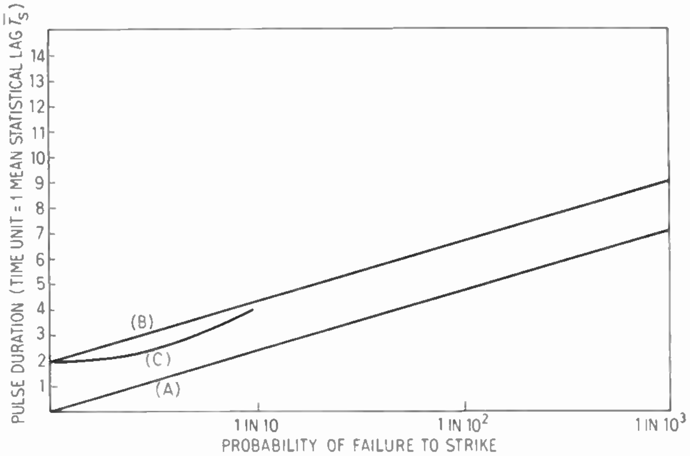


Fig. 3. Relation between pulse duration and probability of failure to strike.

3.2. *Formative Time Lag T_f*

The most accurate calculation of formative time lag is based upon expressions given by Davidson for spatio-temporal growth of an ionization current.[7,8] A knowledge of the relative importance of secondary electron emission due to positive ion bombardment as opposed to photo-electric emission is required.

A rapid approximation valid for the case where secondary electrons are due only to positive ion bombardment has been given by Schade.[9] This mechanism gives longer lags than photo-electric emission and any estimates will err on the pessimistic side.

We have

formative time lag $$T_f = \frac{\tau_i}{\varepsilon} . \ln(1+\varepsilon i_e/i_0)$$

where τ_i is the mean transit time of a positive ion, moving from anode to cathode in reduced field E/p

$$\varepsilon = q-1 = \gamma[\exp(\eta V)-1]-1$$

these symbols being defined earlier.

In the case of a plane parallel diode we have

$$\tau_i = \frac{d}{K_i} \cdot (p/E)$$

where d is the gap between the electrodes,
K_i is the mobility of the ion at 760 torr.

In the case of a commercial tube the value of i_0 may be estimated by assuming that the tube has an acceptable statistical lag. A mean statistical lag of 100 µs implies the liberation of 10^4 electrons per second, thus i_0 is about 10^{-15} amperes.

Taking $i_e = 10^{-6}$ amperes as the criterion of breakdown, then i_e/i_0 is about 10^9 or e^{20}, also $\varepsilon . e^{20} \gg 1$

$$T_f = \tau_{i(s)} \cdot \frac{V_s}{V\varepsilon}(\ln\varepsilon + 20)$$

T_f is insensitive to the ratio i_e/i_0 and the rate of production of primary electrons will normally be unimportant. It will be seen that T_f is given by the product of $\tau_{i(s)}$, dictated by the tube design, and a function of the percentage overvoltage, determined mainly by the circuit conditions.

As many tubes are designed to break down when η is near its maximum value it is of interest to compare $\tau_{i(s)}$ for a 1 mm gap for a rare gas ion moving in its own gas.

Table 1 shows that the differences in mobility K_i are partially nullified by the different values of reduced field. If a tube is to operate at its minimum breakdown potential then changing to another rare gas will not change $\tau_{i(s)}$ to any great extent. A design giving static

Table 1

Comparison of transit times τ_i of rare gas ions at reduced field E/p corresponding to maximum ionization coefficient η

Ion	He_2^+	He^+	Ne_2^+	Ne^+	A_2^+	A^+	Kr_2^+	Kr^+	Xe_2^+	Xe^+
K_i cm^2 V^{-1} s^{-1}	20·0	11·2	6·5	4·2	2·6	1·65	1·2	0·93	0·80	0·63
E/p (for max η) V cm^{-1} torr^{-1}	50	50	100	100	200	200	200	200	300	300
τ_i (for 1 mm gap) μ s	0·13	0·23	0·20	0·31	0·25	0·40	0·55	0·71	0·55	0·70

breakdown well above its minimum value means smaller E/p values and longer lags.

The percentage overvoltage has a dramatic influence on the formative time lag as shown in Fig. 4 and a 10% increase of overvoltage may reduce the formative time lag by a factor of 10. The curve should only be taken as an indication of the way in which T_f will increase if a small overvoltage is used. The expression for T_f is not valid when the overvoltage is high enough to cause space charge effects and values of T_f approaching $\tau_{i(s)}$ must not be estimated in this way.

Very short formative time lags may be obtained when an appreciable fraction of the secondary electrons arise from the incidence of photons upon the cathode. As the relevant transit time is now that of an electron the expression for T_f becomes

$$T_f = \tau_{e(s)} \cdot \frac{V_s}{V}\frac{1}{\varepsilon}(\ln \varepsilon + 20)$$

where $\tau_{e(s)}$ is the mean electron transit time from cathode to anode for the reduced field appropriate to static breakdown.

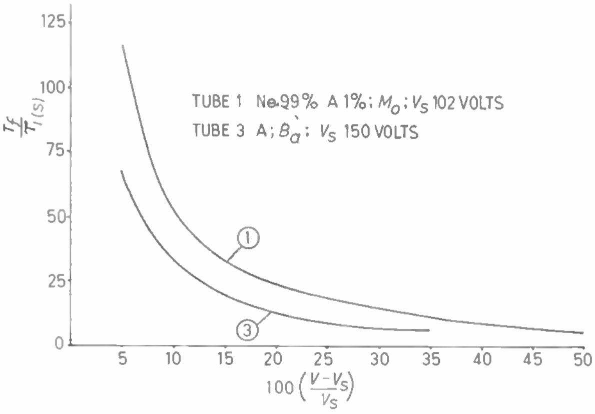


Fig. 4. Graph of $\frac{T_f}{\tau_{i(s)}}$ against percentage voltage $100\left(\frac{V-V_s}{V_s}\right)$ where T_f is the formative time lag and $\tau_{i(s)}$ is the positive ion transit time at static breakdown V_s.

The importance of the photo-electric mechanism will depend upon the probability of raising a gas molecule to an excited state, such that a photon of light is emitted, as compared to the probability of ionizing the molecule.

Work on hydrogen[10] indicates that at values of reduced field E/p of about 50 V cm^{-1} torr^{-1} the photo-electric effect contributes about 75% of the secondary electrons. The effect is less important at higher values of E/p. It is to be expected that gas mixtures including hydrogen will give shorter formative time lags than are given by gases such as neon and helium which show no perceptible photo-effect.[11]

The attempts to obtain these data have shown the necessity for extremely pure gases and clean surfaces when observing phenomena of this nature. It is unlikely that results obtained on one design of commercial tube can be used to predict accurately results on another design.

4. Methods of Priming

4.1. *Illumination Priming*

Priming electrons may be provided by various means. The simplest method is by allowing light to fall upon a low work-function cathode which then emits photo-electrons. A barium cathode emits about 10^7 electrons per second for an illumination of one foot candle, corresponding to a mean statistical lag of about 0·1 μs.

In order to ensure adequate illumination at all times it is necessary to include a light source as part of the equipment. If a small single light source is used then the light will be highly directional and the orientation of the tube when mounted will affect the illumination of the cathode. Differences in performance can arise which would not appear when the tube was used in a 'bread-board' circuit under diffuse lighting.

Where stability is important then a cathode of a high work-function metal (e.g. molybdenum) is indicated. The glass envelope will cut off the light with wavelength less than 3000 Å and the photons striking the cathode will have insufficient energy to release photo-electrons from the cathode. The heavy sputtering required for stability reasons would make the illumination of the cathode even more dependent on tube orientation.

A heavily-sputtered tube with a pure metal cathode will normally rely upon some other means of priming. The statistical lag will still vary slightly with illumination, presumably due to traces of sodium or similar impurities. Tests should therefore be carried out either in darkness or at the lowest level of illumination under which the circuit will operate.

4.2. *Radioactive Priming*

The primary electrons may be provided by including a radioactive material in the tube. This may be either a gas or a solid and the choice is mainly one of practical manufacturing expediency. The previous theory has shown that electrons produced near the cathode are the most efficient. A solid source, e.g. Ni^{63}, would therefore be located on or near the cathode.

The use of tritium gas minimizes the health hazard in manufacture but introduces technical problems. An unknown fraction of the tritium gas is trapped during the sputtering process, which is necessary for the production of a stable tube. Any tritium deposited on the walls of the tube will have greatly reduced efficiency due to beta-particle absorption by overlying layers of sputtered material.

Tritium is a beta emitter, the electrons having a maximum intensity at about 2·5 keV.[12] The fast electrons will produce positive ions and slow electrons in the gap. Since slow electrons produced near the cathode are more effective in producing breakdown than those produced near the anode it is to be expected that an increase in cathode area will reduce the statistical lag. Tube geometry will determine the fraction of beta particles that enter the useful volume and high pressures will increase the number of slow electrons formed. Precise relationships are of little value because of the clean-up by sputtering. In a typical tube the inclusion of 10 μC of tritium is equivalent to about 10^4 priming electrons per second. Very small values of mean statistical lag are thus difficult to achieve and this technique is appropriate only where conditions of use prohibit priming by an auxiliary discharge or by external illumination.

4.3. *Priming Discharge*

Where it is not possible to ensure adequate illumination of the tube then an auxiliary discharge can act as a source of ions and photons. Upon application of a pulse to the main anode some ions will move into the main cathode-anode gap. The number which moves into the main gap will depend upon both the potential of the main anode, the magnitude of the priming discharge current and the proximity of the discharge to the main gap.

If the priming discharge occurs too close to the main gap then space charge effects will reduce the main gap breakdown potential. As a current of 10^{-6} ampere is equivalent to 10^{13} electrons per second, it is clear that only a small fraction need appear in the main gap in order to reduce the statistical lag to an extremely small value. It is therefore possible to arrange the tube geometry so that the priming discharge is slightly away from the main gap.

This method of priming can easily produce much smaller statistical lags than are obtainable with radioactive materials, is applicable to heavily sputtered high stability tubes and does not require the illumination to be controlled. Wherever the circuit permits an auxiliary discharge then this is the easiest way of obtaining vanishingly small statistical lags independent of illumination.

5. Influence of Pulse Parameters

5.1. *Pulse Duration*

Curve B of Fig. 3 shows the relation between pulse duration and probability of failure to strike for a tube which has a mean statistical lag $\overline{T}_s$ and formative lag T_f. For a failure rate of about 1 in 10^4 the required duration is $T_f + 10\overline{T}_s$. If a smaller failure rate of about 1 in 10^8 is required then the pulse duration must be increased to $T_f + 20\overline{T}_s$. If T_f and $\overline{T}_s$ are of the same order then a reduction of $\overline{T}_s$ will bring the greater benefit. Only when $\overline{T}_s$ can be reduced to a very small value (e.g. by a priming discharge) does reduction of T_f become important.

Curve C shows some results obtained with a tritiated tube.[13] When no potential difference exists across the tube then the ions and slow electrons formed by the tritium will recombine or will diffuse slowly to the tube wall and the electrodes. The positive ion concentration will build up to an equilibrium value in between pulses.

When the pulse is applied the ions will move to the cathode, thus increasing the probability of breakdown. This effect is negligible for pulses longer than 100 $\tau_{i(s)}$. If the tube has a bias applied in between pulses then the ions are cleared from the gap and the conventional straight line is obtained.

Measurements taken with a short pulse and no bias will thus have some error. Extrapolation to the region of $T_f + 20\overline{T}_s$ in order to obtain the failure rate under practical circuit conditions would produce a most inaccurate estimate.

5.2. *Pulse Repetition Rate*

The measurement of statistical lag necessarily involves a large number of measurements. Each discharge will produce an after-effect such as ions and metastables in the gap or ions on the cathode. The result should therefore be qualified by specifying the pulse repetition rate, clearing bias, discharge current and duration of current pulse. Deionization measurements are concerned with the depression of breakdown potential due to the space charge arising from ions left in the gap. Statistical lag is affected by the presence of a very small quantity of primary electrons and the after-effects will last long after the normal deionization time.

Paetow[14] gave results for nickel plates after passing a discharge of 1 μA for 1 second. The number of electrons emitted by the cathode fell to 100 per second in about 13 seconds. Llewellyn-Jones and Morgan[15] have shown that positive ions rest on a thin tarnish film on the cathode setting up a strong electric field which extracts electrons from the cathode.

The rate at which the enhanced electron emission will decay will depend upon the rate of leakage of the ions from the cathode surface and cannot be predicted for any proposed design of tube. It is therefore necessary to measure the statistical lag of a new design for various pulse repetition frequencies in order that after-discharge effects can be detected. These effects become less serious as the deliberate priming of the tube is increased.

In the extreme case of a tube with a sputtered molybdenum cathode with no means of priming then these effects can be detected for about 10 hours. A reliable measurement can only be taken after 24 hours in darkness. Commercial tubes usually have sufficient priming to permit measurements being made with 1 pulse per second.

5.3. *Pulse Rise-time*

If the time that the pulse takes to reach the static breakdown potential V_s is much longer than the ion transit time then any ions in the gap will be swept to the cathode before the field is high enough to produce breakdown. A slow-rising pulse will therefore produce the same results as a clearing bias.

5.4. *Rising Potential*

If the applied potential V rises slowly until breakdown occurs then the conditions are much more complex than when the overvoltage is applied rapidly and remains constant. This case has been discussed by Heymann.[16]

From Fig. 1 it will be seen that probability P of an electron producing breakdown is zero when V passes through the value V_s and then increases non-linearly with V.

Considering a tube with mean statistical lag $\overline{T}_{s(\min)}$ and no formative lag.

For small overvoltage only

$$P_{\text{mean}} \simeq \tfrac{1}{2}(\mathrm{d}P/\mathrm{d}t)T_s$$
$$\simeq \tfrac{1}{2}(\mathrm{d}P/\mathrm{d}V)(\mathrm{d}V/\mathrm{d}t)T_s$$

where the values of $(\mathrm{d}P/\mathrm{d}V)$ and $(\mathrm{d}V/\mathrm{d}t)$ at $V = V_s$ are to be used. The probability of a lag exceeding T_s

$$= \exp[-PNT_s]$$
$$= \exp[-(\mathrm{d}P/\mathrm{d}V)(\mathrm{d}V/\mathrm{d}t)T_s^2/2\overline{T}_{s(\min)}]$$

The distribution of lags will follow a different law from the case of constant overvoltage.

As a practical example, a probability of failure of 1 in e^{18} (about 1 in 10^8) corresponds to a lag of

$$T_s = 6\left[\frac{\overline{T}_{s(\min)}}{(\mathrm{d}P/\mathrm{d}V)(\mathrm{d}V/\mathrm{d}t)}\right]^{\frac{1}{2}}$$

The overvoltage at breakdown $(\Delta V)_b$ in this case is

$$(\Delta V)_b = T_s.(\mathrm{d}V/\mathrm{d}t)$$
$$= 6\left[\frac{\overline{T}_{s(\min)}(\mathrm{d}V/\mathrm{d}t)}{(\mathrm{d}P/\mathrm{d}V)}\right]^{\frac{1}{2}}$$

From Fig. 1 it will be seen that $(\mathrm{d}P/\mathrm{d}V)$ will not differ greatly between various designs of tube. The lag and overvoltage will vary with the rate of rise of potential according to the relations.

$$T_s \propto [\mathrm{d}V/\mathrm{d}t]^{-\frac{1}{2}}$$
$$(\Delta V)_b \propto [\mathrm{d}V/\mathrm{d}t]^{\frac{1}{2}}$$

Where the formative lag is small relative to the statistical lag (corresponding to the required probability level) then the formative lag can be estimated by assuming that the overvoltage is constant at $(\Delta V)_b$ as previously calculated. This assumption gives a high value for the formative lag.

Where the statistical lag is small relative to the formative lag then the mean value of ε can be taken as

$$\bar{\varepsilon} \simeq \tfrac{1}{2}.(\mathrm{d}\varepsilon/\mathrm{d}V)(\mathrm{d}V/\mathrm{d}t).T_f$$

giving

$$T_f = [V_s/V]^{\frac{1}{2}}[\tau_{i(s)}(\mathrm{d}\varepsilon/\mathrm{d}V)(\mathrm{d}V/\mathrm{d}t)]^{\frac{1}{2}}[2.\ln(1+\bar{\varepsilon}i_e/i_0)]^{\frac{1}{2}}$$

This expression contains T_f but as the log term varies slowly an approximate value of $\bar{\varepsilon}$ appropriate to the tube and estimated overvoltage can be used.

For example, tube 1 at 5% overvoltage gives $\varepsilon = 0{\cdot}145$.

$$\ln(1+\bar{\varepsilon}i_e/i_0) = \ln(1+e^{20}.0{\cdot}073)$$
$$\simeq 18$$
$$[V_s/V]^{\frac{1}{2}} \simeq 0{\cdot}98$$

Therefore
$$T_f \simeq 6\left[\frac{\tau_{i(s)}}{(\mathrm{d}\varepsilon/\mathrm{d}V)(\mathrm{d}V/\mathrm{d}t)}\right]^{\frac{1}{2}}$$

6. Conclusion

All tubes will have some formative and statistical time lags. These lags may be considered as partly determined by the tube design and partly by the circuit conditions. The reduction of the inherent lags of a design will usually conflict with requirements such as operation at low voltages and stability. It is desirable to re-examine the importance of such requirements in applications where the lags will affect the reliability of operation or the actual breakdown potential.

The most certain way of ensuring very small values of statistical lag is to provide for an auxiliary discharge.

Small values of formative lag may be achieved by operation at high voltages and the relaxation of stability requirements thus permitting the use of faster gas mixtures.

In all cases the use of adequate overvoltages is essential.

7. Acknowledgments

This paper is published by permission of the directors of Hivac Ltd. The author wishes to thank H. G. Brewster, D. Lodge and L. Ince for their valuable help.

8. References

1. F. Llewellyn Jones, "Ionization and Breakdown in Gases" (Methuen, London 1957).
2. J. R. Acton and J. D. Swift, "Cold Cathode Discharge Tubes" (Heywood, London, 1963).
3. J. M. Meek and J. D. Craggs, "Electrical Breakdown in Gases" (Oxford University Press, London, 1953).
4. L. B. Loeb, "Basic Processes of Gaseous Electronics" (Cambridge University Press, London, 1955).
5. R. A. Wijsmann, "Breakdown probability of a low-pressure gas discharge", *Phys. Rev.*, **75**, p. 833, March 1949.
6. F. Llewellyn Jones and E. T. de la Perrelle, "Field emission of electrons in discharges", *Proc. Roy. Soc.*, **216**, p. 267, 22nd January 1953.
7. P. M. Davidson, "Growth of current between parallel plates", *Phys. Rev.*, **99**, pp. 1072–4, 15th August 1955.
8. P. M. Davidson, "Temporal growth of current between parallel plates", *Phys. Rev.*, **109**, p. 1897, 15th September 1956.
9. R. Schade, "Building-up time of glow discharge", *Z. Phys.*, **104**, p. 487, 1936.
10. C. G. Morgan, "Temporal growth of ionization in gases", *Phys. Rev.*, **104**, pp. 566–71, 1st November 1956.
11. Y. Hatta, H. Mase and M. Sugawara, "A new poly-anode counting tube, the 'Polyatron' ", *J. Brit.I.R.E.*, **26**, pp. 383–7, November 1963.
12. S. C. Curran, J. Angus and A. L. Cockcroft, "Beta spectrum of tritium", *Nature*, **162**, p. 302, 1948.
13. A. Noorani, Hivac unpublished internal communication.
14. H. Paetow, "Spontaneous electron emission at electrodes as after-effect of gas discharges", *Z. Phys.*, **111**, Nos. 11–12, pp. 770–90, 1939.
15. F. Llewellyn-Jones and C. G. Morgan, "Surface films and field emission of electrons", *Proc. Roy. Soc.*, **A218**, pp. 88–103, 9th June 1953.
16. F. G. Heymann, "Breakdown in cold cathode tubes at low pressure", *Proc. Phys. Soc. (London)*, **B63**, pp. 25–41, January 1950.

Manuscript received by the Institution on 22nd February 1964. (Paper No. 933.)

DISCUSSION

Mr. W. J. Saysell: Have you any experience of strike time delay using a train of pulses, each pulse being of short duration.

We have made some measurements of the tritium content in finished valves. Variations in valves of the same type occur, some attributable to variation in tritium sources. The effect of absorption by sputtered layers, barium getter films and non-evaporating getters are being investigated.

Mr. Fuller (*in reply*): The statistical lag will be greatly reduced if the pulse repetition rate exceeds a value characteristic of the individual tube. As the tube may also be required to strike first time after a period of inactivity it would be dangerous to rely on a value of mean statistical lag which is based on results observed in this way. The successful operation of a few tubes in prototype circuits could easily mislead a circuit designer into thinking that the pulse width was adequate. If a sufficiently low pulse repetition rate is used (see Section 5.2) then the mean statistical lag quoted will correspond to worst-case conditions and the lags obtained in operation will be better than quoted.

U.D.C. 621.397.621.029.63: 621.391.822

The Reception of Substantially Noise-free U.H.F. Television Signals over Long-distance Paths

By

B. W. OSBORNE, M.Sc.
(Member)†

Presented at a Television Group meeting in London on 4th November 1964

Summary: The factors involved in receiving u.h.f. television transmissions with tolerable noise impairment at ranges exceeding 60 miles are discussed in the light of available and practicable aerial and pre-amplifier techniques, and of the path attenuation. It has proved feasible to use varactor parametric amplifiers fed from aerial arrays of 24 dB forward gain, and this has made it possible to obtain adequate reception at up to 90 miles at some selected sites where the path attenuation does not exceed 180 dB.

Comparison of calculated and measured path attenuations indicates that, where the receiving site is free from local obstructions, the path attenuation calculated on diffraction theory does give a close approximation to the highest measured path attenuations. Field-strength recordings show the largest fading amplitudes during the hours before noon, with a significant decrease in fading amplitude towards the television programme hours of the afternoon and evening.

1. Introduction

The impairment produced by random noise on a 625-line television picture has been discussed by Geddes.[1] For example, for flat noise on a 625-line (5 Mc/s video bandwidth) picture to be imperceptible, the ratio of video signal‡ to r.m.s. noise must be −44 dB. The noise impairment becomes just perceptible at −40 dB, is not disturbing to the viewer at −37 dB, and becomes somewhat objectionable at −29 dB.

In order to relate signal/noise ratios at the receiver input to video signal/noise ratios, it must be remembered that the black to white excursion is only a part of the total signal amplitude, and appropriate allowance must be made for the amplitude of the residual carrier level (during peak white on negative modulation transmission). Thus if the peak voltage of the negative-modulated signal is V, then the maximum black to white excursion may only be 0·6 V if, as is assumed, the residual carrier level is 20%.

Thus to get a 40 dB video signal/r.m.s. noise ratio, we need 45 dB peak signal/r.m.s. noise ratio at the receiver input, or a 42 dB ratio if the signal level is expressed (in the usual manner) as the r.m.s. voltage corresponding to the peak level of the synchronizing pulses (assuming negative modulation). Where signal/noise ratios are quoted below, unless otherwise specified they refer to the latter (see the left-hand column of Table 1).

It is assumed that the voltage applied to the receiver second detector is sufficient for linear working. This condition is generally obtained.

The noise performance of a receiver (at normal temperatures) is conveniently expressed in units of kTB, the noise in a resistor at temperature T over a

Table 1
The visual effect of noise

Signal/r.m.s. noise ratio (dB)		Subjective noise impairment of a 625-line picture (assuming flat noise), after Geddes[1]
From aerial[a]	At video[b]	
30	28	
31	29	Somewhat objectionable
32	30	
33	31	Just tolerable
34	32	
35	33	
36	34	Definitely perceptible but not disturbing
37	35	
38	36	
39	37	
40	38	
41	39	Just perceptible
42	40	
43	41	
44	42	
45	43	
46	44	Imperceptible

[a] The signal level at the aerial is expressed as the r.m.s. level of a sine wave of the same amplitude as that of the synchronizing pulses of a negative modulated television signal assuming 75 : 25 picture/sync ratio and 20% residual carrier.

[b] The video signal amplitude is the black to white voltage excursion (ignoring the synchronizing pulses).

† Rediffusion Research Ltd., Kingston-upon-Thames, Surrey.

‡ Black to white excursion, ignoring sync pulses.

bandwidth B. If the energy bandwidth is taken to be 5 Mc/s with the receiver matched to the aerial, the noise power kTB of a 'perfect' receiver at 290°K is equivalent to 2×10^{-14} W, a noise voltage input of 1·0 µV at 50 Ω impedance.

Thus it is evident that to ensure a 42 dB signal/noise ratio at the receiver input the aerial gain must be such, for a given field strength, as to provide a signal input of $(42+F)$ dB above 1 µV, at 50 Ω, where F is the overall effective noise figure of the receiving equipment (expressed in dB).

F is primarily controlled by the noise figure F_1 of the low-noise pre-amplifier in use, but may also be affected by the noise figure F_2 of the succeeding unit, if the gain G_1 of the pre-amplifier is limited, as mentioned in Section 2.4.1. Note also that any feeder loss reduces the effective value of G_1.

There may also be some degradation of the effective noise figure on site due to the noise temperature of the aerial. (The aerial noise temperature is that at which an equivalent resistance, used in place of the aerial, produces the same thermal noise power as that measured at the receiver output with the aerial connected.) This degradation is small where F is 6 dB or more, but it cannot be ignored where a parametric amplifier of noise figure of about 1 dB is in use. Note that although the Band IV sky temperature away from the galactic plane is only about 100°K, the effective aerial noise temperature for aerial arrays receiving signals nearly tangential to the earth's surface may be much greater (e.g. above 200°K), effectively limiting the lowest usable receiving site noise factor to about 2 dB.

Suppose that an aerial site is being designed to receive signals at field strengths down to E volts per metre, with noise perceptible but not disturbing to the viewer.

As the output voltage across a matched load from an aerial of forward gain G in a field of E volts/metre is $GE\lambda/2\pi$ volts (where λ is the wavelength in metres), we must ensure that G and the overall noise figure F are chosen so that $GE\lambda/2\pi$ is $(42+F)$ dB above 1 µV at 50 Ω impedance.

For example, if F is known to be 8 dB the aerial gain necessary to ensure 42 dB signal/noise ratio for a given incident field strength can be determined. On channel 33 (570 Mc/s) it follows that the use of a pre-amplifier of noise figure 8 dB is permissible, in conjunction with an aerial array of 20 dB forward gain (compared to dipole), provided that the incident electrical field strength equals or exceeds 360 µV/m.

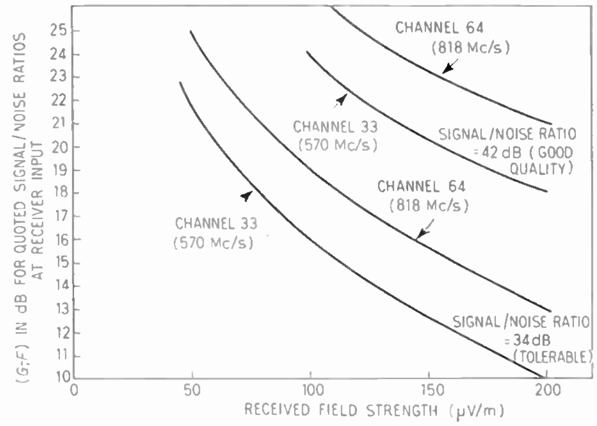


Fig. 1. Aerial gain and receiver noise factor requirements for u.h.f. television reception.

In Fig. 1 the required value of $(G-F)$ in dB necessary to obtain either 42 dB or 34 dB signal/noise ratio is plotted against field strength for two arbitrarily selected u.h.f. television channels (channel 33 at 570 Mc/s in Band IV, and channel 64 at 818 Mc/s in Band V). An economic design practice is to obtain very high aerial gain so as to be able to ensure a satisfactory signal/noise ratio even with a low-cost pre-amplifier of moderate noise factor (e.g. 8 dB), and at the same time to eliminate aerial noise temperature effects which may not be constant and which may not be precisely known. At the more distant sites, however, pre-amplifiers of the best possible noise figures are required, as there are practical limits to the available aerial gain.

Aerial site design will be discussed further in Section 5 below, after consideration of practical values of F the noise factor in Section 2, of the aerial gain G in Section 3, and of path attenuation in Section 4.

2. Low-noise Amplifiers for use on Televison Bands IV and V

2.1. *Choice of Pre-amplifier*

The noise figure attainable at u.h.f. depends on the permissible cost and complexity of the amplifying equipment. Thus whilst various valve or transistor u.h.f. amplifiers[2, 3] can provide noise factors in the range 5 to 9 dB without difficulty and at low cost, to get 1 or 2 dB the present choice is likely to be a varactor parametric amplifier used with a ferrite circulator, such as that described by Pearson.[4]

It is to be noted that at Bands IV and V the performance of the varactor parametric amplifier closely rivals that of the much more costly and complex maser,[2] the latter, however, showing an increasing advantage at higher frequencies.

Tunnel diodes may eventually be useful, particularly over wide bandwidths, but their application appears to be limited by the restriction on maximum signal level. For example, if a tunnel diode has a noise factor of 5 dB, it will be applicable for television reception at signal levels in the range 150 to 300 μV. (At lower levels the tunnel diode noise factor may be inadequate, and at higher levels ordinary valve or transistor units can be used.) Yet in practice, on the distant receiving sites where these levels are obtained, there are periods of low path attenuation when the level of the received signal may be enhanced by as much as 20 dB, so that we must allow for a possible maximum input to the tunnel diode of 1 to 2 mV, which may be more than it can handle. The tunnel diode also suffers from the practical cost disadvantage of needing a circulator, and its advantage of inherently wide bandwidth may be offset by its cross-modulation performance.

It is evident (Fig. 1) that with the aerial gains attainable (Section 3) the range of field strength over which the varactor parametric amplifier may be most useful lies between 30 and 200 μV/m. It was therefore decided to investigate the feasibility of installing and operating parametric amplifiers at remote unmanned receiving sites at distances of between 60 and 100 miles from the B.B.C. channel 33 transmitter at Crystal Palace.

2.2. *Field Tests of Parametric Amplifiers at Remote U.H.F. Reception Sites*

In 1963 field tests of various parametric amplifiers were made at selected receiving sites. The amplifiers tested were either commercially available at the time of the test, or likely to become so shortly afterwards.

The purpose of these tests was not only to determine the noise figure, gain and bandwidth attainable, these being well established from laboratory measurements, but to gain experience on the use of these devices for television reception at remote receiving sites, where engineering maintenance work must be kept to a minimum and consequently where reliability is of the utmost importance. Long-term reliability under field conditions is of prime importance at receiving sites used to feed wired television networks.

One type of parametric amplifier tested on site was a quadrupole amplifier operating in the degenerate mode. The noise figure obtainable from a quadrupole is about 2 dB when used under conditions where the idler circuit noise can be ignored; but for television application the idler noise contribution in a degenerate quadrupole gives a typical noise performance of about 5 dB, inferior to a varactor or other non-degenerate parametric.

A further limitation is that for television the pump frequency must not lie within the range $2(f \pm \delta f)$, where f is the incoming frequency and δf is the highest video frequency, if visible patterning is to be avoided.

Field tests were also made using two different varactor parametric amplifiers. Both of these used a pump frequency of over ten times the incoming frequency and both used a klystron as the pump source. The results obtained on site were similar. Both units proved to be reliable, and effective noise figures on site of about 2 dB were obtained, this including circulator loss, the effect of aerial noise temperature, and the use of a second stage of 8 dB noise figure.

The gain of the parametric amplifiers can generally be set up as desired (by appropriate adjustment of the pump level) within the range 15 to 25 dB; but in order to ensure a bandwidth of 8 Mc/s between 1 dB points it was found convenient to limit the gain to about 16 dB. This gain was sufficient when using a second stage with a noise factor of 8 dB.

. Although the reflector voltage adjustment on the klystron pump source is critical for maintaining a constant tuning point, this did not need re-setting from week to week. There was no deterioration of the *K*-rating of the received picture due to the introduction of the amplifier, and no patterning or other unwanted effects were introduced.

As a result of the field tests, it was considered that the reliability of a varactor parametric amplifier on a remote site is controlled more by that of the klystron pump source and its power supply than by that of the varactor itself; it may be advantageous to replace the klystron with a solid-state pump.

The 'fail-safe' property of the varactor parametric amplifier is a most desirable feature, for if the varactor fails, the incoming signal is still passed to the following amplifier, the noise figure of the latter being only marginally degraded by the small circulator loss. The reduction in overall gain can be taken up by subsequent a.g.c. action.

Typical commercial varactor parametric amplifiers at present in service for u.h.f. television reception are shown in Figs. 2 and 3. It can be seen that in the amplifier of Fig. 2 the klystron pump power supply is separate. The klystron itself, the circulator and the varactor, are mounted together; whilst in the amplifier of Fig. 3 the klystron pump source and its power unit are mounted together, on a separate panel from the varactor. These mechanical considerations are important in relation to feeder loss and to the possible need to position the parametric amplifier close to an aerial, perhaps at the top of a mast, with remote control of the parametric amplifier tuning (see 2.4.2). The amplifier of Fig. 2 has been used on a mast, with remote pump control.

2.3. *Comparison of Noise Figures on Site*

It was found desirable to be able to make a rough comparison of the performance of different pre-amplifiers on a particular receiving site. This usually had to be done during transmission hours, and with-

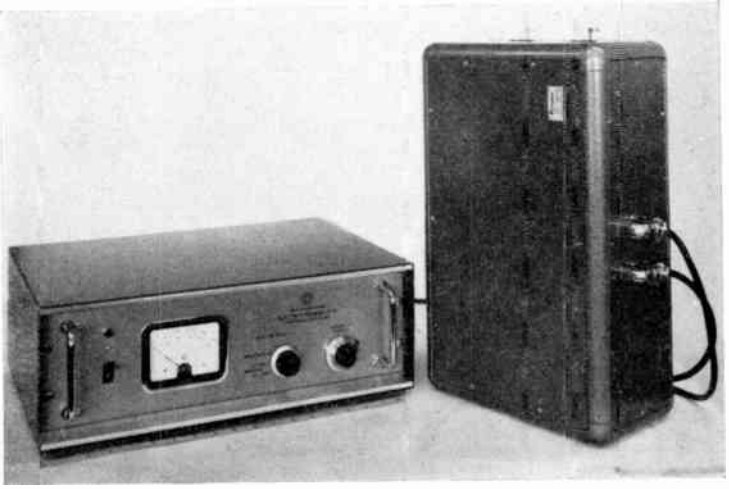


Fig. 2 (*left*). A varactor parametric amplifier with the power supply for the klystron pump in a separate unit, allowing remote control of tuning.

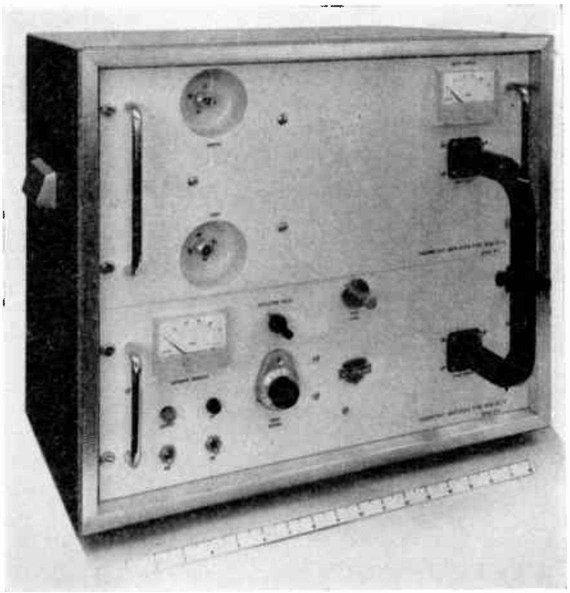


Fig. 3 (*below*). A varactor parametric amplifier capable of operation with the klystron pump remote from the varactor and circulator.

out the availability of laboratory equipment (e.g. noise diode) in a screened room.

Suppose that a receiver, having a noise figure measured and known to be x dB, is used to feed a television display, the input level used being such that noise impairment is clearly perceptible; the same signal input is simultaneously used to feed a second receiver through an attenuator; the attenuator is set at y dB, so that the two television displays simultaneously have the same noise impairment. We can then deduce, all units being correctly matched, that the effective noise figure of the second receiver is $(x-y)$ dB.

This method, though rough, was found to give a useful comparison of the performance of different units on site, during transmission hours.

2.4. *Feeder Loss and Pre-amplifier Gain*

2.4.1. Feeder loss between the pre-amplifier and subsequent units

If amplifiers with individual noise figures of N_1, N_2, N_3 . . . and with power gains P_1, P_2, P_3 . . . are connected in series (the noise figures and the power gains here being expressed in numerical power ratios) then the overall noise figure

$$N = N_1 + \frac{N_2-1}{P_1} + \frac{N_3-1}{P_1 . P_2} + \dots$$

In using this expression, with a pre-amplifier $(N_1. P_1)$ mounted up a mast close to the aerial and feeding a second unit (N_2, P_2) through a feeder cable, the effective gain P_1 in the above expression must be reduced by the feeder loss. For example, suppose that the noise figure of the first unit is 10 dB $(N_1 = 10)$ and that its gain of 30 dB is reduced to an effective 12 dB $(P_1 = 16)$ by the use of feeder cable of attenuation 18 dB; if the noise figure of the second unit is 12 dB $(N_2 = 16)$, then the overall noise figure is degraded to 10·4 dB.

2.4.2. Feeder loss between the aerial and the pre-amplifier

Feeder cable of $\frac{3}{4}$-in diameter might be considered to be about the largest size for convenient use on small remote receiving sites. The attenuation of such a coaxial cable, when made with the conductors spaced by a helical membrane (air dielectric) at a characteristic impedance of 50 ohms, is typically about 1·4 dB per 100 ft at 600 Mc/s. Thus even with this comparatively large diameter feeder it is necessary to mount the pre-amplifier within 35 ft of the aerial if the feeder loss at 600 Mc/s is not to exceed $\frac{1}{2}$ dB. With $\frac{1}{2}$-in solid polythene dielectric feeder cable the corresponding distance is reduced to 10 ft.

In terms of noise impairment to picture, any feeder loss between the aerial and the pre-amplifier is equivalent to a corresponding increase in the overall noise figure of the receiving equipment. It was mentioned in Section 2.2 that at sites where the aerial has to be mounted high, it may be necessary to mount a parametric amplifier up a mast, in close proximity to the aerial, and that it had been found practicable to do this, providing that the pump frequency control is accessible at the foot of the mast. Either the pump energy is fed by a waveguide run up the mast, or the pump klystron power supply leads are led up the mast, both pump and parametric device being mounted high. If the latter method is used the klystron power leads are screened in order to avoid hum modulation of the pump and hence the amplifier signal.

However, the engineering maintenance complications involved in installations of this kind should be avoided whenever possible; so that there are real advantages in choosing distant receiving sites located on hill tops, where the aerial can be mounted close to the ground without loss of field strength.

3. High Gain Aerials

3.1. *Yagi Arrays*

A broadside array of Yagi aerials can provide high gain with low windage. A typical array for channel 33 reception has a forward gain of 21 dB, can be mounted on a single vertical 2-in scaffold pole, and needs only 12 ft vertical × 6 ft horizontal space on the mast. (See Fig. 4.) It was initially considered desirable, when using arrays of 8 elements arranged 4×2, to have the larger aperture dimension vertical, as in Fig. 4. The reason for this was to get a smaller vertical half angle on the main lobe, and thus to reduce any deterioration of effective aerial noise temperature due to noise pick-up from the earth's surface (nearby towns, etc). Against this, it may be that meteorological

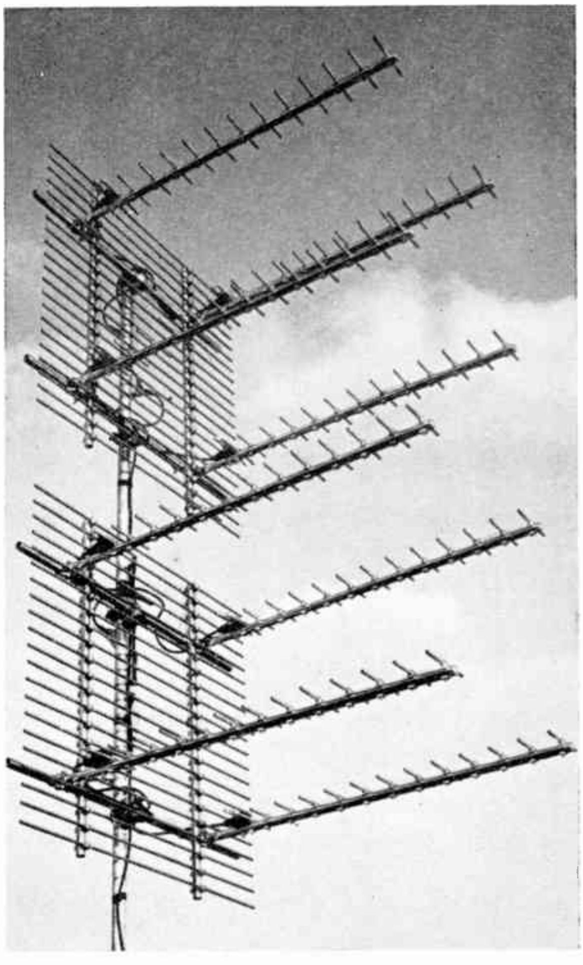


Fig. 4. A 21 dB u.h.f. aerial array for 570 Mc/s, showing the back reflector.

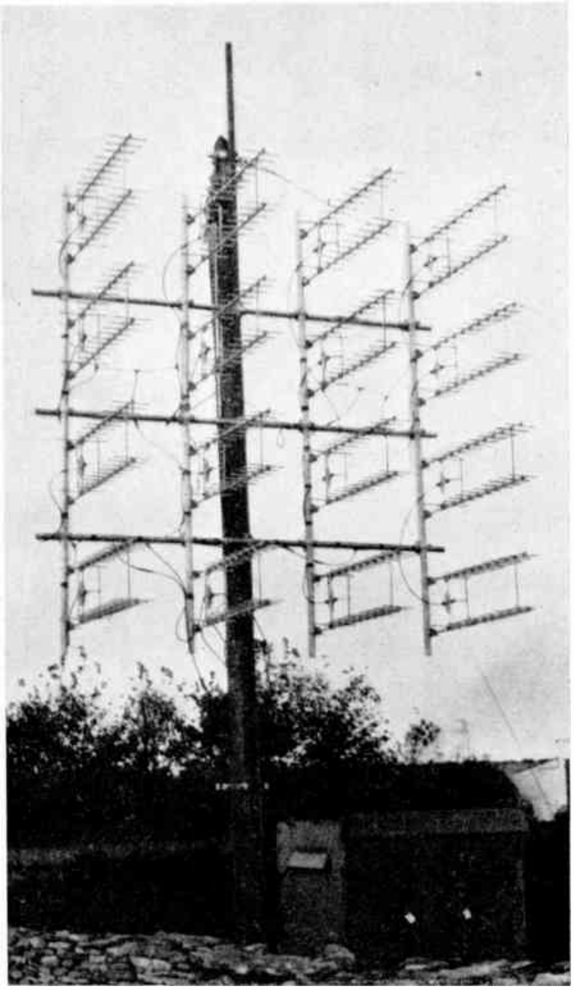


Fig. 5. A 24 dB u.h.f. aerial array (approximately 12 ft × 12 ft).

changes in the propagation path, particularly near the diffraction point, could cause very slight variations in the angle of elevation of the received signal, without change in bearing. To what extent these factors matter when using an array with vertical or horizontal main lobe half angles of about 7 deg (to half power points) is not clear.

An aerial gain of 24 dB has been obtained on site using the array of Fig. 5. However, the attainment of higher gains than 24 dB at u.h.f. by use of a broadside array is likely to be difficult, for increasing the size of the array leads to a rapid increase in the losses due to the feeders and to the combining networks. It may be noted that the comparable gain at Band III (200 Mc/s), using an array of the same overall dimensions and of the same type, is about 18 dB.

The combining of the elements of a multiple Yagi array can conveniently be done by use of a quarter-wave combiner, such as that shown in Fig. 6. This is simply a quarter-wave transformer constructed so that the inputs are in parallel, the transmission line being made with a square outer section so that the coaxial connectors can be conveniently mounted to it. When combining four inputs the characteristic impedance of the combiner is half that of the feeder impedance.

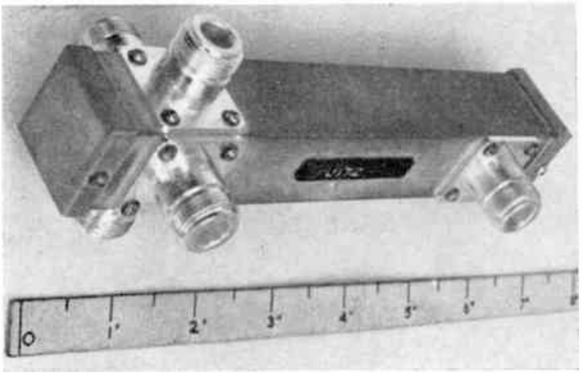


Fig. 6. A 4-way combiner for use on u.h.f. aerial arrays.

The spacing of the elements of an array is not critical, provided that the spacing is sufficient to reduce coupling between elements to a low level. The spacing may be determined, on particular sites, by the need to reduce pick-up from a co-channel transmitter at some particular back angle.

The phasing of an array, such as that of Fig. 4 or of Fig. 5, can be most conveniently performed and the radiation pattern determined, by mounting the array near the ground at an angle of 45 deg to the vertical, with a locally generated signal source mounted at a high level in line with and radiating towards the array. This avoids unwanted ground reflection, and the array can afterwards be moved to a remote site or mounted on a mast. However, it is sometimes desirable, for example on rounded hill-top receiving sites, to use a large array close to the ground. It is then necessary with arrays of large aperture to phase on site, using a steady distant signal.

3.2. *Paraboloids and Corners*

The aperture of a paraboloid of comparable gain, for television reception at very distant sites on Band IV, would need to be over 16 ft. A paraboloid of this size is expensive and unsightly, and furthermore is larger than the Yagi array of similar forward gain, more visible and has higher windage.

A practical 45 deg corner with sides of 3λ can give 14 dB forward gain (compared with a half-wave dipole). The use of a co-linear or stacked array of four such corners to obtain 20 dB gain results in an aerial of high windage, not suitable for use on towers, although one possible application is on low-level sites where the front-to-back ratio is of paramount importance.

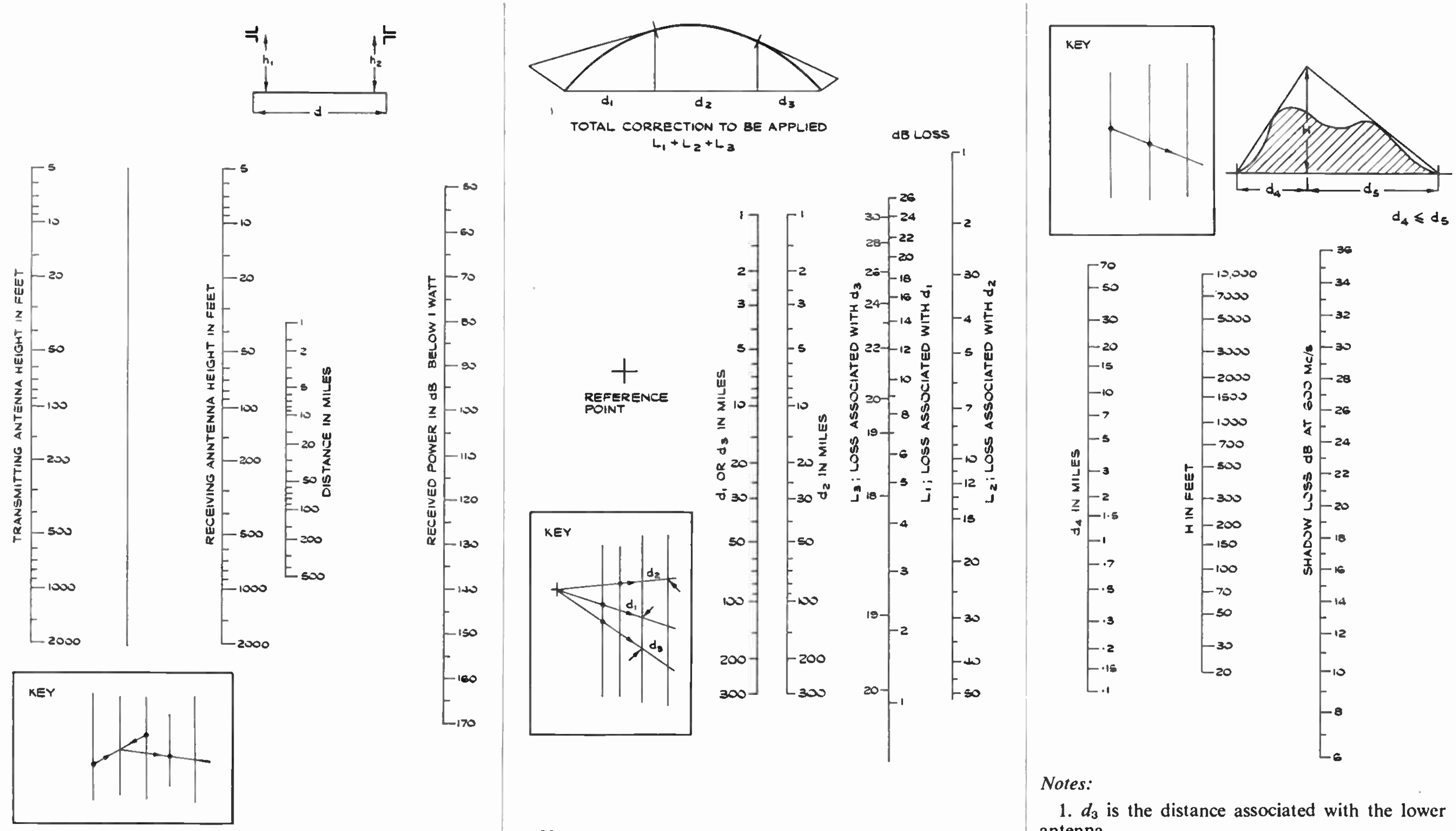
3.3. *Aerial Gain*

Measurement of aerial gain has been discussed elsewhere.[5, 6, 7] Whilst it is conventional to use the aerial under test as a transmitting aerial and to measure the radiated power, the effective gain on site of a large array near the ground may be affected by local conditions, ground reflections, etc., and it may be more convenient to determine the effective gain under existing site conditions by comparing the voltage output from the aerial array with that obtained from a smaller aerial of known gain, with the same incident field strength. This avoids the complication of determining the field strength variations over the aperture of a large array mounted near to the ground, and does give the information required for the assessment of the performance of the receiving site as a whole.

3.4. *Wideband Aerials*

Hitherto we have ignored the possible eventual requirement that a high-gain television receiving array would have a wide bandwidth, perhaps 80 to 100 Mc/s, in order to receive a group of four u.h.f. television signals radiated from a common transmitter site.

As the gain of an array is increased, the difficulty of getting adequate performance over a wide frequency band becomes greater, due not only to the inherent bandwidth of the components of the array but also to all frequency sensitive components in the feeder combining networks. To work over a wide bandwidth the forward gain must be nearly constant with frequency, and the back-to-front ratio and the impedance seen looking into the feed point must not vary greatly with frequency.



Notes:

1. Use flat earth diagram to compute shadow loss.
2. This diagram is only applicable to 600 Mc/s.

Notes:

1. d_3 is the distance associated with the lower antenna.
2. $d_3 \leqslant d_1$.
3. This diagram is only applicable at 600 Mc/s for 4/3 earth's radius.

(a) Uncorrected attenuation between half-wave dipole aerials.

(b) Diffraction loss.

(c) Shadow loss.

Fig. 7. Nomograms for calculating u.h.f. path attenuation over a long distance diffraction path. (After Bullington.[11])

These requirements, although easily met by a helix or by a suitable Yagi with a folded slot-fed element, are not easily met on a multiple array, whether of helices or of Yagis, and may favour a physically large corner array or paraboloidal section. As the windage would be high, the use of such aerials is more suited to rounded hilltops than to high masts.

4. Path Attenuation

4.1. *Method of Calculation*

The following discussion is restricted to consideration of the u.h.f. path attenuation to selected receiving sites, these being free from local obstructions and, where possible, on round hill-tops where the ground falls away in the direction of the transmitter. For more general discussions of u.h.f. television reception see Smith-Rose,[8] Robb,[9] and Whitbread.[10]

It was found possible to adapt the method used by Bullington[11] to obtain a quick and easily applied method for determining the likely u.h.f. path attenuation on a long-distance diffraction path.

Although the method does not take into account any effects due to the sides of valleys or other three-dimensional ground configurations, it has been found to give good forecasts of the likely minimum u.h.f. field strength at selected long distance receiving sites.

In the nomograms of Fig. 7(a)–(c), in the interest of simplicity, a frequency of about 600 Mc/s is assumed.

The procedure used, a simplification of Bullington,[11] is as follows:

(i) The path profile between transmitting and receiving aerials is drawn and plotted on 4/3 earth's radius paper, and the 'average ground level' estimated.

(ii) The uncorrected attenuation is obtained from Fig. 7(a). On some paths it is necessary to refer to 'average ground level' in deciding on the height of the transmitting and receiving aerials to be used in the calculation.

(iii) The diffraction correction is obtained from Fig. 7(b), the tangents from the top of the transmitting and receiving aerials being drawn to the 'average ground level' in order to determine the distances d_1, d_2, d_3.

(iv) The shadow-loss correction is determined by drawing the path profile on linear paper (not 4/3 earth radius as in (i)) and by drawing tangents to this profile from the bases of the transmitting and receiving aerials (see Fig. 7(c)). The height H and the shorter distances d_4, d_5, are used to determine the shadow loss from Fig. 7(c). A minimum shadow loss of 6 dB is to be assumed.

(v) The total path attenuation is the sum of the uncorrected attenuation (ii), the diffraction loss (iii), and the shadow loss (iv).

(vi) Add transmitter e.r.p. (in dB above 1 W) to the gain of the receiving aerial (in dB, referred to half-wave dipole). Subtract this sum from the total path attenuation to get the total received power, in dB referred to 1 watt.

Table 2

Calculated and measured u.h.f. path attenuations (570 Mc/s)

Transmitting aerial at about 700 ft above ground, the ground height being 300 ft above sea level. Receiving aerial height generally at 20 ft above ground is assumed.

Receiving site	Distance (miles)	Receiving site height above sea level (ft)	Calculated path attenuation (dB)	Measured path attenuation (dB)[a]	Remarks
A	109	750	219		No reception.
B	98	650	211		Intermittent reception, but no regular service possible.
C	90	700	178	175–179[b]	Usable signal for a television service (noise perceptible but tolerable).
D	82	640	180	176–180[b]	Usable signal for a television service (noise perceptible but tolerable).
E	78	400	183	—	No measurements.
F	78	700	171	168–175	Steady signal, little fading.
G	65	275	167	165–170	Steady signal, little fading.
H	60	350	170	168–175	Greater fading amplitude at times than at site G. Usable signal.

[a] The figures quoted are those for normal days. Abnormally high field strengths are at times obtained on all sites.

[b] Figures obtained from analysis of several weeks' field strength records apply for over 90% of the time.

The application of this method is sufficiently rapid for it to be of use in selecting suitable sites from a map, in advance of any experimental site checks. The latter may in fact be considerably less reliable than the theoretical diffraction path information, unless field-strength recordings are made over a sufficiently long period.

4.2. *Calculated Path Attenuation to Various Receiving Sites*

The path attenuations to various distant sites, determined by the above method, are tabulated in Table 2, and compared with the corresponding experimentally obtained figures. It is evident from Table 2 that the measured figures are close to those calculated (the measured figures being those for normal days, the received signals often being higher than the quoted levels due to meteorological changes).

Measured field strengths at two particular sites (sites C and D in Table 2) will be described in the following section.

4.3. *Field-strength Recordings at Certain Distant Sites*

4.3.1. Measurements at site C

This receiving site, in the East Midlands, is situated at a height of about 700 ft above sea level, and 90 miles from the transmitter. The estimated uncorrected attenuation was 132 dB, with an additional 40 dB diffraction loss and 6 dB shadow loss, the total estimated path attenuation thus being 178 dB (to an accuracy of perhaps ± 6 dB). The receiving aerial was used for measurements at a height of less than 20 ft above ground.

Field-strength recordings were obtained on channel 33 during programme hours over some 4 weeks in April and May 1964. The estimated received field strength corresponding to the 178 dB path attenuation was 77 μV/m. The measurements showed that the received field strength was normally in the range 70 to 100 μV/m, the received level only going below 40 μV/m for 0·3% of the time. It was noticed that field strengths in excess of 300 μV/m were frequently obtained, particularly before noon, and that the fading amplitude was typically 18 to 20 dB during the hours before noon (e.g. 0800 to 1100 G.m.t. in April–May 1964).

It was noted that the received signal was free from any echoes or other visible multipath effects even during the large amplitude fades.

4.3.2. Measurements at site D

This receiving site, in South-West England, is situated at a height of about 640 ft above sea level, and 82 miles from the transmitter.

The estimated path attenuation was 180 dB (again to an accuracy of perhaps ± 6 dB). The receiving aerial (shown in Fig. 5) was mounted close to the ground. The site was on the top of a conical hill, situated in the centre of a wide valley lying along the propagation path. In spite of the possible danger of multi-path reception from the hills on either side of the valley, no echoes were observed on the received picture.

Field-strength recordings, made during a total of 72 hours in February 1964, showed that the measured field strength only fell below the calculated value for 1% of the time. For over half the time the received levels were over 6 dB above that estimated.

Again, the occasional presence of very high field strengths during the hours before noon was noted and, as at site C, the fade amplitudes were significantly less during the afternoon and evening (programme) hours than during the hours before noon.

4.3.3. Diurnal changes in propagation conditions

The information from the field-strength recordings at site C was used to obtain information on:

(a) the diurnal variation of maximum fading amplitude;

(*b*) the diurnal variation of the difference between the median received field strength and that estimated on diffraction theory.

The points plotted are the medians of the maximum fade amplitudes for the particular half hour (using all records available). The resultant curve of Fig. 8 shows the marked reduction in fading after noon. It should be mentioned that experimental results were restricted to the period between 0800 and 1900 G.m.t.

The median received field strength could be determined from the recordings since these were made on a

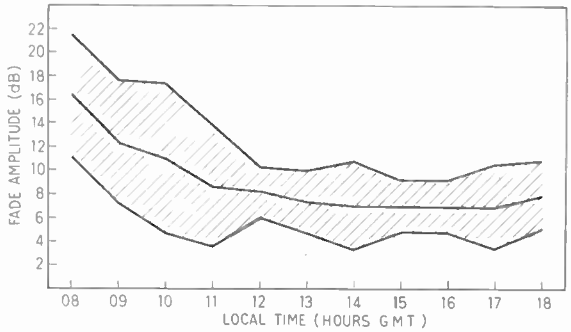


Fig. 8. Variation of maximum fade amplitude with local time. Results obtained using 168 figures of maximum fade amplitude over 30-min periods at 570 Mc/s over 90-mile path, April/May, 1964. Standard deviations are shown by the boundaries of the shaded area.

sampling basis, one point being recorded every 5 seconds.

In Fig. 9 the mean difference between the median (for any hour) and the 'normal' diffraction path level is plotted (in dB) against local time. The 'normal' level was that obtained from diffraction theory.

It is evident, by comparing Fig. 8 with Fig. 9, that on long distance u.h.f. diffraction paths when the field strength is high, the fade amplitude is also high; and that these conditions are obtained during morning hours, the fading and the level returning to 'normal' after noon.

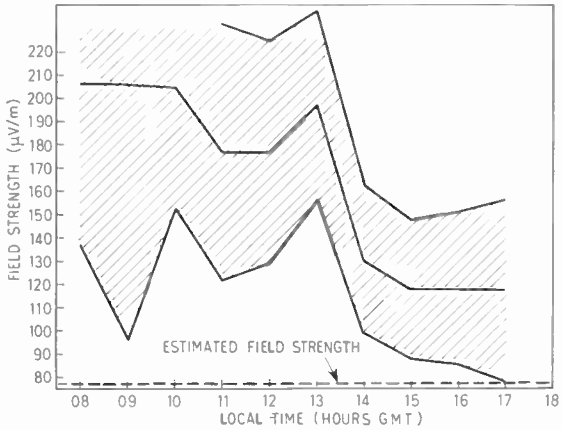


Fig. 9. The diurnal variation of median received field strength (April/May, 1964, path length of 90 miles at 570 Mc/s). Standard deviations are shown by the boundaries of the shaded area.

On the sites used to obtain the information of Figs. 8 and 9, the fading is minimized by the selection of a receiving site where the ground falls away rapidly in the direction of the transmitter. It is known that on neighbouring sites where there is hilly or even level ground in the foreground, the fade amplitude is often higher and the fading more persistent.

Although the experimental material referred to here is sparse, the results do not appear to be inconsistent with those of Bean and Horn.[13]

5. Choice of Aerial and Pre-amplifier for a Given Path Attenuation

5.1. *Necessary Conditions*

If the effective radiated power from the transmitter is P dB above 1 watt, the path attenuation x dB and the receiving aerial gain (over a half-wave dipole) is G dB, then the power available from the receiving aerial is $(P+G-x)$ dB, referred to 1 watt.

It follows from Section 1 that for Q dB signal/noise ratio at the receiver input, $(P+G-x)$ must equal or exceed the power level corresponding to $(Q+F)$ dB over 2×10^{-14} watts. As 2×10^{-14} watts is -137 dB relative to 1 watt, the requirement is that

$$(P+G) \geqslant (x+Q+F-137).$$

If the effective radiated power P and the maximum practical aerial gain G are known, then we can determine the desirable value of the receiver noise figure F for a given path attenuation x, remembering that aerial noise generally limits the effective overall noise figure of the receiving site equipment to values of 2 dB or greater.

If for acceptable signal/noise ratios (Table 1) we require $Q = 42$ dB, then we must satisfy

$$(P+G) \geqslant (x+F-95);$$

for example, if $P = 57$ dB and $G = 24$ dB, the condition is that $(x+F) \leqslant 176$ dB. However, the signal may still be tolerable ($Q = 34$ dB) if $(x+F) \leqslant 184$ dB (see Fig. 1).

The path attenuation x can be taken as known if the calculated value (Section 4) is within 3 dB of that measured on site under normal conditions, the site measurements being made over a period of several days. On distant sites where there is severe fading the field strength should be recorded over as long a period as possible, in order to determine x before deciding on aerial and pre-amplifier requirements.

5.2. *Examples*

If an aerial with 24 dB gain is used it is evident that a noise factor of 8 dB will suffice when $P = 57$ dB and the path attenuation does not exceed 168 dB. However, if a parametric amplifier of 2 dB (or better) noise factor is available, substantially noise-free signals are attainable at path attenuations of up to 174 dB. The signal may, of course, still be of tolerable noise impairment at path attenuations of up to 182 dB.

If we assume F to be 8 dB (avoiding the use of the more costly parametric amplifier) then we can determine the necessary aerial gain for a given path attenuation x, since we must ensure that $(x-G)$ does not exceed 144 dB for 42 dB signal/noise ratio into the receiver, the transmitted e.r.p. again being assumed here to be 57 dB above 1 watt.

To take some practical examples, first consider site F at 78 miles, where the path attenuation of 171 dB was closely confirmed by experimental field strength measurements.

With 24 dB aerial gain, to satisfy the requirement $(x+F) \leqslant 176$ dB we need a noise figure of 5 dB. Thus it was decided to use a parametric amplifier on this site, giving an effective F of about 2 dB; with this noise figure it was then evident that the use of 21 dB aerial gain satisfied the requirement

$$(P+G-x) \geqslant (F-95).$$

It was therefore decided to use a parametric amplifier on site F in conjunction with a 21 dB aerial.

However, on site D, over an 82-mile path, the calculated path attenuation of 180 dB indicated that, with the highest aerial gain (24 dB) easily practicable, with a parametric amplifier giving an effective noise figure (including aerial noise) of 2 dB, and with $P = 57$ dB, the signal/noise ratio at the receiver input would still be as low as 36 dB, giving noise impairment (see Table 1) that would be classified as tolerable. Fortunately, field-strength measurements showed that the actual path attenuation to site D was typically nearer to 176 dB, giving slightly better signal/noise ratios than would have been predicted.

6. Conclusions

Present techniques make it possible to achieve 24 dB aerial gain and a 2 dB effective noise figure at 570 Mc/s on small, remote u.h.f. aerial sites. Good quality television pictures free from multi-path effects can be received, often at up to 80 or 90 miles distance, with an e.r.p. of 57 dB above 1 watt where the path attenuation is not more than 174 dB; tolerable television pictures can be received where the path attenuation is not greater than 182 dB. The calculated attenuation of these long distance diffraction paths is in close agreement with measured field strengths, except when meteorological conditions cause temporary enhancement of the level of the received signal. Fading is more prevalent before noon, when the highest peak field strengths are obtained.

7. Acknowledgments

Thanks are due to Mr. J. D. Pearson and Mr. V. M. Pastore, both of Ferranti Ltd., to Mr. C. Aitchison of Mullard Research Laboratories, and to Mr. C. Davenport of the M-O Valve Co. Ltd., for loans of parametric amplifiers for site tests during 1963; to Mr. D. Gill of the Marconi Co. Ltd., for the loan of a circulator for the same purpose; to Mr. C. A. Burgess of Associated Aerials Ltd., in connection with aerials; to Dr. G. J. Phillips, Mr. R. V. Harvey and others at the B.B.C. Research Department for some most useful discussions; to Mr. S. W. Wells, Chief Engineer of Rediffusion (South West) Ltd., at whose aerial sites many of the parametric amplifiers and experimental aerials were tested; to Mr. R. A. Frankum, Mr. C. M. Payne and Mr. A. M. Peverett for their contributions at Rediffusion Research Ltd. in the fields of aerial work, of path attenuation calculation and of low noise amplifier design respectively; and to the Directors of Rediffusion Research Ltd., for permission to publish this paper.

Figures 7(a), (b) and (c) are based on figures previously published in the *Proceedings of the Institute of Radio Engineers* (America)[11] by permission of Mr. K. Bullington of the Bell Telephone Laboratories.

8. References

1. W. K. E. Geddes, "The relative impairment produced by random noise in 405-line and 625-line television pictures", *E.B.U. Review*, Part A, No. 78, pp. 46–8, 1963.
2. R. Hearn, R. J. Bennett and B. A. Wind, "Some types of low noise amplifier", *J. Brit.I.R.E.*, **22**, pp. 393–403, 1961.
3. Ulrich L. Rohde, "Very low noise transistor amplifiers in the u.h.f. band using the parametric conversion mode", *J. Brit.I.R.E.*, **24**, No. 3, pp. 223–8, September 1962.
4. J. D. Pearson, "Parametric amplifiers using semiconductor junctions", *Research*, **15**, pp. 483–8, 1962.
5. "Methods of measurement of essential electrical properties of receiving aerials in the frequency range from 30 Mc/s to 1000 Mc/s", I.E.C. Publication 138, 1962.
6. E. V. Jull, "The estimation of aerial radiation patterns from limited near field measurements", *Proc. Instn Elect. Engrs.* **110**, pp. 501–6, 1963.
7. R. G. Manton, "The calculation and measurement of the gains of end-fire v.h.f. and u.h.f. aerials", *Electronic Engineering*, **36**, pp. 8–11, 1964.
8. R. L. Smith-Rose, "Radio wave propagation and the problems of television bands IV and V", *J. Television Soc.*, **8**, pp. 59–60, 1956.
9. A. C. Robb, "U.h.f. television reception", *Wireless World*, **69**, pp. 385–90, 1963.
10. C. F. Whitbread, "Receiving aerials for u.h.f. television" *J. Television Soc.*, **10**, pp. 243–53, 1963.
11. K. Bullington, "Radio propagation at frequencies above 30 Mc/s", *Proc. Inst. Radio Engrs*, **35**, pp. 1122–36, 1947.
12. R. L. Smith-Rose, "A survey of British research on wave propagation with particular reference to television", *Proc. Instn Elect. Engrs*, **99**, Part IIIA, pp. 270–80, 1952.
13. B. R. Bean and J. D. Horn, "Radio-refractive-index climate near the ground", *J. Res. Nat. Bur. Stand.*, **63**, pp. 259-71, 1959.

Manuscript received by the Institution on 8th April 1964. (*Paper No. 934/T29.*)

U.D.C. 621.396.96

A Side-lobe Suppression System for Primary Radar

By

J. CRONEY, B.Sc.†

AND

P. R. WALLIS, B.Sc.‡

Reprinted from the Proceedings of the Symposium on "Signal Processing in Radar and Sonar Directional Systems", held in Birmingham from 6th–9th July, 1964.

Summary: An auxiliary omni-directional receiving aerial and a logarithmic receiving channel identical to the radar receiving channel are used. A refinement of the system provides a sharp cusp in the omni-directional polar diagram coincident with the main lobe of the radar aerial. Non-linear processing of the signals in the radar and omni-directional channels gives complete suppression of side-lobe signals. The operation of the system in the presence of jamming is examined. Illustrations are given of the performance of the system under several conditions. It is especially effective in suppressing pulse interference from other radars.

1. Introduction

The experiments described in this paper were made early in 1955[1,2] and are thought to be the first application of side-lobe suppression techniques to primary radar. Publication has not been possible until recently. The radar displays and receivers from which the photographs in this paper were taken were developed well over 10 years ago, using the valves and cathode-ray tubes of that day, and no apology is therefore made for the poor definition of the pictures, compared with those of modern cathode-ray tube displays. Such imperfections are immaterial to the efficacy of the system described.

2. System Description

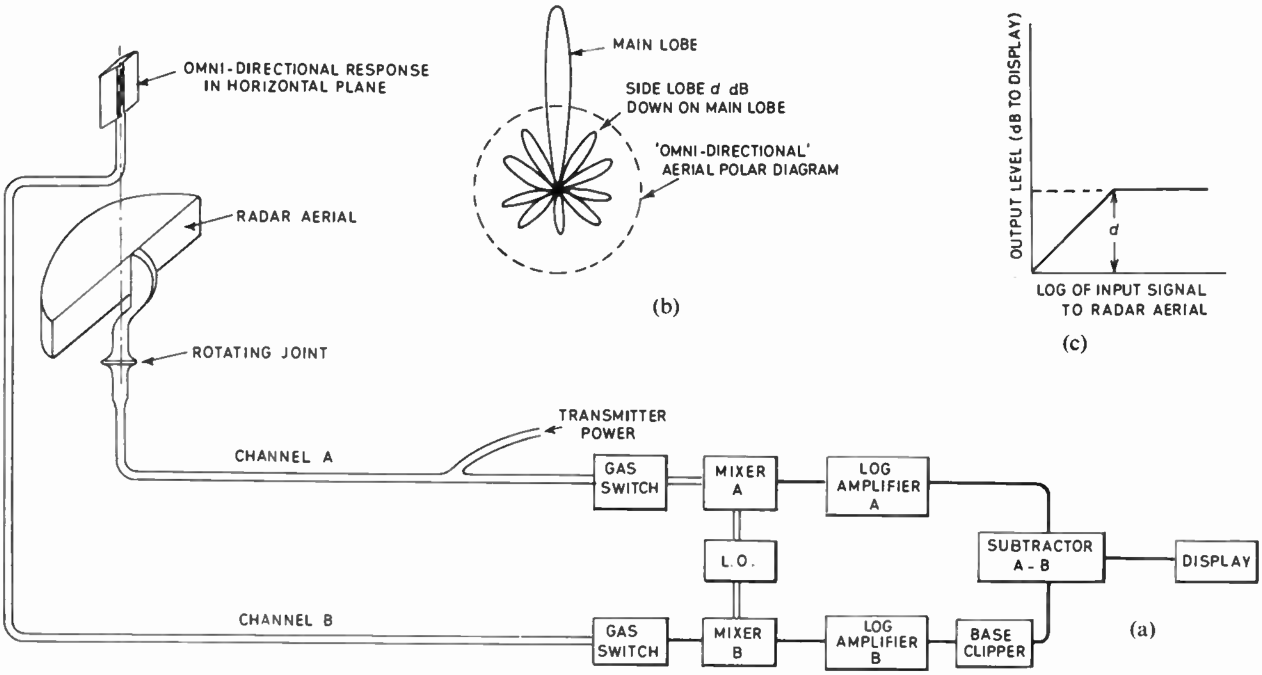
Two versions of the system were proposed in reference 1, and experiments were made with both. The first version is shown in Fig. 1(a). An 'omni-directional' aerial of the type described in reference 3 is mounted approximately above the centre of revolution of the radar aerial so that it is subject to the same blind arcs (caused by nearby obstacles) as is the radar aerial. No transmission is made from this aerial and it need not rotate. The signals it receives are fed to a microwave superheterodyne amplifying channel identical to the one fed by the radar aerial. The i.f. amplifiers of both these receiving channels have logarithmic input-output laws and are matched as closely as possible in their response characteristics.

The gain of the omni-directional aerial is designed to be at least equal to, and if possible in excess of, that associated with the worst sidelobe of the radar aerial (d dB down on the main-lobe). This is shown pictorially in Fig. 1(b). Its vertical pattern should agree generally with that of the radar aerial. A side-lobe echo picked up by the omni-directional aerial will now always be greater than the same side-lobe echo picked up by the radar aerial. The signal at the output of channel B logarithmic amplifier is therefore greater than that at the output of channel A amplifier. For the worst side-lobe (d dB down) the channel B and channel A signals can be adjusted to equality by clipping off the base of the channel B signals. The subtractor unit then gives zero signal at the display, the side-lobe echo being completely cancelled.

The base clipper adjustment is set permanently to give cancellation of the worst side-lobe echo. For other side-lobe echoes there will then be some over-cancellation, so that a signal 'blacker than black' is fed to the display. This implies an unnecessary loss of sensitivity to wanted main-beam signals which happen to be painting in the black region of a suppressed side-lobe. Such over-suppression possibilities are considered in detail in Section 5. In the case of main-lobe signals the omni-directional aerial feeds through channel B to the subtractor unit, a signal which is d dB down on the signal from channel A, provided that the target echo is strong enough to exceed the noise and clipper level in the omni-directional channel. The input/output characteristic of the combined receiving system is shown at Fig. 1(c). The limiting level of d dB (usually about 20 to 25 dB) is still well in excess of the limiting level of the normal p.p.i. so that there is no loss of intensity in echo painting.

† Admiralty Surface Weapons Establishment, Hambrook, near Chichester, Sussex.

‡ Formerly at A.S.W.E.; now with the Admiralty Underwater Weapons Establishment, Portland, Dorset.



(a) Suppression circuit arrangement for simple 'omni-directional' characteristic.
(b) Diagram illustrating radar and 'omni-directional' aerial radiation diagram.
(c) Receiving characteristic of side-lobe suppression system.

Fig. 1. Side-lobe suppression system.

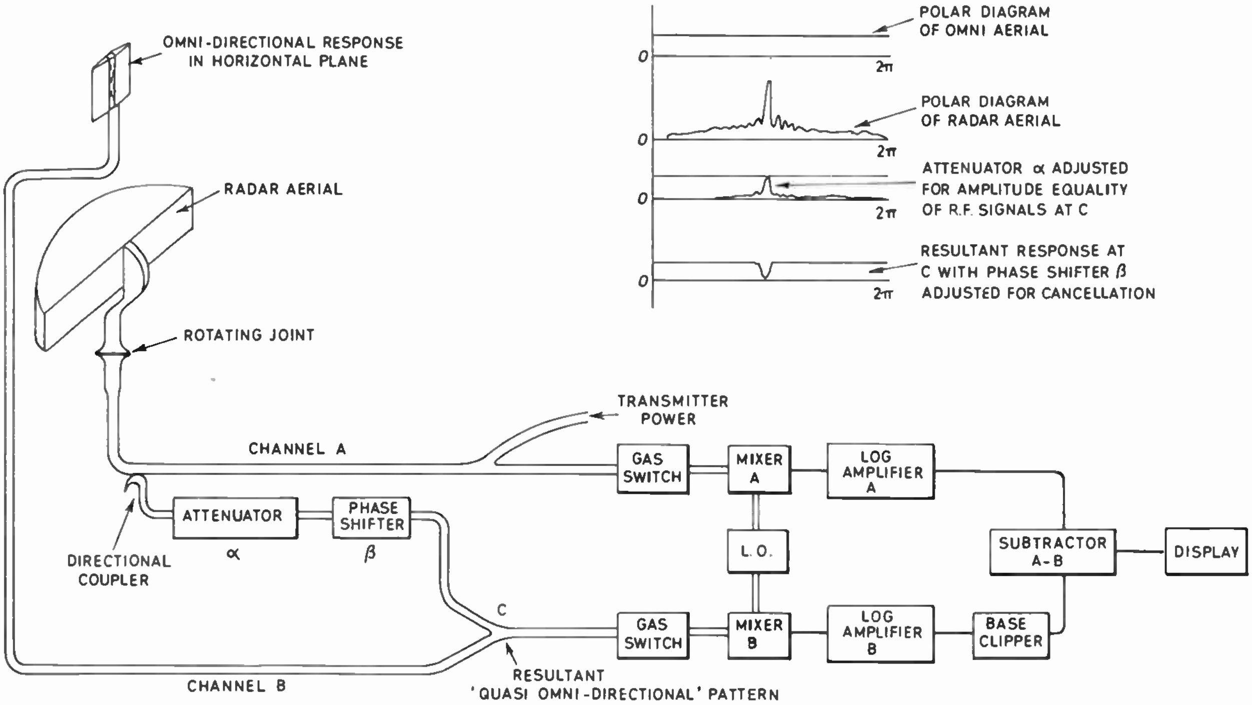

OMNI-DIRECTIONAL RESPONSE IN HORIZONTAL PLANE
RADAR AERIAL
ROTATING JOINT
POLAR DIAGRAM OF OMNI AERIAL
0
2π
POLAR DIAGRAM OF RADAR AERIAL
0
2π
ATTENUATOR α ADJUSTED FOR AMPLITUDE EQUALITY OF R.F. SIGNALS AT C
0
2π
RESULTANT RESPONSE AT C WITH PHASE SHIFTER β ADJUSTED FOR CANCELLATION
0
2π
TRANSMITTER POWER
CHANNEL A
GAS SWITCH
MIXER A
LOG AMPLIFIER A
ATTENUATOR
α
PHASE SHIFTER
β
DIRECTIONAL COUPLER
L.O.
SUBTRACTOR A-B
DISPLAY
C
GAS SWITCH
MIXER B
LOG AMPLIFIER B
BASE CLIPPER
CHANNEL B
RESULTANT 'QUASI OMNI-DIRECTIONAL' PATTERN

Fig. 2. Suppression circuit arrangement for 'quasi-omni-directional' characteristic.

The extra receiving channel (channel B) introduces additional noise to the display and in the absence of 'base clipping' would degrade the signal/noise ratio by at least 3 dB. Where the radar aerial parameters allow the omni-directional aerial to be designed to have a gain in excess of that of the worst side-lobe of the radar aerial, it is of advantage to do so, since 'base-clipping' can subsequently be used to eliminate the effect of this excess aerial gain. In so doing, some of the noise of channel B is also eliminated and the degradation of signal/noise ratio of the system can be made substantially less than 3 dB.

Figure 2 shows the second version of the scheme. In this case the omni-directional aerial is mounted above the radar aerial with its vertical axis exactly coincident with the axis of rotation of the radar aerial. A fraction of the received radar signal (at the microwave frequency) is coupled into the omni-directional channel B, in amplitude equality and phase opposition with the omni-directional received signal. The signal at the waveguide junction C is therefore that received by an omni-directional aerial having a sharp cusp in its characteristic which rotates with the main beam of the radar aerial. This is referred to as a 'quasi-omni-directional' characteristic. The effect of this cusp in the omni-directional pattern coincident with the main beam is to raise the limiting level of the output signal at the display. If a cusp of x dB can be produced, the limiting level in Fig. 1(c) is raised from d dB to $(d+x)$ dB. The extent to which the phase and amplitude cancellation necessary to produce the cusp can be maintained over the whole 360 deg of rotation of the radar aerial, is discussed in Section 4.

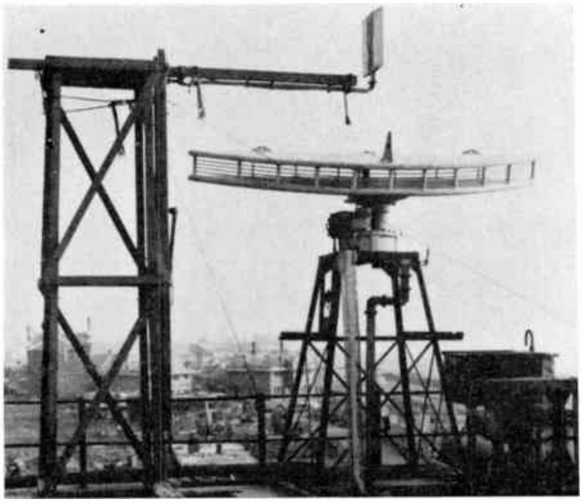


Fig. 3. Aerial arrangement showing the side view of an 'omni' aerial.

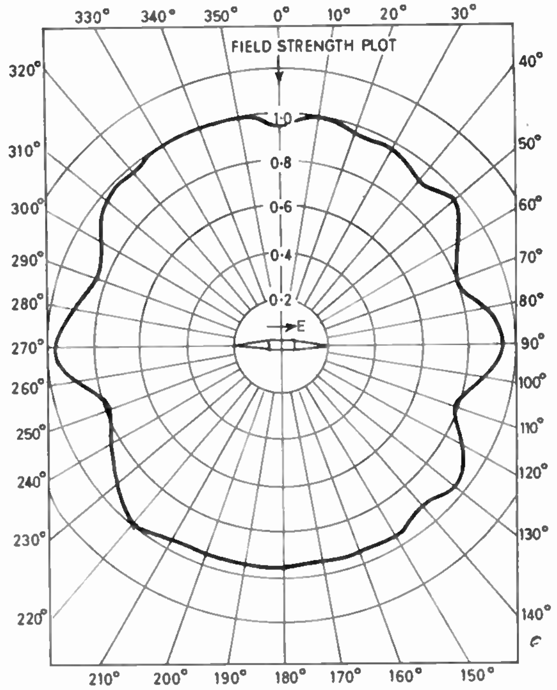


Fig. 4. Radiation diagram of the omni-directional aerial.

3. Details of Circuit Arrangement

The system was applied to an S-band naval radar situated on the coast at Southsea. Very serious side-lobe echoes were obtained at this site since large blocks of buildings stretch for miles along the coast from zero range outwards. Side-lobe echoes painted on the p.p.i. as continuous rings out to a range of 5 miles.

The complete circuit arrangement of Fig. 2 was installed, but the attenuator and phase-shift link between channels A and B could easily be disconnected to give the simple system of Fig. 1(a).

The omni-directional aerial was designed to give a pattern of the same vertical beamwidth as the radar aerial, the gain being then about 20 dB down on the radar aerial, and about 4 dB more than that associated with the worst side-lobe of the radar aerial. The omni-directional aerial is shown mounted above the radar aerial in Fig. 3. The radiation diagram of the omni-directional aerial in the horizontal plane is given in Fig. 4.

When the simple system of Fig. 1 was in use, the gas switch and mixer of channel B was placed just below the omni-directional aerial, the local oscillator being 'piped up' from the office through a low-loss microwave cable. In this way a long run of waveguide from the omni-directional aerial (entailing many bends and corners) was avoided. The resulting increase in effective gain of the omni-directional aerial, allowed more base clipping to be employed after the logarithmic amplifier in channel B, as described in Section 2 above.

The logarithmic i.f. amplifiers were of the successive detection type (i.f. = 13·5 Mc/s, bandwidth = 2 Mc/s) described in reference 4. They were each preceded by a head amplifier and filter unit giving an overall bandwidth of 1 Mc/s in each channel. Each receiver chain was logarithmic down to about 20 dB below the r.m.s. level of the output noise, and up to about 80 dB above. A typical circuit of the logarithmic amplifiers may be found in reference 4. The input/output laws of the two amplifiers are shown in Fig. 5. No procedure other than careful 'lining-up' was involved in producing this degree of equivalence. The 'base clipping' unit shown in channel B was in fact incorporated in the subtraction unit, the circuit of which is given in Fig. 6.

4. Results

4.1. *Simple Omni-directional Aerial*

The first experiments were made using the simple system, the s.h.f. link between channels A and B being disconnected. Figure 7(a) shows the normal p.p.i. display (range scale 12 nautical miles) obtained from the naval radar by feeding the output of the logarithmic amplifier of channel A (Fig. 1(a)) direct to the display. Figure 7(b) shows the picture obtained using the suppression system; the dark rings against the noise background can be clearly seen where side-lobe echoes have been suppressed. The outline drawing of Fig. 7(c) gives a key to the echoes in Fig. 7(b).

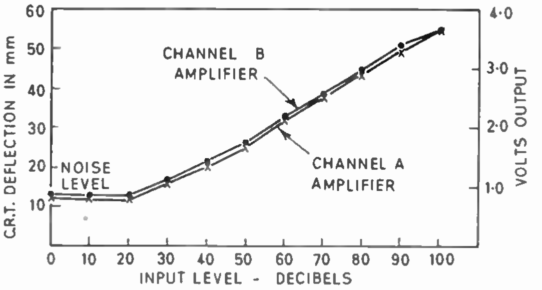


Fig. 5. Input-output curves of logarithmic amplifiers for suppression system.

Troublesome pulse interference was experienced on the radar at Southsea from the radars of ships entering and leaving Portsmouth and Southampton harbours. The authors were impressed by the effectiveness of the system in eliminating this interference. Figure 8 shows a comparison of the pictures received (under conditions of pulse interference) on the normal radar system, compared with those from the side-lobe suppression system, for the three range scales, 12 n.m. and 30 n.m. and 60 n.m. In Fig. 8(f) the discrete arcs may be seen over which main-lobe interference is still received.

It is estimated that, by mounting the head amplifier of the omni-directional channel B directly below its aerial, the gain of the omni-directional aerial was effectively raised to more than 6 dB above that associated with the worst side-lobe of the radar aerial (−24 dB). This enabled about half the noise at the output of the omni-directional channel to be removed by base clipping in the process of matching the worst side-lobe signal at the output of the omni-directional

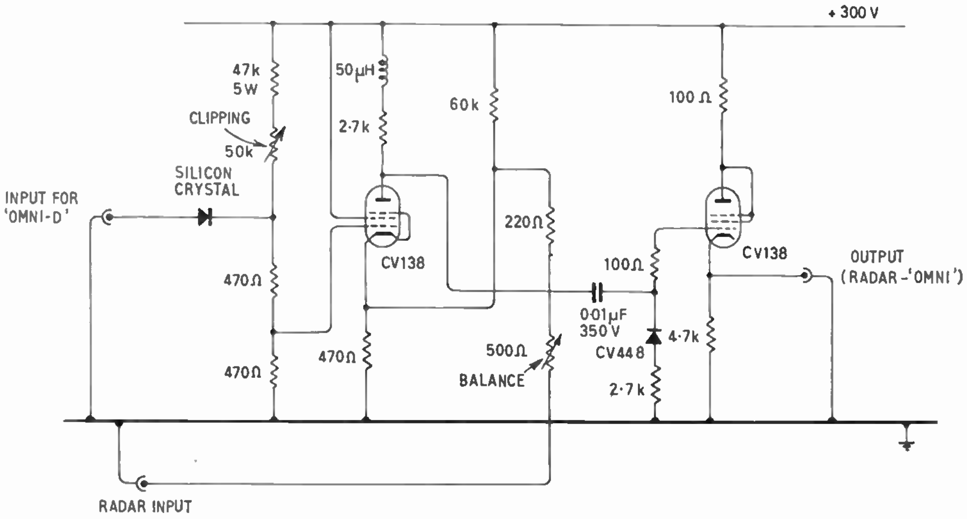
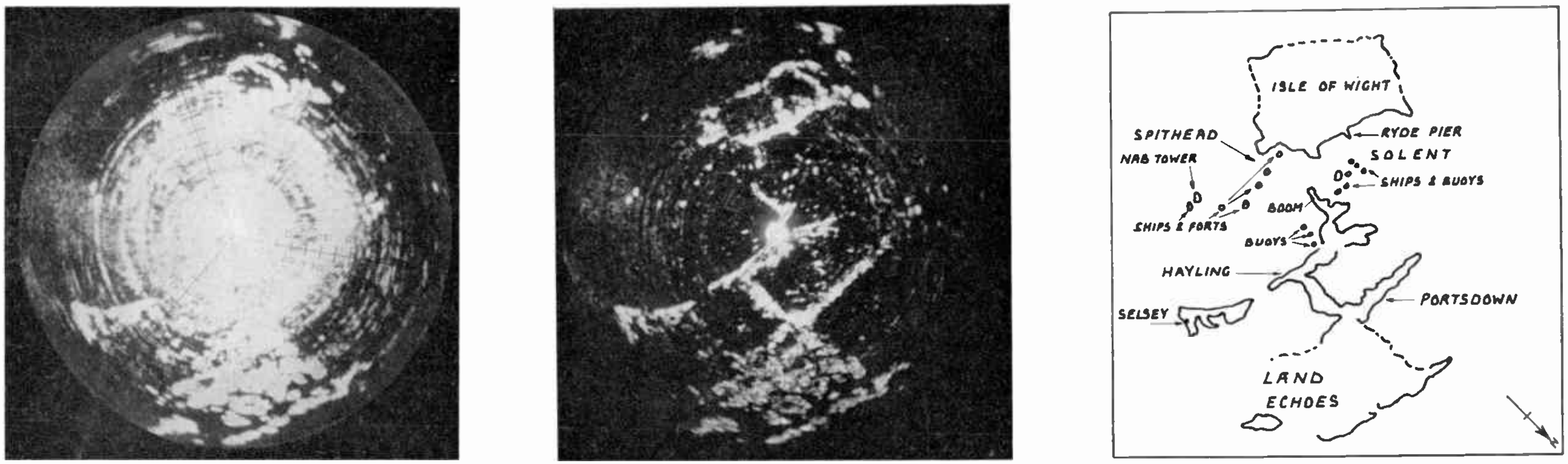


Fig. 6. Circuit diagram of subtraction unit.

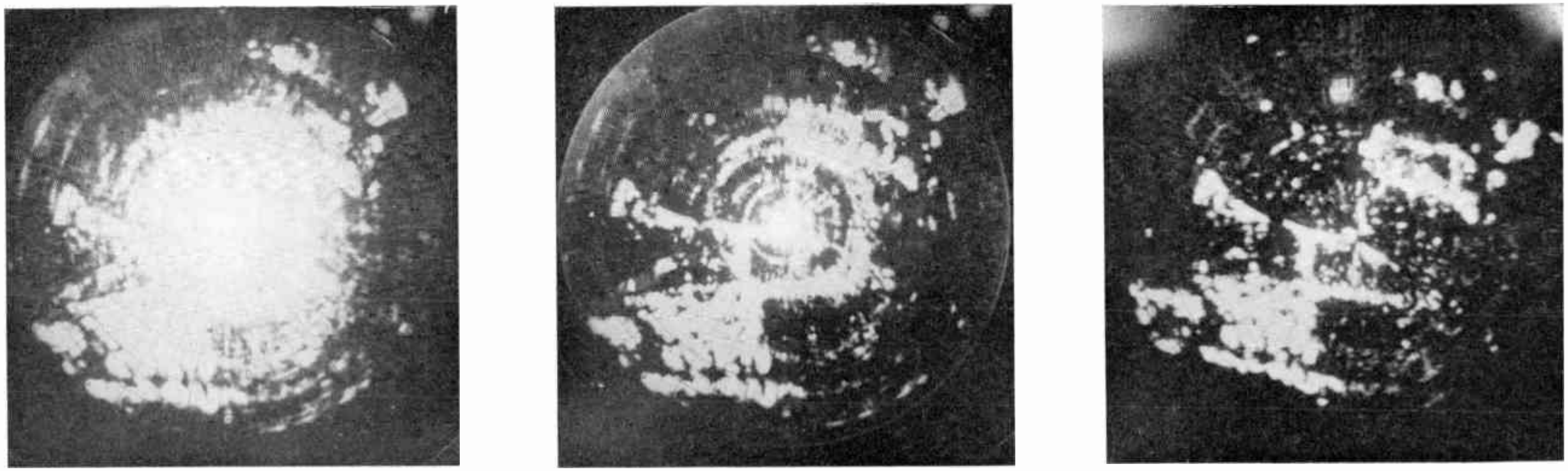


(a) Normal p.p.i. picture from Naval radar without swept-gain (range scale 12 miles).

(b) P.p.i. picture with side-lobe suppression system in operation.

(c) Key to Fig. 7(b).

Fig. 7. P.p.i. display without a swept-gain facility.

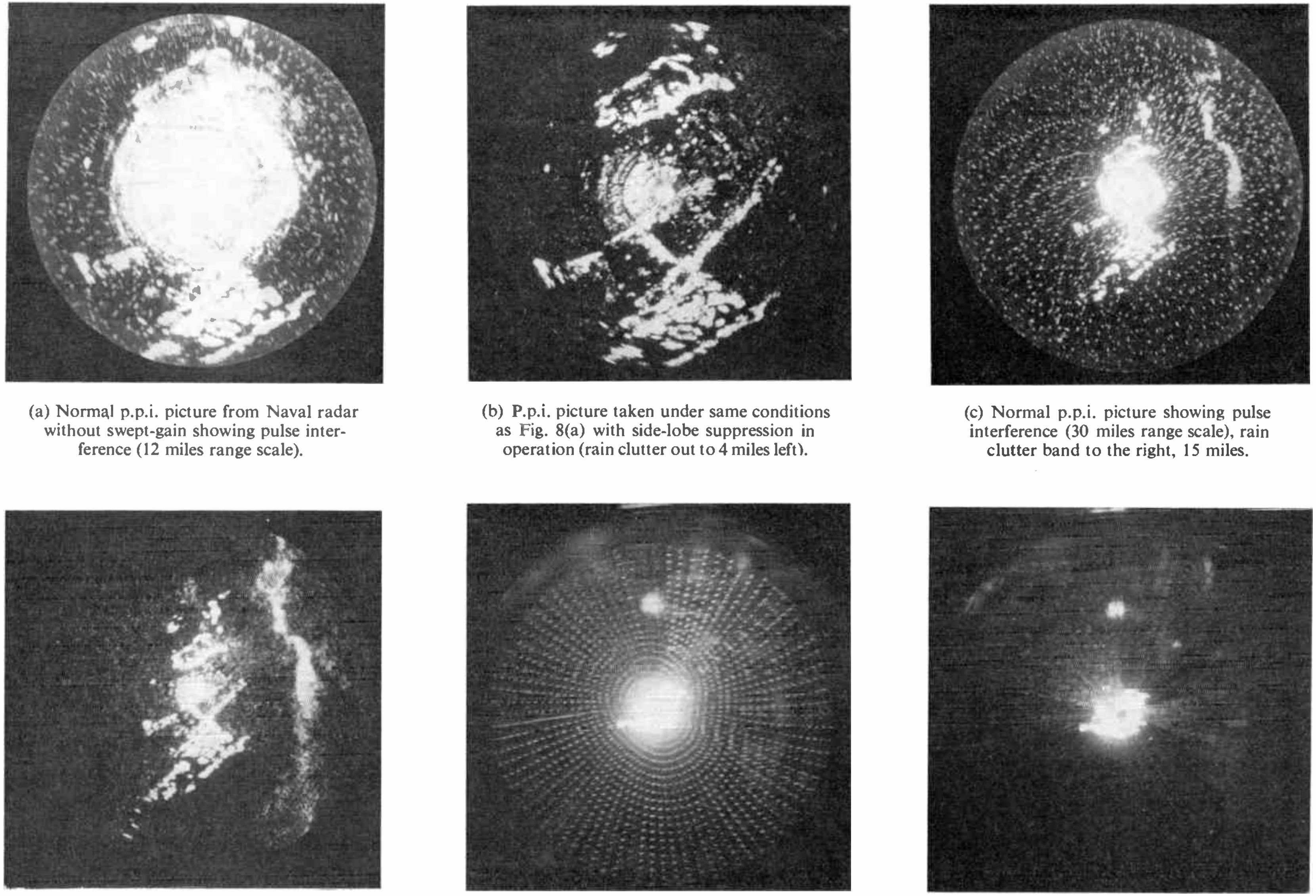


(a) Without swept gain (12 miles range scale).

(b) With optimum swept gain setting for tracking an aircraft echo.

(c) With side-lobe suppression in operation.

Fig. 9. Type B display.



(a) Normal p.p.i. picture from Naval radar without swept-gain showing pulse interference (12 miles range scale).

(b) P.p.i. picture taken under same conditions as Fig. 8(a) with side-lobe suppression in operation (rain clutter out to 4 miles left).

(c) Normal p.p.i. picture showing pulse interference (30 miles range scale), rain clutter band to the right, 15 miles.

(d) As Fig. 8(c) but with side-lobe suppression in operation. Rain clutter band extending.

(e) Normal p.p.i. picture showing pulse interference (60 miles range scale).

(f) As Fig. 8(e) but with side-lobe suppression in operation.

Fig. 8. Comparison between pictures received on normal radar system and those from the side-lobe suppression system.

channel to its counterpart at the output of the radar channel. The resultant signal/noise reduction, caused by the presence of the second receiving channel was measured, under these conditions, as about one decibel. The second channel can obviously be 'pulsed off' at long ranges where side-lobe echoes are not experienced.

The photographs shown in Fig. 7 (a) and (b) were taken from a p.p.i. display in which a 'swept gain' (s.t.c.) facility was not available. It was thought interesting to compare the performance of a display, in which an optimumly adjusted swept-gain system was in use, with that of a display fed from the side-lobe suppression system. An experimental display (referred to hereafter as Type B display) was available which incorporated a swept-gain control. In this display the sawtooth waveform of the swept-gain voltage had been designed to have the correct slope for maintaining the received echo strength of an approaching target constant at the c.r.t. as the range closes. The law was checked carefully in the following manner. The output from an i.f. signal generator (pulsed by a variable range strobe) was connected to the receiving channel. With the strobe set successively at a number of descending values of range, the signal generator output was increased in accordance with the inverse fourth power of the range (free-space radar equation). A preset control was adjusted until the swept-gain law maintained the signal from the generator at constant intensity on the display for all ranges. The amplitude of the swept-gain voltage was then adjusted so that the echo of an approaching aircraft target just reached the limiting level of the p.p.i., remaining so into zero range. Under these conditions the background receiver noise reached its normal painting level on the p.p.i. at a range of about 14 miles.

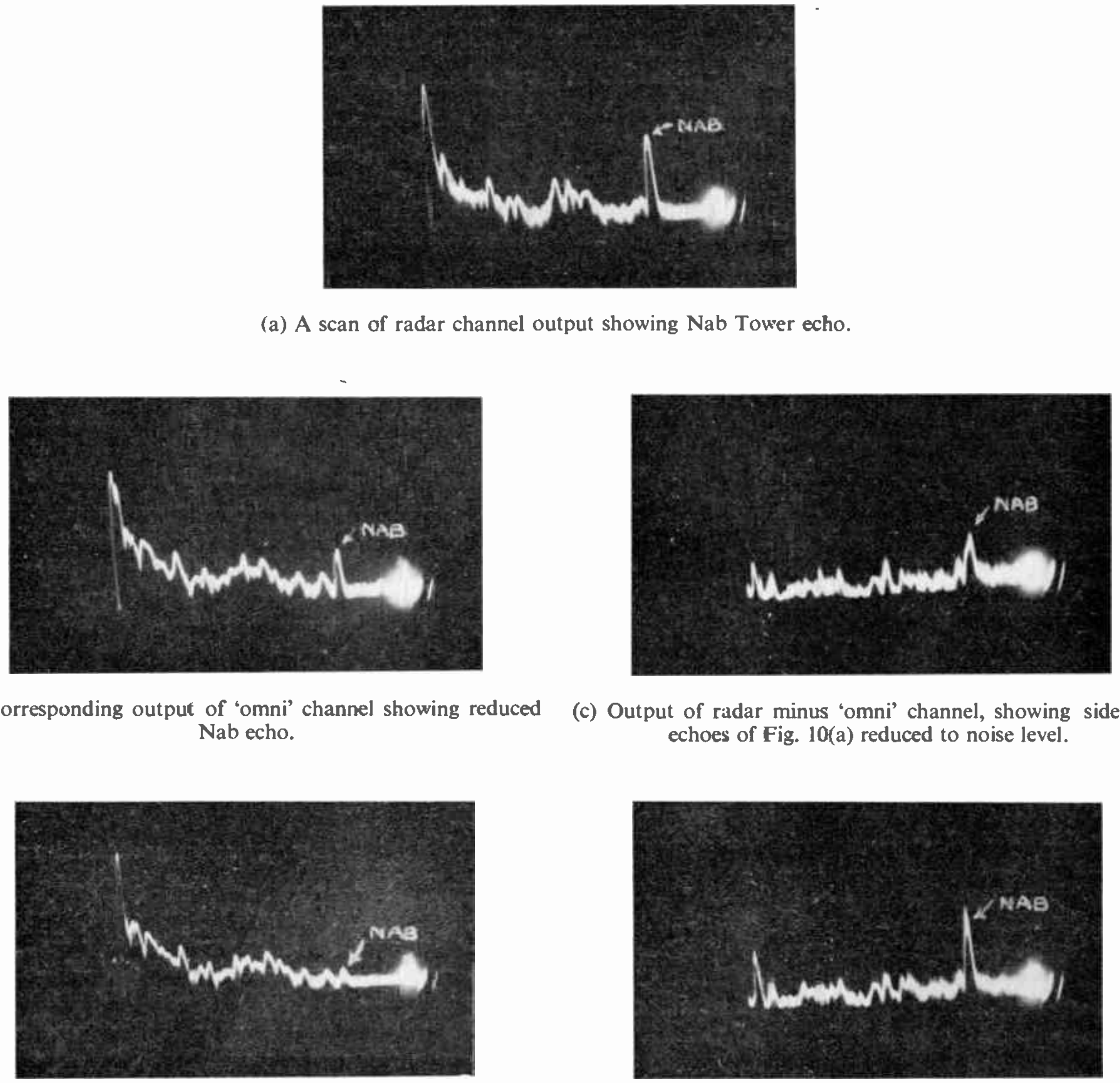
Figure 9(a) shows the picture on this Type B display without swept gain and Fig. 9(b) shows the picture with the swept gain correctly adjusted as explained above, for closing aircraft runs. Figure 9(c) shows the picture on the same display without swept gain, but with the side-lobe suppression system in operation. Figure 9(c) cannot be compared point for point with Fig. 9(b) because there was an interval of an hour or more between the exposures, and the range scale of the p.p.i. is slightly more open for Fig. 9(b) than for Fig. 9(c). Unfortunately the c.r.t. spot in this experimental display could not be sharply focused. This in turn made focusing of the camera difficult, and in the case of Fig. 9(c) the camera was clearly out of focus. The faint specks in the background of Fig. 9(b) do not represent inherent noise (since this is completely suppressed by the swept gain control) but interference pulses from other radars. It will be noticed from Fig. 9(b) that the optimum setting of the swept gain control still allows side-lobe echoes to paint out to a range of about 5 miles. Further gain suppression would cause reduction in the intensity of paint of the echo from the approaching aircraft.

4.2. *Quasi-omni-directional Aerial*

Practical work with this system was restricted to determining to what extent it was possible to produce the cusp in the omni-directional radiation diagram for a fixed bearing of the radar aerial, and the degree to which this could be maintained over a complete revolution of the radar aerial.

Since the omni-directional aerial does not rotate it is essential, if phase cancellation is to be maintained over a complete revolution of the radar aerial, to mount the omni-directional aerial so that its vertical axis is coincident with the axis of revolution of the radar aerial. To achieve this condition a plumb-line was hung from the omni-directional aerial (coincident with its vertical axis); the plumb-bob was located as near as possible over the centre of revolution of the radar aerial. No fine adjustment was available on the mounting of the omni-directional aerial, and its supporting structure was moved about by levers at its base to achieve this centring. It is not surprising that the nearest achievable coincidence of the two axes was about $\frac{1}{4}$ in ($0\cdot06\lambda$) by this method. Departures of at least 20 deg of phase from perfect phase opposition could therefore be expected as the radar aerial rotated.

For setting-up the phase and amplitude cancellation the radar aerial was stopped on the bearing of the Nab Tower (Fig. 7(c)). Figure 10(a) shows the A-scan picture obtained from the radar channel A; the Nab Tower echo is marked. Except for any shipping which may be passing through the same bearing, only sea intervenes between the Nab Tower and the radar site at Eastney, and all the intermediate echoes are therefore, with a few possible exceptions, side-lobe echoes. Figure 10(b) shows the corresponding picture received on the omni-directional aerial channel B, the phase and amplitude link between channels A and B being disconnected. The side-lobe echoes are clearly being received at about the same strength as in channel A (Fig. 10(a)). Figure 10(c) shows the result after subtracting the output of channel B from channel A. The Nab echo is now reaching a limiting level of about 20 dB above noise. The side-lobe echoes are all reduced to noise level. For Fig. 10(d) the phase and amplitude link has been introduced between channels A and B and progressive adjustments of phase and amplitude made in an endeavour to cancel the Nab echo in channel B. Comparison of Fig. 10(b) with Fig. 10(d) shows that a good degree of cancellation has been obtained. The side-lobe echoes have not been affected, i.e. a cusp has been effectively introduced in the omni-directional pattern coincident with the main beam of the radar aerial. Subtraction of the



(a) A scan of radar channel output showing Nab Tower echo.

(b) Corresponding output of 'omni' channel showing reduced Nab echo.

(c) Output of radar minus 'omni' channel, showing side-lobe echoes of Fig. 10(a) reduced to noise level.

(d) Output of 'omni' channel with phase and amplitude cancellation of Nab echo to produce 'quasi-omni-directional' aerial.

(e) Output of radar minus 'quasi-omni' channel showing side-lobe cancellation with higher limiting level of Nab echo.

Fig. 10. Traces showing echoes from the Nab Tower.

omni-directional channel B from radar channel A now gives the result shown in Fig. 10(e) where the limiting level of the Nab Tower echo is about 35 dB above noise level. The phase and amplitude cancellation adjustment was quite simple, and held surprisingly well with time, in view of the continual vibration of the omni-directional aerial support by wind.

To see how well the adjustment would hold with rotation of the radar aerial, the aerial was set on to targets at other discrete bearings over the 360 deg of revolution, and the readjustment necessary to the phase changer to maintain cancellation for each bearing was noted. The results are plotted in Fig. 11, with the aspect of the omni-directional aerial superimposed. It will be seen that cancellation can be held to ± 15 deg of phase over two discrete arcs of about 60 deg, but that large and rapid variations of phase are experienced elsewhere. There are several possible reasons for this; first, the rotating joint of the radar aerial forms part of the phase loop and this was an early type of joint where rapid phase variations with rotation are possible. Secondly, the omni-directional aerial is not a circularly symmetrical structure and the phase axis is likely to shift in the direction of the fins when signals are being received from those directions. Thirdly, there is the difficulty already referred to of

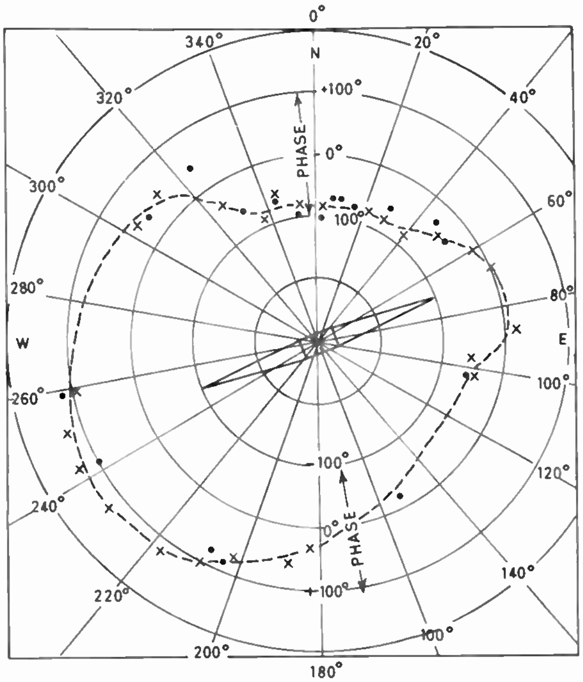


Fig. 11. Plot of phase variation with bearing, required to maintain cusp in 'omni' polar diagram over one complete revolution of radar aerial.

ensuring that the rotational axis of the radar aerial and the vertical axis of the omni-directional aerial coincide.

The above three difficulties could all be avoided by mounting the omni-directional aerial rigidly on the radar aerial, so that it rotates with the radar aerial, and feeds to channel B via an additional rotating joint. The necessary fraction of power from the radar channel for the cancellation process could be fed to the omni-directional channel by a pick-up probe inserted in the primary feed to the radar 'cheese' reflector, and the phase and amplitude adjustment would then be affected on the aerial side of the rotating joints, so that variations of phase in the rotating joints would be of no consequence.

In any system involving phase cancellation of this type, installed in a ship, stabilization of the aerials along either a single vertical axis or constantly displaced vertical axes (both aerials rotating) would probably be essential. For instance, if in an unstabilized system the omni-directional aerial were mounted 1 ft above the radar aerial, a roll even of ± 5 deg would involve a path difference (from the target to the two aerials), of about one-quarter of a wavelength at S-band, with the radar aerial pointing athwartships. In these circumstances there would be a reinforcement of the signal in channel B rather than a cancellation.

5. Suppression of Target Echoes by Over-cancellation of Side-lobe Echoes

As stated in Section 2 there is a possibility of suppression of targets by over-cancellation. This section examines this problem quantitatively and compares the proposed system with conventional techniques.

5.1. *Behaviour of the Logarithmic Receivers*

The output voltage is related to the input signal power S by

$$v = k \log_{10} S,$$

where k is a constant determined by the amplifier design. In the amplifiers used in the experiment k had a value of 0·4. To simplify the following discussion, however, it is convenient to take k equal to 10. The output voltage then becomes equal to the level of input signal power, S, measured in decibels relative to some convenient reference; the replacement of the equation by a proportionality would not affect the argument.

This law will not, however, be followed indefinitely. When the signal power is smaller than the input power of head amplifier noise, the latter will determine the output level. If this noise power is used as our reference in measuring the signal power level, the receiver output will then be approximately equal to either S or 0 whichever is greater. Furthermore, if there are two signals present simultaneously with signal/noise power ratios in decibels of S_1 and S_2 respectively, then only the larger will determine the receiver output. We shall write this as

$$v = \left\{\begin{matrix} S_1 \\ S_2 \\ 0 \end{matrix}\right.$$

it being understood that only the largest term is used.

5.2. *Operation of Normal Radar*

In the normal air search radar the signals of interest will come from aircraft targets with echoing areas in the range 1 to 100 m^2. The echoes from land can be very much greater, depending on the radar pulse duration and beamwidths. Echoing areas up to 10^7 m^2 can occur, although 10^6 m^2 are more likely. We write L as the ratio of land echo to receiver noise, measured in decibels, when the beam is trained on it in azimuth, while S is the corresponding signal/noise ratio for an aircraft target. Typically $L-S$ may be about 50 dB.

If in a normal radar, using a logarithmic amplifier, the beam is directed at the aircraft but not at the land responsible for L, then the receiver output will be

$$\left\{\begin{matrix} S \\ L-2D \\ 0 \end{matrix}\right.$$

where D is the side-lobe level of the radar in decibels, relative to the main lobe, for the particular angular separation. We shall assume that the smallest value of D found is 24 dB as for the actual aerial used in the experiment. D may in general be some 30 or 40 dB.

In the absence of swept gain (s.t.c.) which will be discussed later, the wanted signal will only be distinguishable if there is no land echo present; this arises from the very limited dynamic range of cathode-ray-tube displays. The condition for a visible signal is therefore

$$L < 2D$$

Figure 12 presents a diagram of the plane of S against L. It may be observed that only in the region to the left of the chain-dotted line AB can the target be seen. Although the target signal is actually above the land signal in the region BAF, it will not be detectable there.

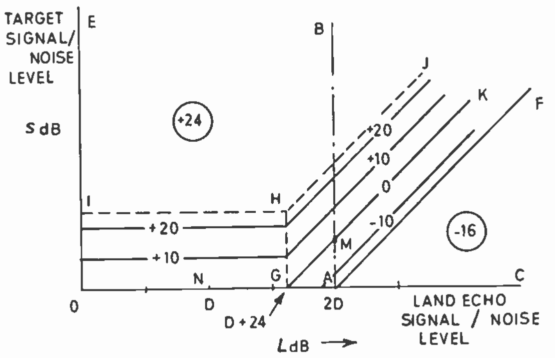


Fig. 12. Plane of target versus land signals.

It might be expected that the target signal could be made detectable in the region BAF, where it exceeds the land signals, by including a high-pass filter between the receiver and the display. The land signal will, however, have a rather large variation in amplitude even after such a 'differentiation'. Even with a uniform distribution of scatterers the probability distribution of the land echo will have a similar form to that of the receiver noise, but with a markedly greater correlation over the beamwidth. When the considerable inherent variation of land echoing area is included, it is found that the target is likely to be undetected unless its level relative to the land is very high. This technique has been described in an earlier paper.[5]

5.3. *Normal Radar with Swept Gain*

When a swept-gain facility is incorporated, the receiver gain may be deliberately varied during the early part of the transmission interval with a law corresponding to the expected law of signal variation with range, as described in Section 4.1 and demonstrated in Fig. 9(b). To avoid suppressing the smaller targets it is necessary to adjust the law to correspond with the smallest expected target, say 1 m^2. Targets may therefore be lost if the land echoes exceed $2D$ decibels relative to 1 m^2. This can occur as may be seen from the presence of many residual echoes on Fig. 9(b).

Adjustment of the swept gain to this ideal setting is difficult, particularly when the propagation conditions, including vertical beam shape and sea reflections, lead to appreciable variation of target signals, and there is a significant risk of unwanted suppression of both weak and strong signals.

5.4. *Radar with Side-lobe Suppression Sytsem*

The outputs of the main and omni-directional channels may in this case be written as

$$\text{Main} \begin{cases} S \\ L-2D \\ 0 \end{cases}$$

$$\text{Omni} \begin{cases} S-24 \\ L-D-24 \\ 0 \end{cases}$$

where the system has been adjusted to a worst side-lobe level of 24 dB.

The dotted line in Fig. 12 indicates how the regions in which the target or land signals preponderate in the omni-directional channel. Above IHJ the target signal is strongest; to the right of GHJ the land signal is strongest, below IH, the omni signal is zero.

We then subtract the two channels. In IHJ, where target signals dominate both channels, a steady level of +24 dB is obtained and signals are seen. To the right of AF, where land dominates both, a steady negative level of $(24-D)$ dB results. This will be clipped in the amplifiers and display system. In between these lines the output depends on S and L and contours of output signal are shown in the figure, which is drawn for $D = 40$ dB. It will be seen that the output is positive above the line OGK; in this region the targets should be visible.

It may be added that if the quasi-omni-directional aerial is used the line IHJ is merely moved upwards on the diagram, with a consequent increase in peak output signal.

Compared with the original basic radar, all signals in the region enclosed by BMK have been rendered visible. However, it is at the cost of losing those in the triangle GMA. This area is small compared with that covered by signals. It can be seen, moreover, that the line GK actually corresponds to

$$L-S = D+24$$

which equals 64 dB in the case shown in Fig. 12. Comparing this with the typical value of $L-S$ of 50 dB quoted above, we see therefore that signals of interest can generally be expected to be above the line GK. The better the side-lobe levels the less the over-cancellation losses to be expected. Thus side-lobe suppression has rescued a majority of targets from oblivion at small cost. It also has the fundamental advantage of removing *all* side-lobe echoes from the display.

Part of the former advantage can be obtained from the earlier mentioned techniques. However, the 'differentiation' system can still be expected to lose targets to the right of GK, and indeed owing to the variation of the land echoing area it may lose more if the display is not adjusted correctly. The swept-gain system even when correctly set, is as bad, and in practice, as the setting is an operator adjustment, is substantially inferior to the result obtained automatically by the side-lobe suppression scheme. Neither of these systems will suppress all the side-lobe echoes and these spurious targets may well be a most serious interference to target detection.

6. Performance of the System in the Presence of Jamming

We now consider the behaviour of a radar fitted with the side-lobe suppression scheme in the presence of s.h.f. jamming. We shall again write the radar signal/noise ratio on an aircraft target as S dB, when the radar is pointing at it. We write the jamming/noise ratio in the radar as J dB when the radar is on the bearing of the jammer.

We have to consider two cases: first, when the target and jammer are on the same bearing, and secondly, when they are separated by at least an aerial beamwidth. Jamming in these situations is usually described as 'main-lobe' and 'side-lobe' jamming respectively. The tactical value of side-lobe jamming is obviously much greater than main-lobe jamming, as the interference is then not confined to a relatively narrow sector, but may completely nullify the radar's function. We shall attach greater importance therefore to the performance of the system in the presence of side-lobe jamming.

6.1. *Main-lobe Jamming with Normal Radar*

If no precautions are taken in the radar receiver, the onset of jamming will lead to saturation in the receiver. In this case the requirement for target detection can be put crudely as

$$J<0$$

This is, however, an unnecessarily severe criterion, as precautions can usually be taken to prevent the saturation. Methods include manual gain control, automatic gain control and the use of a logarithmic i.f. amplifier with its output differentiated.[5] Under these conditions the requirement for detection becomes:

$$S>J$$

6.2. *Main-lobe Jamming with Side-lobe Suppression Radar*

The outputs of the two channels, using the previous convention, will now be

$$\text{Main}\begin{cases} S \\ J \\ 0 \end{cases}$$

$$\text{Omni}\begin{cases} S-24 \\ J-24 \\ 0 \end{cases}$$

Provided J is less than 24 the system will behave identically to a normal radar. Again, it is desirable to avoid display saturation. This can conveniently be obtained by introducing a high-pass filter (a differentiation circuit) between the subtraction unit and the display; alternatively, the circuit may be placed between the two valves shown in Fig. 6.

When J exceeds 24, however, we observe that the signal/jammer ratio is identical in both channels. When subtracted, a steady level of +24 dB results. Differentiation of this gives a zero result. Thus the saturation characteristic shown in Fig. 1(c) leads to a loss of signal in main-lobe jamming when $J>24$ dB, although it is visible in a normal radar fitted with the anti-jamming devices mentioned above.

When a quasi-omni-directional aerial is used this saturation is delayed by the magnitude of the cusp in the radiation diagram.

6.3. *Side-lobe Jamming with Normal Radar*

The jamming signals are in this case reduced relatively by D. Thus our condition for seeing the target, if we can design to prevent saturation effects, becomes

$$S > J-D$$

Due to the variability of D in azimuth it is not practicable to use a manual gain control to avoid saturation and the automatic methods mentioned above are necessary.

6.4. *Side-lobe Jamming with Side-lobe Suppression Radar*

The outputs of the two channels now become

$$\text{Main}\begin{cases} S \\ J-D \\ 0 \end{cases}$$

$$\text{Omni}\begin{cases} S-24 \\ J-24 \\ 0 \end{cases}$$

We observe that as D is greater than 24 dB the jamming signal will first enter the omni channel. Thus the subtracted signal, if S is less than $J-24$, can be brought to a negative result and lost. As J increases it will later enter the main aerial, after which the signals work from a steady negative 'baseline' of $24-D$. A maximum suppression of $D-24$ dB can therefore occur.

This effect can be illustrated by Fig. 12 as the channel outputs may be observed to be the same as in the previous land echo discussion if we put J equal to $L-D$. Thus the diagram will apply with J measured on the abscissa from point N.

However, this situation may simply be cured by incorporating the differentiation circuit in the subtractor output, as mentioned before. The circuit will not now pass the negative level and the loss will be avoided.

Some degradation of performance by a few decibels will, however, occur for the values of J between 24 dB and D dB, because the jamming noise in the omnidirectional channel and the receiver noise in the radar channel will be uncorrelated, and the resultant noise level will therefore be increased. The effect should not be serious, however.

When J is greater than D and the jamming appears in both channels, it is interesting to note that the two signals will be appreciably correlated. With complete correlation, the jamming signals would then subtract completely. Provided the differentiation is included to remove the bias, this would leave the signal (and the original radar noise) against a non-noisy background. This would permit an increase in video gain in the display and suggests a prospect of detecting signals which are actually less than the jamming. The amount of correlation will depend on the match in amplitude, delay, and frequency response of the two channels, and this will set a practical limit to the extent to which the scheme will allow detection of signals below jamming. (The theoretical limit for perfect correlation would be close to the normal unjammed radar performance.)

This technique will not operate on main-lobe jamming as shown above, but has possibilities provided D is not zero. It may even be able to operate on the skirts of the main lobe. It would not be expected to work when more than one jamming signal is present.

7. Conclusions

The two versions of the system described give automatic suppression of side-lobe echoes and side-lobe interference on a primary radar. The simpler version can be applied to an existing radar with little additional complexity.

The suppression scheme should not lead to any important deterioration in the presence of side-lobe interference. Some over-suppression can occur with main-lobe interference; this can be reduced with the phase cancellation scheme.

8. Acknowledgments

The authors thank Mr. C. R. Shearston who performed the bulk of the experimental work and Mr. D. R. Roessler for the design of the omnidirectional aerial. The paper is published by permission of the Ministry of Defence.

9. References

1. J. Croney, "A New Proposal for Eliminating Side-lobe Echoes from Radar Displays", Unpublished Technical Note AX-55-1, January 1955.
2. J. Croney and P. R. Wallis, "Trials of a System Giving Automatic Suppression of Side-lobe Echoes and Interference on a Naval Radar", Unpublished Technical Note AX-56-3, April 1956.
3. O. Böhm and G. O. Kibbler, "An Aerial with Omni-direction Radiation in the Equatorial Plane for Horizontally Polarized Centimetric Waves", Unpublished Technical Note TX/51/11, September 1951.
4. J. Croney, "A simple logarithmic receiver", *Proc. Inst. Radio Engrs*, 39, pp. 807–13, July 1951.
5. J. Croney, "Clutter on radar displays", *Wireless Engineer*, 33, pp. 83–96, April 1956.

Manuscript received by the Institution on 26th February 1964 (*Paper No. 935/RNA 32.*)

DISCUSSION

Under the chairmanship of Dr. E. V. D. Glazier

Dr. D. H. Davies: Have the authors considered the possibility of using two halves of a linear array connected to provide sum and difference outputs—the sum pattern being used for the normal radar aerial and the difference pattern as the 'omni' aerial?

Dr. R. Benjamin: For a sin x/x main beam pattern, a pattern giving a good match to the side-lobes but having a null at the main-lobe peak can be obtained from the sum of two sin x/x patterns, whose axes coincide with the first nulls on either side of the main pattern. Unfortunately, this scheme gives a very poor match at the first two side-lobes.

The two opposite end elements of a linear array, combined *in phase-opposition* (to give a null on the axis of the array), will also theoretically match the nulls of the sin x/x array pattern, and can be made to give a good match to all the side-lobe peaks. This latter suggestion is due to Mr. T. E. Mitchell of A.S.W.E.

The Authors (*in reply*): The points raised by Dr. Davies and Dr. Benjamin are best answered if taken together. Our original work was concerned with systems which might be added easily to existing radar aerials and we did not therefore consider schemes which would involve redesign of the radar aerial. We accepted the fact that the omni-directional pattern could only be made to match the peaks (assumed equal) of the worst pair of side-lobes of the radar aerial and that some over-suppression would occur for lesser side-lobes. At that time Mr. Wallis did propose that perhaps a passive plate or rod reflector fixed to the top of the radar aerial, in such a way that it rotated round the static omni-directional aerial might produce a crude asymmetry in the omni-directional pattern, to match more nearly the falling levels of the outer side-lobes of the radar aerial, but this was not tried. From the consideration given in the paper we did not, and still do not, consider the likely extent of over-suppression to justify any significant complications.

Regarding the specific proposal of Dr. Davies, and assuming an ideal $\frac{\sin K\theta}{K\theta}$ pattern, the sum and difference signal amplitude can be shown to be

$$2\,|\sin K\theta . \cos K\theta / K\theta|$$
$$\text{and } 2\,|\sin K\theta . \sin K\theta / K\theta|$$

respectively. Thus for 50% of the angles the sum signal will exceed the difference, which does not meet our requirements. Regarding the proposals of Dr. Benjamin and Mr. Mitchell, schemes of this type do assume very good repeatability in the side-lobe structure which is rarely obtained in practice, especially when the radar aerial has to be mounted close to other superstructure. Over-suppression troubles might therefore still be experienced with these sophisticated techniques, which involve double rotating joints in the radar aerial and extra t/r circuitry.

Professor D. G. Tucker: Although it does not seem to be stated in the paper, I presume it is the *rectified* signals which are subtracted (e.g. in Fig. 1(a) or Fig. 2(a)).

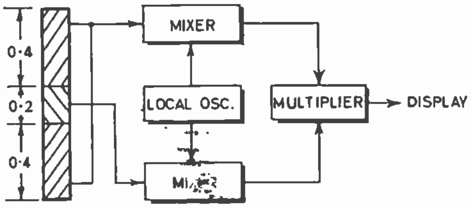


Fig. A.

Now it is easy to convert this system into a multiplicative one, and a particularly good arrangement—which has advantages in sonar, at any rate—would be to do away with the separate omni-directional aerial and use merely the central one-fifth of the longer aerial in its place, as in Fig. A. Thus only one aerial would be needed, although with two feeds. This system causes almost all the secondary responses to be negative, and those which remain positive, i.e. of the same polarity as the main beam, are at least 44 dB down.[†] So roughly the same result is achieved as in the authors' system.

The Authors (*in reply*): Professor Tucker is correct in assuming that the rectified signals are subtracted.

On Professor Tucker's proposal the remarks of the first paragraph of the previous answer apply again. Professor Tucker does not show how the three aerial sections are energized during the transmission phase, but a very unsatisfactory pattern would result if the centre one were left idle, so presumably some moderately complicated hybrid t/r circuitry would be involved in the proposal. We then still have residual side-lobes of 44 dB left on reception. As impulsive interference from other radars may well rise to 80 dB or more above noise, this side-lobe level would still be quite inadequate for the situation, and in fact, side-lobe levels of 40–50 dB can be obtained in simpler ways by suitably tapered amplitude distribution across an aerial aperture. We do not therefore regard this proposal as specially suited to the present context, but of course this is not to say that multiplicative processing in antennas is not of value for other requirements.

Mr. A. G. Halliday: I would like to put forward an alternative to the 'quasi-omni-directional' antenna just described which also prevents saturation by a jammer in the main beam. The method requires, however, a less critical positioning of the omni-directional element relative to the main antenna.

[†] See D. G. Tucker, "Multiplicative arrays in radio-astronomy and sonar systems", *J. Brit.I.R.E.*, **25**, p. 113, February 1963.

Referring to Fig. 1 of the paper, the logarithmic action of the receivers equalizes the fluctuations from a noise-like jammer in the two channels prior to subtraction. It effectively balances out the difference in gain between omni-directional and main antennae (for jamming much greater than receiver noise). The balance does not apply to the low frequency components. They are removed by a high-pass filter. This balancing action is, however, also responsible for the signal saturation effect which occurs when a jammer is present in the main beam.

Now clearly one easy method of preventing this saturation effect is to switch out the omni-directional channel whenever the total received power is substantially due to the main beam. This condition can be detected by an auxiliary comparison circuit which compares the outputs of the two amplifiers of Fig. 1(a) and operates an inhibiting gate, placed between amplifier B and the subtractor, whenever a threshold is exceeded.

The amplifier should ideally be d.c.-coupled to the comparator. Circuit tolerances are, however, fairly easily met since the main antenna gain will normally exceed that of the omni-directional channel by about 20 dB. The response time of the inhibiting circuits can usually be made shorter than the pulse width.

Since the system I have just described involves only video comparison of the two channels, co-location of the two antennas is not so critical as with the 'quasi-omni-directional' system of the paper. The required degree of co-location is within about a tenth of the radar range-resolution distance.

The Authors (*in reply*): We accept Mr. Halliday's comments. The quasi-omni-directional proposal is certainly of smaller practical application than the simpler scheme because of the critical positioning involved. This could only be avoided, as suggested in the paper, by fixing the omni to the radar aerial so that it rotates with that aerial and using a double rotating joint to bring down the additional channel. This arrangement ought to work over limited angles of target elevation. Certainly the scheme was of great interest to us as a means of producing a very unusual type of radiation pattern, but Mr. Halliday's video arrangement may well be a more convenient method of reducing the deterioration possible with main-lobe jamming. From the point of view of side-lobe echoes and side-lobe impulsive interference, however, the simple scheme appears to be quite adequate.

U.D.C. 681.32/33

A Hybrid Computer as a Training Simulator

By

P. A. R. WRIGHT, B.Sc.[†]

AND

D. S. TERRETT, B.Sc.[†]

Presented at a Joint Colloquium of the Computer and Radar and Navigational Aids Groups of the Institution and the Computer Design Group of the Institution of Electrical Engineers on "Simulators for Training Purposes" held in London on 21st April 1964.

Summary: This paper gives a general review of the various computational techniques employing conventional analogue, precision analogue and digital methods, and indicates the approach to the ideal hybrid techniques. It illustrates the present and future trends in training simulators, involving the use of precision analogue computers, fully integrated with digital control logic and high-speed arithmetic to form a true hybrid simulator.

1. Introduction

The digital computer has been adequately described elsewhere in this Colloquium and its advantages as a training simulator cannot be discounted. It can be an extremely accurate machine, and fast in its operation, but perhaps its greatest merit lies in its power of memory and logical operation. Nevertheless it is inherently a sequentially operating machine performing even the simplest mathematical calculation by a relatively complex series of additions and subtractions. The more complex the problem the larger the number of individual operations and the slower the complete computation, and hence the inability to deal accurately with the solution of any dynamic problems other than those in which the variables are changing slowly with respect to time. The computer, given time, may approximate by iteration to an accurate result, but extended time is rarely permissible in a training simulator. Another important limitation appears when program changes are introduced, as is often the case in an experimental or research phase of training; here the inflexibility of the machine becomes apparent. For fixed program static calculation the digital computer is ideal; for variable, fast, dynamic problem solution the computer is cumbersome, inaccurate and inflexible.

The conventional analogue computer is essentially a parallel machine in which there is a one-to-one relationship between the complexity of the problem and the complexity of the computer. A simulation problem of a dynamic nature can be successfully represented on an analogue computer subject only to limitations in accuracy and bandwidth of the computing elements and reliability. It must be remembered that the step-by-step or iteration process in the digital computer is now replaced by an electronic analogue of the system under study, where the general design, circuitry and component accuracy are of critical importance; if these problems are overcome satisfactorily, however, the resultant machine will be flexible, reasonably accurate and extremely fast in operation. Major disadvantages and limitations may appear with very large simulation problems such as in the solution of complex partial differential equations (heat transference and the like) since the basic analogue has no more than a single point memory capability and has limited power to perform logical decisions other than by human intervention.

The precision analogue computer is not a superficially improved analogue computer; it is designed from its concept as a machine which will reduce the inherent errors of analogue computers to a minimum within reasonable cost limits. Typically, computing resistors of 0·002 to 0·005% accuracy, precision polycarbonate (or sometimes polystyrene) capacitors, with amplifiers employing guard ring and screened circuitry, wide use of solid-state oven-mounted circuits for non-linear functions and gold-plated plug and prepatch panel contacts will be used throughout. High-bandwidths and solid-state switching permit high-speed iterative calculation and the inclusion of simple programming systems permit wide individual selection of time scales. The precision analogue is a well worth while achievement, but is nevertheless still limited by its small capacity to remember or decide, and by its sheer size when dealing with the more complex simulation exercises.

Table 1 shows speeds of typical machines of both types simulating a moderately sized six-degree simulation problem.

Table 1

Speed comparison for 3½–4 decimal precision	
Serial D.D.A.	0·001 c/s
I.B.M. 650	0·0005 c/s
I.B.M. 7090	0·25–0·5 c/s
Serial-parallel D.D.A.	0·1–0·5 c/s
Analogue Computer	5–30 c/s

[†] Electronic Associates Limited, Burgess Hill, Sussex.

2. The Hybrid Computer

On one hand, therefore, is the digital computer, severely limited as a simulator but having unique advantages in memory logic and fast arithmetic. On the other hand is the precision analogue computer, ideal for real (or condensed) time simulation but superficially limited in accuracy and severely limited in its memory and logic capability. It is to overcome these inherent limitations of the two approaches, that the hybrid computer was conceived. As far back as 1955 there was a partial acceptance of the problem, but unfortunately there appeared to be rivalry between the designers of digital and the analogue machines with each trying to prove the superiority of their approach. The first compromise came on I.C.B.M. trajectory studies with the ADDALINK (analogue/digital/digital/analogue/linkage) in which a large digital and a precision analogue computer were coupled by conversion and multiplexing interfaces, but this was of only limited success due principally to the totally different philosophies, programming methods, and operating techniques of the two computers.

Some minor moves to add analogue circuits to commercial digital computers were tried but fundamentally the digital computer was unsuited to dynamic scientific research problems. A diversionary move to the digital differential analyser met with little success; the resultant machines were costly, and combined not only the advantages but also the disadvantages of both analogue and digital techniques, although the d.d.a. did find a restricted field of application in the integral navigation type of problem.

In early 1963 the first integrated hybrid computer appeared, based in its design on the precision analogue computer with an integral parallel digital complement providing logic, memory and timing. Here was one machine, operating in parallel like the pure analogue and therefore as a true representation of the physical system, but with the ability to provide higher accuracies on non-linear operations, to tackle more involved and complex problems, and to store partial or complete computations.

Typical of the problems now solvable were:

(i) Simultaneous differential equations with widely different parameters producing both low and high frequencies in the solution.

(ii) High-speed solution of differential equations, employing prediction, iteration or optimization processes.

(iii) Sampled data or simulation of computer-controlled systems.

(iv) Perturbation analysis about slowly changing, precisely established solutions.

(v) Statistical analysis requiring repeated solution of differential equations, including Monte Carlo methods for deterministic problems; essentially a data storage and simple evaluation task around the solution of differential equations.

(vi) Filtering and processing continuous and sampled data for evaluation purposes.

(vii) Partial differential equations to be solved by serial integration procedures.

(viii) Ordinary differential equations accompanied by transport delays.

The precision analogue with a varying degree of digital equipment is the ideal approach to training simulation. The training simulator in its most elementary form will be fixed program, but to achieve any worthwhile benefit there must exist a means of introducing discrete or random variations. The effectiveness of the simulator will increase, as the introduction of variations in performance is facilitated, and ultimately the fast, flexible and easily varied analogue or hybrid computer will provide the ideal solution.

3. Some Applications for Hybrid Computers

A simulator at Edwards Air Force Base in the United States simulates the flight performance stability control during re-entry studies with the X-20 *Dyna-Soar* project and enables human intervention to be tested and assessed.

The U.S. Navy has a programme to provide a fleet of some sixty *Polaris* submarines each with a crew of 100 and a similar reserve crew of 100. The training of 12 000 men by conventional instruction at sea is uneconomic and hence at Charleston and at Pearl Harbour there are being installed two large precision analogue computers to simulate actual 'at-sea' problems in real time. A model moving-platform control room provides inputs to the computer and can pitch and roll in response to the computer outputs, which are fast enough to represent force-7 hurricane conditions.

All functions of water management are simulated and model ballast control tanks are flooded and drained in sympathy. Steering, course control, hovering, depth control and even the dynamic control aspects of missile firing are simulated. Emergency and disaster conditions are incorporated in this training simulator, which adequately provides a full six-equation-of-motion program and utilizes 140 operational amplifiers, 24 high-speed servos and 400 coefficient potentiometers. The computer is, of course, solid state in its design.

In the training of aerospace crews for manned space vehicles there is virtually no other method of simulation than the hybrid. One typical installation is concerned with the guidance and attitude control system for a space vehicle attempting to dock with a target satellite, where the analogue represents the motion of the space vehicle and the digital represents the control (see Fig. 1). This docking manœuvre might be manually controlled by the pilot, in which case the design of the control mechanisms must take into consideration the problems of the coupled man-machine system and the real-time analogue approach is clearly indicated. Similarly the selection of control parameters and the arrangement of feedback sensors is particularly easy when mechanized on such an analogue computer. There are, however, some features of the control system which are discrete or sampled, and therefore suggest the use of digital circuitry. The reaction jets which adjust the attitude of the vehicle are 'on/off' rather than continuous; the manual 3-axes control stick energizes in a series of pre-selected ways, some, but rarely all, of 16 jets. These features are, of course, readily established with the logic capability of digital equipment and the complete training and research simulator must therefore, be of a hybrid nature.

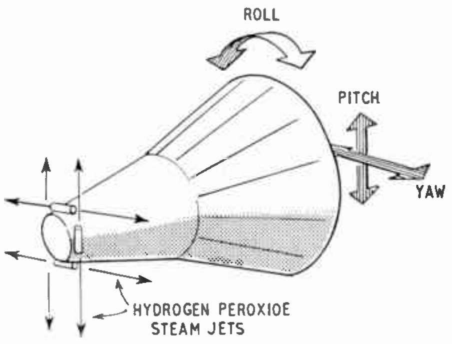


Fig. 1. Attitude control by reaction jets.

In the analysis of rocket parameters and staging times for maximum efficiency in orbit injection, the hybrid computer is again indicated. A multi-stage rocket's flight into orbit with the necessary considerations of aerodynamic forces, control systems and staging with the consequent abrupt and sudden changes in mass, inertia and thrust, is a complex simulation problem. High-speed simulation of a typical six-minute flight to orbit injection is essentially an analogue computation. Automatic iteration of system parameters and flight path designs, where very high computational accuracy is required, is most appropriately achieved using a digital computer. This kind of optimum-search problem is a common one having many and varied applications; its implied computation is the solution of differential equations plus a need for up-dating stored values and the making of logical decisions according to some criterion, strongly suggesting the hybrid computer approach.

Two further illustrations of the versatility of the hybrid approach are found in the successful solution of:

(i) Aerospace vehicle long-range trajectory simulations where high precision is essential in the terminal phase and real- (or faster than real) time scaling is necessary.

(ii) Real-time simulation of a radar system and actual terrain data in connection with high-speed terrain following vehicle simulation (Figs. 2 and 3).

The successful simulation of the motions of a space capsule on re-entry to the earth's atmosphere and the retro-rocket effects, is a vital step in the training of astronauts to control the craft in the event of automatic control failure, and in the designing of the automatic control gear itself. Providing a precision analogue computer is employed the problem is relatively straightforward in its solution, and such a computer of 600-amplifier capacity is operational at N.A.S.A., Langley Field, Virginia, to provide complete simulation of the conditions experienced in such an exercise.

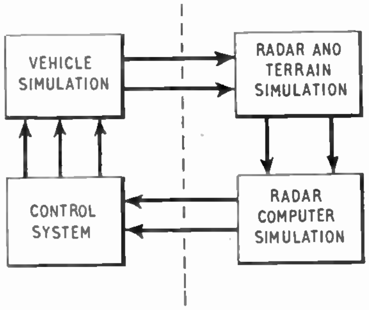


Fig. 2. Simulation of radar return signals.

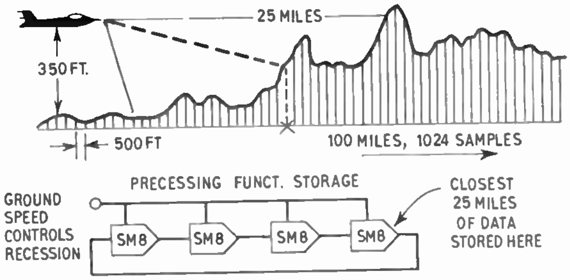

25 MILES
350 FT.
500 FT
100 MILES, 1024 SAMPLES
PRECESSING FUNCT. STORAGE
GROUND SPEED CONTROLS RECESSION
SM8
SM8
SM8
SM8
CLOSEST 25 MILES OF DATA STORED HERE

Fig. 3. Inverse function generation (repetitive).

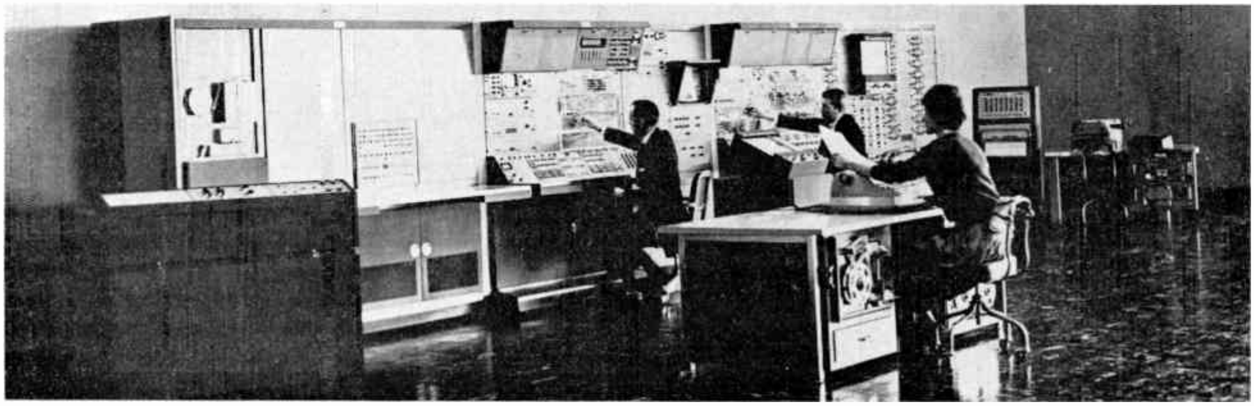


Fig. 4. HYDAC 2400 scientific hybrid computer.

Électricité de France have recently installed a computer system principally for the solution of heat exchanges and boiler problems. A German oil-field survey company uses a similar machine on oil-field geography, indicating that the hybrid computer is not confined only to flight and missile simulation, but could typically find application in operational training in other areas.

The importance of these typical applications lies primarily in the fact that none has needed a special-purpose computer. Complete flexibility is achieved by the use of a general-purpose precision analogue for high-speed computation, a digital-control logic system and a general-purpose high-speed digital computer for accurate slower computation such as flight path representation, and up-dating instructions such as engine failure. The only equipment special to the simulator lies in a separate control console as with a nuclear reactor trainer, or a separate model as with the cockpit and loading rigs of an aircraft simulator. This approach is undoubtedly more economical than the earlier special purpose simulators, since it permits changes in the configuration simply by rearrangement of the pre-patch panel, without the necessity of re-building the simulator.

4. Future Trends

Already the future trend of training simulators is clearly defined, but as technological advances in the precision analogue computer are leading to considerably higher computational speeds, so must the proportion of slower operating digital equipment increase in a truly balanced hybrid machine. The established hybrid computer is tending nowadays to require additionally a full fast-arithmetic general-purpose digital computer as an integral part of the precision analogue and digital memory and logic control system.

The HYDAC 2400 scientific hybrid computer is perhaps the only such machine in commercial existence today built as a completely integrated computing system (see Fig. 4). It incorporates three basic groups:

(*a*) A high-speed precision analogue computer with solid-state mode switching and complete solid-state non-linear elements.

(*b*) A fast arithmetic general-purpose digital computer.

(*c*) A digital operations system, acting as the control and communications centre for the complete hybrid system.

The HYDAC 2400 offers facilities for real and condensed time, continuous and discrete system simulation and makes possible the tackling of a new range of problems which have not previously been solvable by either analogue or digital techniques alone. Iterative calculations for heat and material balances, simulation of discrete phenomena, system optimization, high-speed prediction and heat transfer boundary problems are but a few of the problems capable of solution. The parallel programming of the digital logic section is so closely analogous to the conventional analogue approach that the major hurdle of different programming philosophies has been successfully overcome, and it is obvious that the next stage must be an even closer integration of the analogue and digital sections. Higher speeds, greater bandwidths, greater flexibility, are all natural trends in this third discipline of scientific computation.

5. Acknowledgment

The authors wish to thank Electronic Associates for permission to present this paper.

Manuscript received by the Institution on 21st April 1964. (Contribution No. 82/C72.)

U.D.C. 623.55.027: 681.33

Simulators for Manually Controlled Missiles

By

T. B. BOOTH, M.A.[†]

G. HARRIES, Ph.D.[†]

AND

E. J. STANNARD[†]

Presented at a Joint Colloquium of the Computer and Radar and Navigational Aids Groups of the Institution and the Computer Design Group of the Institution of Electrical Engineers on "Simulators for Training Purposes" held in London on 21st April 1964.

Summary: The training of missile operators in the manual guidance of line-of-sight command missiles is carried out in the operational environment in which a simulated missile is to be used against a real target. Using a joystick, the operator has to send command signals to keep the missile and target coincident in the visual field of view. Joystick signals are shaped in a transistorized electrical analogue unit to represent the missile and the output signals of the unit are used to precess a gyro carrying a small mirror. A spot of light is reflected from the mirror into the operator's field and represents the motion of the missile demanded by the operator. Such a system ensures that the missile line-of-sight is stabilized in space against the motion of the ship or aircraft.

The three simulators described are: (1) a pilot model for proving the system; (2) a shipborne trainer (ship to air); (3) an airborne trainer (air to surface).

1. Introduction

The firing of manually guided missiles for the purpose of training operators or initial proving of a weapon system is expensive. Much of the work in respect of both requirements can be achieved by simulators if the simulation is highly realistic. But realism can only be achieved if the following conditions are satisfied:

the simulation is performed in an operational environment;

the dynamic response of the overall system compares with that achieved in practice with the real missile.

These conditions are best reproduced by aiming a simulated missile viewed through an optical sight through which a real target (ship or aircraft in this context) is also viewed directly.

2. Pilot Model

In this pilot scheme a system was developed to prove the simulation technique and thereby provide a model for the further development of training equipment for service use. The system is essentially experimental in nature and no severe limitations were imposed by production potential or space available.

The system comprises a joystick, an electrical analogue of the missile, and an optical head (with recording camera).

A collimated spot of light is projected into the aimer's field of view to represent the missile. The position of this spot is controlled by the aimer using the weapon joystick whose signals are shaped in the analogue unit. The aerodynamic responses of the missile are reproduced in this unit with two supplementary units to initiate the firing sequence and the dispersion of the missile at launch. A schematic is given in Fig. 1.

The line of sight is stabilized in space by the gyro which eliminates the motion of the platform carrying the aimer's sight. In this experimental equipment the gyro had a 4-in heavy-duty wheel.

2.1. *Validation*

The proving or validation of the system was done as follows. Missiles were flown carrying telemetry equipment to record the performance of the missile in terms of wing deflection and acceleration in pitch and yaw. Simultaneously a recording was made on tape of the operator's joystick signals. These recordings were played into the simulator and the corresponding parameters recorded from the missile analogue. For the sample of rounds which were flown for the purpose good agreement was established between the telemetered records and simulator. A typical comparison is shown in Fig. 2.

This preliminary work justified the use of the system extensively in the acceptance trials of the *Seacat* missile at sea and in trials to define the tactical use of the weapon system. In this role the simulator operated successfully at sea for several months without fault.

3. Shipborne Trainer

There are three phases of a surface-to-air missile flight which have to be simulated: dispersion at launch; guidance during flight; terminal phase.

[†] Admiralty Surface Weapons Establishment, Portsdown, Portsmouth.

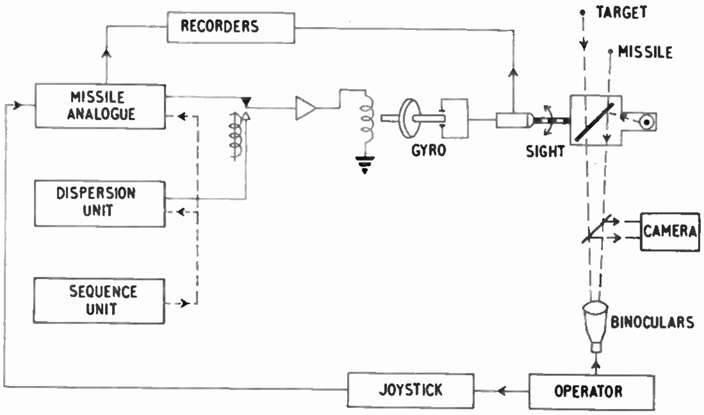


Fig. 1. Pilot simulator.

These are preset on selector switches by the controller before each training run, the settings being target range at launch, target speed selector, dispersion (in lateral and vertical planes) and launch button.

A short period after launch the missile 'spot' appears in a dispersed position in the operator's field of view and comes under joystick control. The joystick provides signals which are smoothed, shaped and processed electronically into currents to precess a line-of-sight gyro carrying a mirror which deflects a spot of light representing the missile. The sight uses the compact gyro unit from a gyro gun sight.

3.1. *Patching*

In the guidance phase the missile characteristics during flight are simulated by using four operational d.c. amplifiers patched as shown in Fig. 3.

These sections represent signal shaping; missile receiver, aerodynamics, and kinematics.

Consider the patching of amplifier 4 forming the kinematics circuit. The lateral acceleration (a) of the missile and the rotation rate ($\dot{\theta}$) of the line of sight between operator and missile are related by the equation

$$a = R\ddot{\theta} + 2\dot{R}\dot{\theta}$$

where R is the missile range.

As a transfer function this may be written

$$\dot{\theta} = a \, . \, \frac{\dfrac{1}{2\dot{R}}}{1 + \dfrac{Rp}{2\dot{R}}}$$

It has been shown in practice firings that the velocity $\dot{R}$ may be assumed constant thus reducing the expression to

$$\dot{\theta} = \frac{a}{2U\left(1 + \dfrac{tp}{2}\right)}$$

where U is missile speed and t is elapsed time of flight.

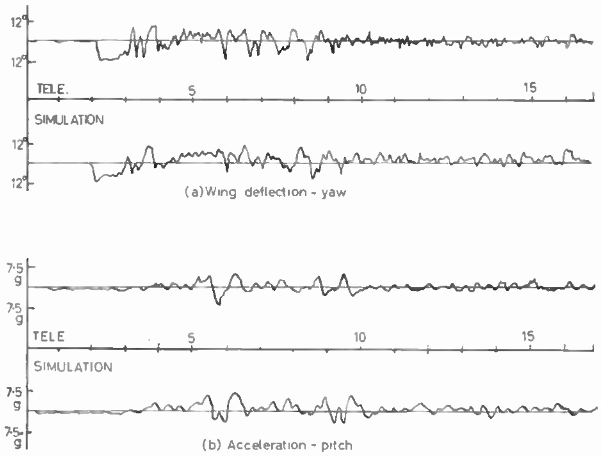


Fig. 2. Flight records.

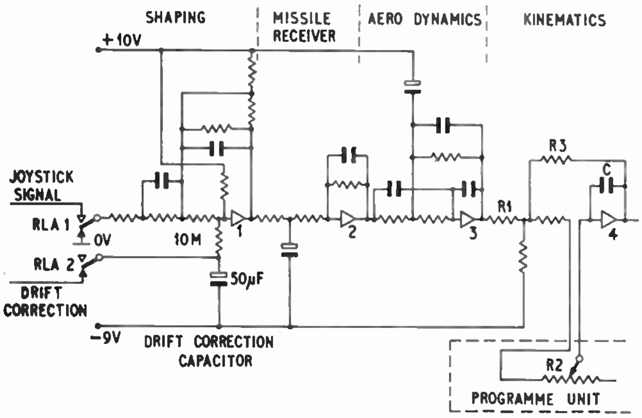


Fig. 3. Amplifier unit patching. (One plane only.)

This expression was set up as shown in Fig. 3 to give the transfer function of

$$\frac{V_{\text{out}}}{V_{\text{in}}} = -\frac{R_3}{R_1} \cdot \frac{1}{\left(1 + \left[R_3 + R_2\left(1 + \frac{R_3}{R_1}\right)\right] Cp\right)}$$

In this expression the value R_2 is increased at the rate appropriate to the $tp/2$ term. This is a lag with a response becoming progressively more sluggish with time.

3.2. *Drift Correction*

Associated with the patching of the first amplifier is a 50-μF capacitor in series with a 10-MΩ resistor connected to the virtual earth. Between training runs the input signal is clamped to zero volts and this capacitor is charged to a potential which is a function of the error voltage produced at the output of the simulation. Once the launch button is pressed a relay connects the joystick to the amplifier input and disconnects the capacitor from its charging circuit. During the time of flight this capacitor partly discharges through its long time-constant but provides adequate drift correction. This scheme eliminates the need for zero-set controls.

3.3. *The Amplifiers*

The basic operational amplifier was designed with the minimum number of components (see Fig. 4). The circuit has an open-loop voltage gain of about 60 dB and may be patched with impedances of the order of 100 kΩ. The bandwidth at a voltage gain of 10 is not less than 200 c/s and at this frequency with unity gain the phase shift is less than 5 deg.

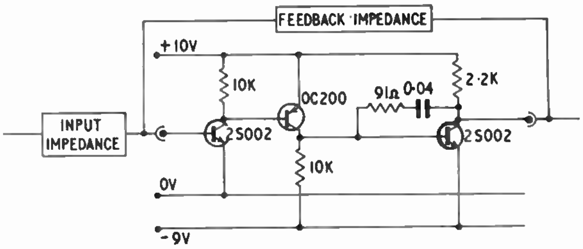


Fig. 4. Basic operational amplifier.

The shaping and kinematics amplifiers, however, require a higher input impedance; for this an emitter follower stage precedes the amplifiers.

The simplicity of the basic amplifier lends itself to micro-miniaturization and such circuits have been made.

The amplifiers may be patched assuming a high input impedance and thus standard computing techniques may be used to re-patch the analogue if changes are made in the missile or its characteristics.

3.4. *System Schematic*

The block diagram of Fig. 5 shows the relation between the units discussed. The joystick feeds the amplifier units of both planes through the drift correction relays to the driver units which convert the voltage signals into the currents required to move the missile spot.

This movement in each of the four directions up, down, left and right is controlled by energizing the appropriate gyro coil. Substantial currents are required by the coils, each of which is powered by its

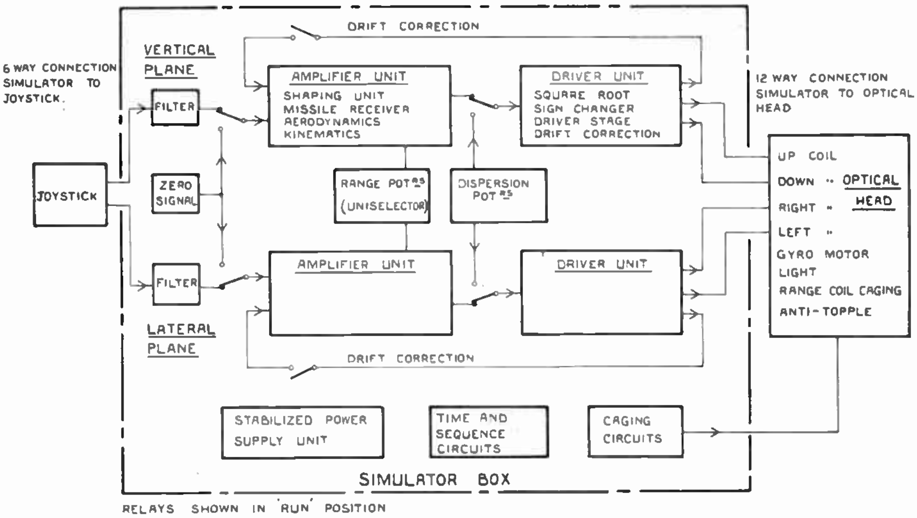


Fig. 5. Block diagram of the simulator.

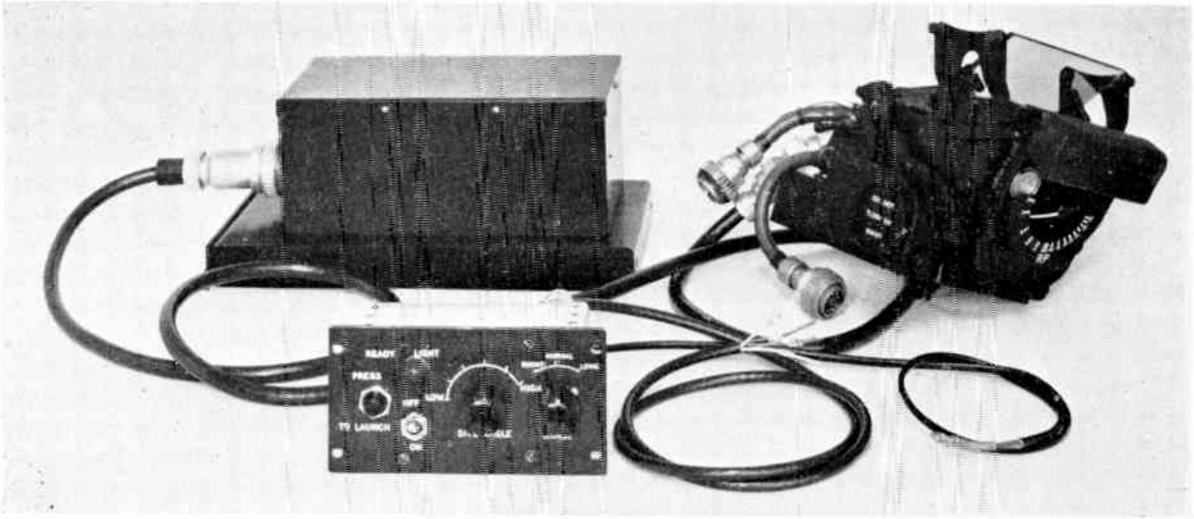


Fig. 6. *Bullpup* system.

own driver stage. The response of the gyro to a current through a coil is to precess in that direction at an angular velocity proportional to the square of the coil current. The square root of the signal from the kinematics section is therefore produced before being applied to the driver stage.

The other items in the block diagram are self-explanatory. The timing circuits are in fact programming the simulation. They operate the relays which uncage the gyro, connect the drift correction circuits and inject voltages to simulate the dispersion phase discussed earlier.

4. Airborne Trainer

This trainer was built for use with the air-to-surface missile *Bullpup*. This is a short range supersonic missile, and is steered visually down the line of sight to the target by the pilot of the parent aircraft. The need for training in the art of controlling both aircraft and missile simultaneously is obvious. The same design features are used in this trainer as in the ship-borne system and therefore this description will be confined to characteristics which are particular in the air role.

The principal limitation on equipment to be fitted in aircraft is space although weight, fire risk and vibration are also important. In order to conserve space not only was the fitted weapon joystick used but the existing gyro gun sight was adapted for use as an optical head.

The electronic equipment again uses the simplest analogue techniques capable of giving adequate accuracy. It is compact and is sited remote from the cockpit. In the cockpit are located the pilot's control box with settings for range and dive angle and the necessary indicators.

Special features of this equipment are the simulation of gravity drop (which the missile experiences in flight), dimming of the missile spot as range increases (to simulate the sensory effect of a missile in flight) and the extinction of the spot at the time estimated for the missile to reach its target. This latter feature gives the pilot an indication of his terminal accuracy.

5. Conclusions

The basic design employed in the simulators allows for versatility; the system is readily modified to keep abreast of modifications to the missile.

The simplicity of the system enables the equipment to be used in an operational environment without interference with the operational requirements. It is noteworthy that although transistorized electronic circuits reduce the size of the equipment the greater savings in space and cost derive from the use of a gyro as a means of stabilizing a line-of-sight.

Experience has shown that the value of these training aids is out of all proportion to their cost.

6. Acknowledgment

The authors wish to thank the Admiralty for permission to publish this paper.

Manuscript first received by the Institution on 21st April 1964. (Contribution No. 83/C73.)

U.D.C. 621.396.677: 621.396.965.8

An Analogue Polarization Follower for Measuring the Faraday Rotation of Satellite Signals

By

GOTTFRIED VOGT,
Dipl. Ing.†

Reprinted from the Proceedings of the Symposium on "Signal Processing in Radar and Sonar Directional Systems", held in Birmingham from 6th-9th July 1964.

Summary: Conventional methods for measuring polarization angles, as they are employed in measuring the Faraday rotation of satellite signals, determine the time elapsed between two zeros of the polarization fading. The instrumentation described in this paper, however, makes continuous measurement and recording possible by application of a fast electronic scanning and orientation method to the steerable polarization pattern of a crossed-dipole antenna. The information obtained by the intersection of the incident polarization and the pattern scan controls a servo system that in turn steers the system pattern in order to follow the incident polarization angle. System function, theory of operation, and test recordings that are obtained from v.h.f. satellite signals are presented and discussed.

1. Introduction

The instrumentation described in this paper was designed to measure continuously and record automatically the rotation angle of a linear polarization. When a signal is transmitted from a satellite to a ground receiving station, the radio wave propagates through the ionosphere. Because of the ionospheric electron content and the earth's magnetic field, a change of the polarization angle takes place. This phenomenon is known as the Faraday effect and is thoroughly treated in the literature.[1,2,3,4] In order to understand the system's performance better, some of the characteristic features of this type of wave propagation are reviewed. The electromagnetic wave radiated from the satellite antenna is assumed to be linearly polarized, with a polarization angle constant with respect to the north-south direction during the observation period.

All linear polarizations can be considered as a combination of two circular polarizations with the respective vectors rotating in opposite directions. Because of the influence of the Faraday effect, the two modes (called the ordinary and extraordinary ray) travel with slightly different phase velocities. At the receiving antenna, a vector addition of the two modes takes place, resulting again in a linear polarization. However, the resultant vector now exhibits a different polarization angle, Δa, which has accumulated over the path length S by vector addition of the ordinary and extraordinary components Δa_0 and Δa_x. For a static ionosphere and a constant path S (motionless satellite), the shift of the polarization angle can be expressed as:

$$\Delta a_0 - \Delta a_x = \Delta a + n2\pi = \frac{K}{f^2}\overline{B\cos\phi}\int_0^S N\,\mathrm{d}s \qquad \ldots\ldots(1)$$

where f is the operating frequency; $\overline{B\cos\phi}$ is the effective component of the geomagnetic field in the direction of the ray path; N is the local electron density (along the path); and the line integral is formed along the radio path 0 to S.

The magnitude of the integrated electron density cannot be determined from eqn. (1) by measurement of the polarization angle, since the reference angle and the value of N are unknown. However, by making two simultaneous measurements at two different frequencies, the difference between $\Delta a_1 - \Delta a_2$ will furnish the desired result. This method, detailed by Swenson,[5] is incorporated in the system described in this paper.

Conventional methods for the determination of the electron density are based on the count of average fading periods in order to measure the rate of change of Faraday rotation, $\mathrm{d}a/\mathrm{d}t$, during a passage of the

† U.S. Army Electronics Research and Development Laboratory, Fort Monmouth, New Jersey.

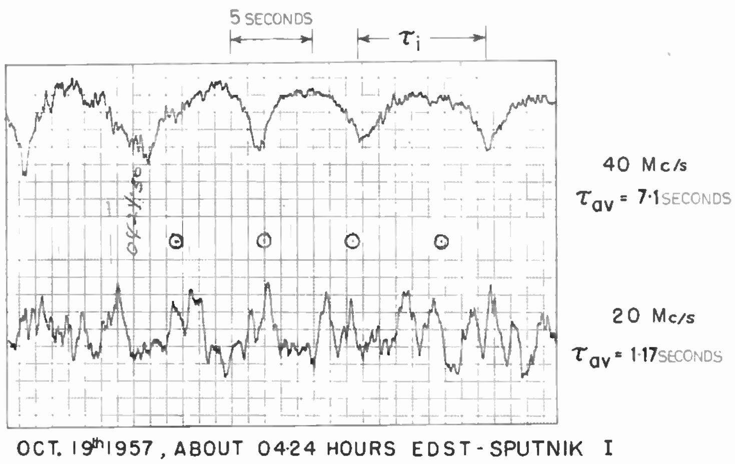


Fig. 1. Faraday rotation, conventional fading recording.

satellite. A simultaneous measurement of $\mathrm{d}a/\mathrm{d}t$ and the corresponding Doppler shift on one or two frequencies has been performed or proposed.[6, 7] These hybrid methods utilize very precise measurement of the Doppler frequencies employing a period-count method; however, they exhibit a lack of accuracy for the Faraday-rotation measurement. In Fig. 1 a strip-chart recording shows the rectified envelope of the high-frequency output of a linearly polarized antenna receiving a satellite signal with a $\frac{\mathrm{d}a}{\mathrm{d}t} = \frac{\pi}{\tau} \simeq 26$ deg/s. Single measurements can be made only every $\tau = 7$ seconds as compared with Doppler measurements, which deliver precise results every second. The sense of rotation cannot be determined. Lack of fine structure and the necessity of the manual transfer of measuring points call for an improved instrumentation and recording.

With this design requirement in mind, the polarization follower receiving system was developed. In addition, the experiment was aimed at the proof of the feasibility of novel electronic circuitry designed to achieve rapid pattern steering.

2. Measuring Principle

The fundamental method of measuring the polarization angle in this system is analogous to that of conventional direction finders. In both cases, an angular position is intersected with the characteristic pattern of an antenna. By rotation of the antenna and observation of the changing output, a maximum or minimum amplitude can be obtained. This manipulation can generally be performed mechanically or electronically, and both of these manipulations can be performed either manually or automatically. At the final adjustment, the calibrated indicator of the pattern control delivers the incident angular position. In our case, the sinusoidal polarization pattern of a dipole antenna is employed which should not be confused with the directional pattern as found in direction finders.

In the latter case the electromagnetic wave travels in the plane formed by the antennas, whereas the polarization pattern is perpendicular with respect to the antenna plane. The polarization-follower recording system under discussion permits high accuracy combined with faster response time, as compared with the conventional measuring methods mentioned in the introduction. With an accuracy of several degrees combined with a response time of less than one second, the system meets the following specifications:

(1) Electronically controlled, inertialess pattern steering.

(2) Capability of very slow positioning of the polarization pattern.

(3) Fast scanning and search process capable of controlling a servo system that simultaneously performs the orientation indicated under (2).

The following is a discussion of these three objectives (see Fig. 2). The incident electromagnetic wave, with the Poynting vector P perpendicular to the antenna plane and its angle of linear polarization α, is received by two crossed dipoles X and Y. The corresponding h.f. output voltages of these polarization-sensing antennas are E_X and E_Y. They are fed through two amplifiers to two corresponding modulators, designated as the X- and Y-multipliers. The X-multiplier forms the product of E_X with a magnitude cos β; the Y-multiplier forms the product of E_Y with a magnitude sin β. These two magnitudes are in the simplest case plus or minus direct voltages. The angle β represents, as we will see later, the adjusted polarization pattern angle of the system. A function

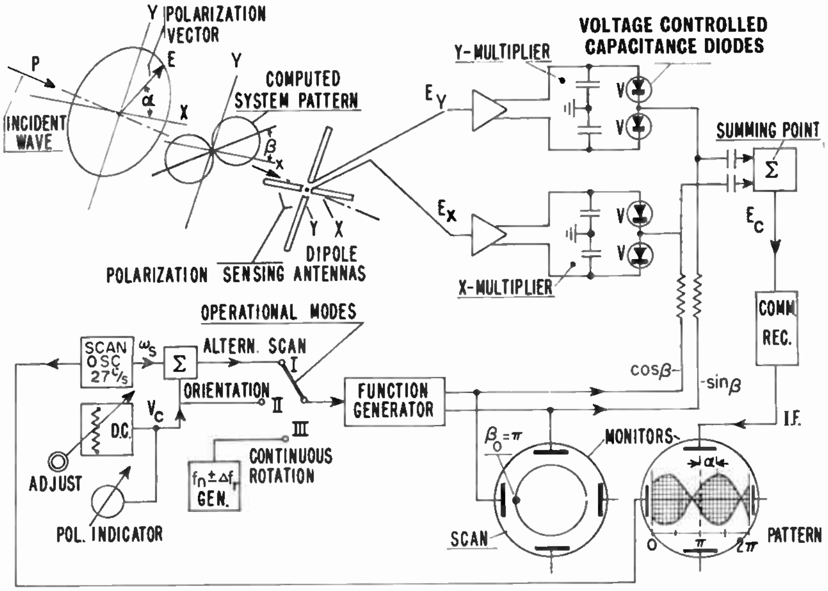


Fig. 2. Principle of electronically scanned polarization indicator.

generator computes cos β and sin β by receiving at its input a direct voltage that is proportional to the angle β. Let us now return to the output products of the two multipliers. In a summing device, the two products are added together to form the sum voltage E_c. In the communications receiver, E_c is amplified and the h.f. converted to the i.f., which is available at the receiver output. This output can either be applied to the vertical deflection plates of an oscilloscope (pattern monitor), or undergo a demodulation process in order either to indicate or to follow automatically the incident polarization.

Let us next perform a pattern measurement, with a fixed direct voltage, for an arbitrary setting of β, applied to the input of the function generator. If the polarization angle α of the incident wave rotates from zero to 360 deg, a figure-eight pattern is obtained. As indicated in Fig. 2 under 'computed system pattern', the maximum amplitude of this pattern appears under an angle β. If the input direct voltage to the system is varied, angle β changes and the system pattern rotates.

Thus, the first objective of the polarization follower system is achieved: an electronic, inertialess steering of the polarization pattern.

Regarding the measuring process, provision is made for three operational modes: If the switch in Fig. 2 is turned to position II (orientation), the β-pattern can be adjusted as described before. If the incident wave has constant polarization angle α, an intersection takes place between the polarization vector and the arbitrarily adjusted β-pattern. For a manual searching process the d.c. input voltage V_c has to be varied in order to maximize the i.f. output. When the maximum is reached, angle β is equal to angle α. For slowly varying polarization angles, continuous readjustment of V_c is necessary. This ability to position an antenna pattern (d.c. adjustment) represents the second objective for the follower system.

From an engineering point of view the following objection is now justified. If the search and follow-up process is so slow, why cannot a mechanical motion of one dipole or that of a goniometer be used to serve the same purpose? Mechanical antenna motion is eliminated because of the large size of the antenna configuration, which, in the present case, has to be designed for frequencies in the h.f. or v.h.f. band. Basically, goniometers in combination with crossed dipoles could be employed. In fact, the modulation scheme of the system follows the same fundamental theory as is applied to the goniometer function. The reason for not following the goniometer concepts is related to the requirement for a fast scan of the slowly changing pattern. A slow search and follow-up process would exhibit the same disadvantages of inaccuracy as the method discussed in Section 1 (Fig. 1). In order to obtain optimum detection of the polarization information in the presence of noise, the integrated information of all pattern angles has to be evaluated. Further, the scanning process involved has to be performed with high angular velocity, thus eliminating all mechanically operated devices. Two electronic scanning methods meet this requirement: continuous rotation and alternate scan.

Continuous rotation of the β-pattern represents the simplest scanning method. This operation takes place in switch position III where a generator of frequency $f_n \pm \Delta f_r$ beats with another generator (not shown) of frequency f_n. The rotation frequency Δf_r synchronizes the horizontal linear sweep of the pattern monitor. The entire pattern is now displayed and its maximum appears with an offset of the oscilloscope's centre equivalent to the angle α. In this case the control functions are time functions of the form cos $\omega_r t$ and sin $\omega_r t$ with $\omega_r t = 2\pi\Delta f_r$. These two functions displayed on an X-Y oscilloscope will display a closed circle (Fig. 2). This 'continuous rotation mode' is the mode conventionally used for electronic pattern control.[8] However, the lack of orientation capability makes the incorporation of an alternate-scanning method more feasible.

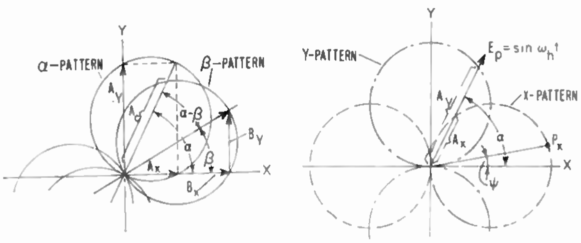
The measuring principle used in this system is explained in Fig. 2, showing the mode switch on position I (alternate scan). The control voltage V_c is superimposed with an alternating voltage. A typical value for this scanning frequency is $f_s = 27$ c/s; $\omega_s = 2\pi f_s$. It is also used as the horizontal sweep in the pattern monitor. The sum $(V_c + S \sin \omega_s t)$ is now applied to the input of the function generator. The two outputs form a semicircle when displayed on the screen of the scan monitor. For this experiment, as indicated in Fig. 2, the orientation voltage V_c is adjusted to deliver a pattern angle of 180 deg ($\beta_0 = \pi$) and the scan amplitude is such that the maximum deviation is π. With these adjustments the semicircle will just close to a circle. Thus, the β-pattern will move ± 180 deg back and forth around a centre value of 180 deg. With the polarization vector received, the pattern displayed on the screen of the pattern monitor will indicate the angle α.

Instead of measuring α by reading its magnitude from the oscilloscope and keeping the orientation constant we can keep the maximum of the pattern at the centre of the screen by manual adjustment of the control voltage. An instrument (shown in Fig. 2) calibrated from 0 to 360 deg will then deliver the reading of the measured angle α.

This method of simultaneous scanning and orientation represents the third objective. It meets all requirements for the incorporation of a servo system. In this case the pattern monitor, functioning as an indicator, has to be replaced by an envelope detector followed by a phase discriminator with the scanning frequency as reference. When β differs from α, a positive or negative error voltage is generated. This voltage controls the direction and speed of a servo motor which drives a potentiometer by means of a reduction gear. This potentiometer is equivalent to the adjustment potentiometer in Fig. 2. In closed-loop operation, the entire system has the tendency to keep $\beta = \alpha$; thus, the control system follows any polarization change of the received signal. The servo output voltage V_c can be recorded. Thus, automatic tracking and recording is achieved, being the purpose of the design of this polarization follower.

3. Theory of Pattern Control

Since we are now familiar with the measuring principle, we will continue with basic theoretical considerations that will prove the feasibility of the multiplication scheme incorporated in our system. In this section the rotation of the polarization pattern of combined dipole antennas will be discussed. The configuration of this pattern is bi-directional and can be described by a figure 'eight'. Figure 3(a) shows such a pattern, designated as β-pattern. Expressed in polar co-ordinates and unity amplitudes, it has the form:



(a) Theory of pattern rotation (b) Antenna components

Fig. 3. Pattern control.

$$p = \cos(\psi - \beta) \qquad \ldots\ldots(2)$$

where p = amplitude for the angle β

ψ = variable pattern angle

β = angle of lobe maximum

In rectangular co-ordinates

$$\tan\beta = \frac{\sin\beta}{\cos\beta} = \frac{B_Y}{B_X} \qquad \ldots\ldots(3)$$

Since we are interested in the intersection of the polarization vector, appearing under angle α, with the β-pattern, the amplitude at the point of intersection (with $\psi = \alpha$)

$$A_c = \cos(\alpha - \beta) \qquad \ldots\ldots(4)$$

Using the well known trigonometric identity,

$$\cos(\alpha - \beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta \qquad \ldots\ldots(5)$$

or

$$A_c = A_X . B_X + A_Y . B_Y \qquad \ldots\ldots(6)$$

we obtain the basic equation for the sum amplitude A_c. The functional building blocks (see Fig. 2) of the modulation scheme under consideration have the following meaning: the components B_X and B_Y are generated by the function generator, while A_X and A_Y are the components delivered by the X- and Y-antennas. The X- and Y-patterns of the two crossed dipoles are shown in Fig. 3(b). Their corresponding patterns using eqn. (2) can be expressed as

$$\begin{aligned} p_X &= \cos\psi \quad \text{for } \beta = 0 \\ p_Y &= \sin\psi \quad \text{for } \beta = \pi/2 \end{aligned} \qquad \ldots\ldots(7)$$

For $\psi = \alpha$ we obtain

$$\begin{aligned} A_X &= \cos\alpha \\ A_Y &= \sin\alpha \end{aligned} \qquad \ldots\ldots(8)$$

Thus, the sum of the two products indicated in eqn. (5) results in the combined amplitude A_c of eqn. (4).

Since this amplitude is obtained by the intersection of the β-pattern with a polarization arriving under the angle α, it can be maximized by rotation of the β-pattern so that $\beta = \alpha$. Equations (8) and (5) consequently become:

$$A_{c(\max)} = 1 = \cos^2\alpha + \sin^2\alpha = A_X^2 + A_Y^2 \qquad \ldots\ldots(9)$$

The multiplication of the antenna components A_X and A_Y with amplitudes of $B_X = A_X$ and $B_Y = A_Y$ respectively, results in an α orientation of the β-pattern.

Since the received polarization vector delivers an h.f. signal of unity amplitude for all angles α, a voltage E_p will be induced in the maximum of the β-pattern that represents the carrier in our modulation scheme:

$$E_p = \sin\omega_h t \qquad \ldots\ldots(10)$$

The h.f. output voltages of the two dipoles can be obtained by multiplication of eqn. (8) by eqn. (10):

$$\begin{aligned} A_X(t) &= \cos\alpha\sin\omega_h t = E_X \\ A_Y(t) &= \sin\alpha\sin\omega_h t = E_Y \end{aligned} \qquad \ldots\ldots(11)$$

The summing point output consequently becomes, using eqn. (4):

$$A_c(t) = \cos(\alpha-\beta)\sin\omega_h t = E_c \qquad \ldots\ldots(12)$$

4. Scanning-modulation and Error-voltage Generation

The expression for E_c in eqn. (12) assumed a constant angle α and a slow positioning procedure for the β-pattern. In the following, a periodic scanning function for β will be incorporated. In the simplest case of continuous rotation, angle β is a linear function of time, and by substituting $\beta(t) = \omega_r t$ into eqn. (12), the sum voltage has the form:

$$E_{c,r} = \cos(\alpha - \omega_r t)\sin\omega_h t \qquad \ldots\ldots(13)$$

This mode, as mentioned before, is not employed in the present system; it provides, however, a simple explanation of the fundamental features of the modulation scheme.

The sum voltage E_c is represented by a double-sideband (d.s.b.) suppressed carrier signal with $\sin\omega_h t$ as carrier. The two sidebands are f_r above or below the carrier frequency f_h. The envelope carries the phase information of the polarization angle α with reference to the original scanning frequency.

Incorporated in the system is the 'alternate scan'. This scanning time function can be written as:

$$\beta_s(t) = \beta + \Delta\psi \sin\omega_s t \qquad \ldots\ldots(14)$$

where $\Delta\psi$ = scanning amplitude (maximum deviation). Equations (14) and (12) produce the envelope function:

$$E_{c,s} = \cos[\alpha - (\beta + \Delta\psi\sin\omega_s t)]\sin\omega_h t \ldots\ldots(15)$$

This again represents a d.s.b. suppressed carrier signal. The envelope again contains the α information. If this envelope is demodulated by means of a product detector, a low-frequency signal (E_d) is obtained of the form:

$$E_d = \cos[(\alpha-\beta) - \Delta\psi\sin\omega_s t] \qquad \ldots\ldots(16)$$

This waveform can be analysed by means of Bessel functions. In order to generate an error voltage for controlling a servo system, the fundamental component can be used for the phase comparison with the reference scan frequency.

The system described here employs the simplest method of demodulation: envelope detection. Since the mathematical treatment for envelope detection of a function as given in eqn. (15) is rather complex and lengthy, this process is illustrated with the help of a polar diagram, as shown in Fig. 4(a).

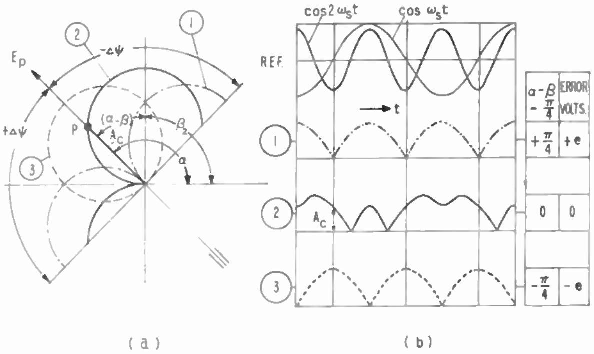


Fig. 4. Alternate scan, generation of error voltage after envelope detection.

The polarization vector (E_p) is shown with an angle $\alpha = \frac{3}{4}\pi$(135 deg). Three β-patterns (1), (2) and (3) are presented; each differ from the next by $\pi/4$ (45 deg). The centre pattern (2) has its maximum at

$\pi/2$ and intersects E_p at P, thus producing an instantaneous amplitude A_c. The superposition of the scanning function (in this diagram) is performed by causing a deviation of the angle α, instead of sweeping the β-pattern. This is done for the sake of easy comprehension and simplification of the diagram. From eqn. (15) or (16), it is evident that the scan is added to $(\alpha-\beta)$. Thus, no distinction can be made between both angles, and it is permissible to apply a scan amplitude of $\Delta\psi = \pm(\pi/2)$ to angle α.

Considering the time function of A_c for any $(\alpha-\beta)$ in between the indicated extreme values of patterns (1) and (3), a point-by-point construction will result in time functions of the detected envelopes for patterns (1), (2) and (3), as shown in Fig. 4(b). On the top portion of the representation of the scanning frequency, $\cos \omega_s t$ is indicated together with its second harmonic. Below, the reference frequencies are shown as they appear after envelope detection, corresponding to the three pattern configurations. Waveform (1) contains a large amount of second harmonics in phase with the second harmonic of the reference frequency. Waveform (3) shows a second harmonic that is in opposite phase with the second-harmonic reference. The centre configuration (2) delivers a complex waveform which, according to a frequency analysis, exhibits a smaller amount of second harmonic, compared with case (1) and (3), but with a phase shift of 90 deg to the second-harmonic reference.

The generation of an error voltage to control a servo system can now be performed as follows (Fig. 7): the i.f. output of the receiver, delivering the converted summing-point function $E_{c,s}$ of eqn. (15), is rectified and the resulting low-frequency waveform, which is identical to those of Fig. 4(b), is passed through a filter, the centre of which is tuned to the second harmonic of the scanning frequency. The filter output and the second harmonic of the reference scanning frequency are fed to a phase comparator.

The extreme and centre voltages of the discriminator curve produced by this output correspond to the three β-pattern orientations (1), (2) and (3). These voltages are:

$$e_{1,2,3} = e > 0 > -e$$

for $$(\alpha-\beta-\pi/4)_{1,2,3} = +\frac{\pi}{4} > 0 > -\frac{\pi}{4} \quad \text{......(17)}$$

Since the servo system tends to make the error voltage equal to zero, the pattern angle β adjusts itself, according to eqn. (17), to $\beta_2 = \alpha-\pi/4$. This offset of 45 deg is represented by the indicated pattern position on the screen of the two monitors in Fig. 7. Appearing as a constant angle, this offset is of no significance to the function of the entire system.

5. Multiplier and Function Generator

Since the multipliers and the function generator can be considered as being the most essential functional units of the electronic pattern control system, a brief explanation of their function is advisable. More details can be found in Reference 9.

The multiplier represents a specific type of modulator. In contrast to the conventional type, this modulator has to perform a multiplication in a strict mathematical sense. In particular no clipping of the carrier is allowed. In this application the output generates a carrier-suppressed d.s.b. signal with a high degree of freedom from distortion. The design exhibits (Fig. 2) a capacitive bridge arrangement with two voltage-controlled variable-capacitance diodes in one bridge arm and two equal fixed capacitors in the other arm. The diodes are biased in the non-conducting region (not shown). This multiplier performs with high stability and a carrier suppression in the order of 50 dB. Excellent linearity for h.f. inputs in the order of 0·1 V with modulation inputs of ± 1 V has been achieved.

The high input impedance of the multipliers is matched to the capacitive outputs of the function generator. In Fig. 5, an attempt is made to explain the generation of the control functions. A clock frequency f_m generates simultaneously a sawtooth waveform and two auxiliary sine waves of the frequency $f_n = 2f_m$. Their relative phase angle is 90 deg. The output functions $\sin\beta$ and $\cos\beta$ are now computed in the following way: an instantaneous amplitude V_{c1} of the control voltage is fed to an analogue comparator in which the sawtooth waveform is intersected by the control voltage. At the time instant T_{c1}, both voltages are equal and a very short pulse (50 ns) is generated. This control pulse drives two coincidence gates into conduction. One gate is

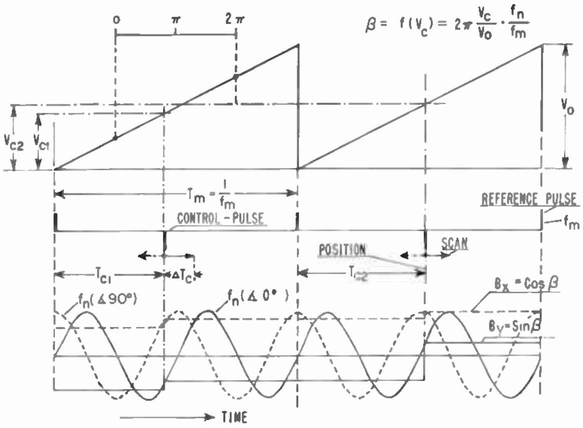


Fig. 5. Generation of control functions.

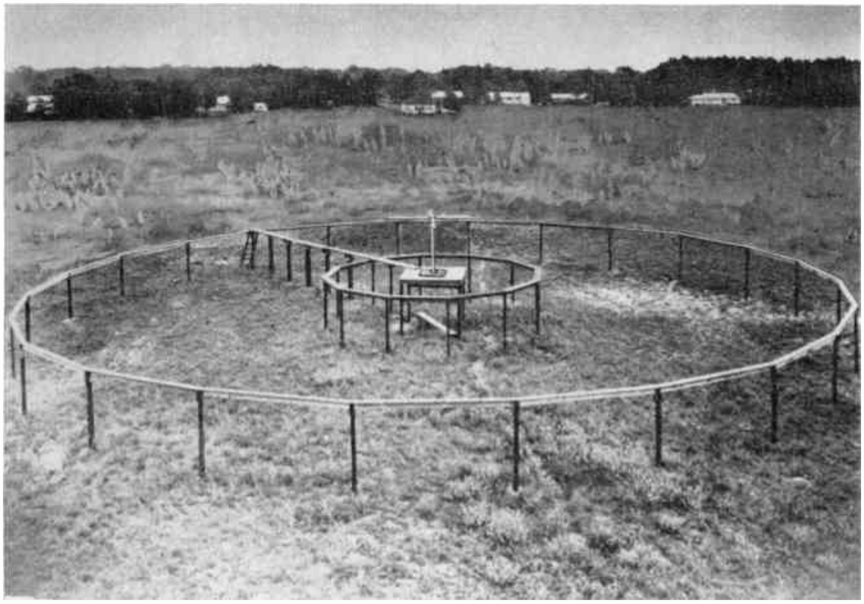


Fig. 6. Polarization sensing antenna.

connected to the sine generator of f_n, the other to the cosine generator. The outputs of each of the gates have a capacitor as a load. Thus, the control pulse intersects the sine and cosine waveform depicting at the instant T_{c1} the amplitudes of sin β and cos β respectively. The capacitors are instantaneously charged to the corresponding voltages and remain at these voltages until a new set of values is obtained: for example, V_{c2} changes the pulse position to T_{c2}, generating different values of sin β_2 and cos β_2. The arrows ΔT_c indicate the motion of the control phase caused by the alternate scan. The geometry of the waveshapes makes it evident that β is proportional to V_c. The time functions of f_m and f_n are completely eliminated and the conversion of d.c. magnitudes is one of the striking properties of this computation method.

6. The Complete Receiving System

The described functional units of the scanning and follower method can now be combined with the sensing antenna, frequency converters, and communications receiver to form a receiving system.

The sensing antennas form a functional unit with the pattern control system or vice versa. In Fig. 6 the actual antenna is shown. The crossed dipoles were designed for the 40–41 Mc/s frequency band, and great care was taken in the antenna design to maintain mechanical and electrical symmetry. Therefore, the transformer feed system is arranged inside the antenna mast. The dipole group is then mounted on a 2-metre elevated counterpoise of 30-metres (4λ) diameter. This eliminates pattern distortion caused by secondary radiators.

The functional interconnections of the remainder of the system's sub-units can be followed by using the block diagram of Fig. 7. In order to keep the system noise figure below 5 dB, it was necessary to keep the antenna-feed cable as short as possible. Thus, the first frequency conversion from 40–41 Mc/s to 10 Mc/s is performed in a hut about 62 metres away from the antenna configuration. Here the simultaneous satellite transmission of c.w. signals at 40 and 41 Mc/s, as was specified in the Introduction, are received in time sequence. Two oscillators, one at 30 Mc/s, the other at 31 Mc/s are alternately switched to the double converter at 1-second intervals. The two converter outputs consequently generate a first i.f. of 10 Mc/s for both input frequencies. The common oscillator preserves the correct phase and amplitude information of the antenna voltages.

A pair of 250-metre h.f. cables connects the converter station to the main instrumentation that is located in the receiver building. It is evident that a front-end design of this distributed type, if not carefully designed, is very sensitive to differential phase and amplitude changes. However, because of environmental influences, some instabilities remain; consequently, provisions had to be made for a calibration procedure. A calibration transmitter, operating at 1-second intervals, either on 40 or 41 Mc/s feeds a calibration antenna. A horizontal dipole is located at a distance of 30 metres from the sensing antenna and its ray path forms an angle of

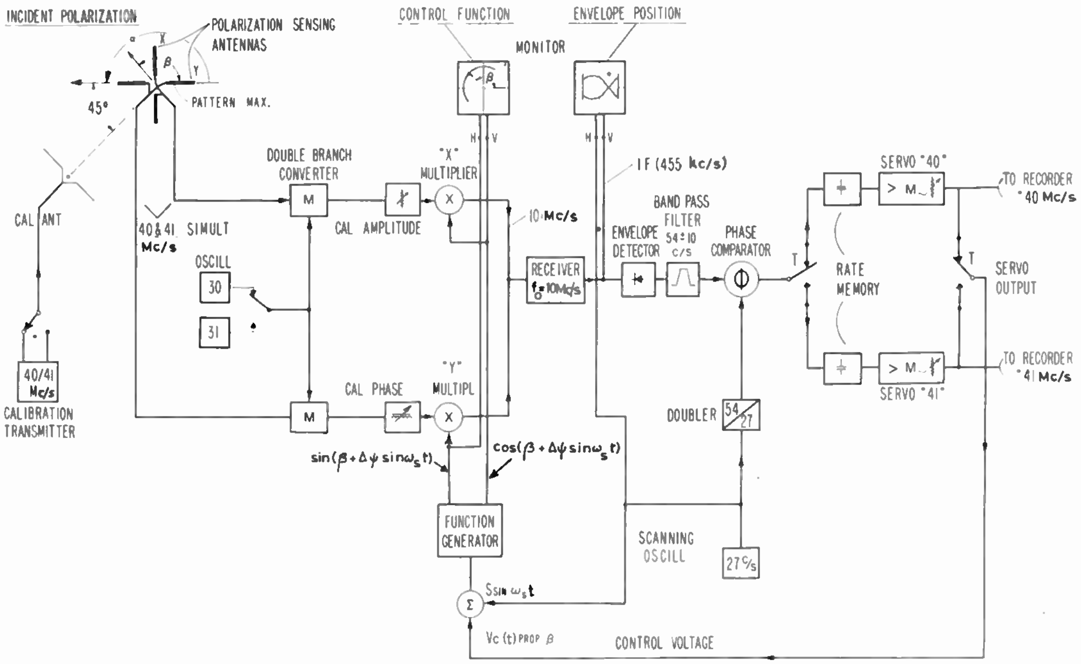


Fig. 7. Polarization follower receiving system for two frequencies.

45 deg to the dipole cross (Fig. 7). Thus voltages of equal phase and amplitude are induced to the X- and Y-antennas. In order to equalize eventual phase and gain differences, a time-delay line is inserted in the X-branch and an attenuator in the Y-branch. The correct adjustment is observed on the pattern monitors. The amount of correction necessary in the experimental system did not exceed 15 deg for the v.h.f. phase and 3 dB for the amplitudes.

The pattern-control system incorporated in Fig. 7 has been described before (Fig. 2). The communications receiver that follows is a high-quality type, and is fixed-tuned to the 10 Mc/s first i.f. Its i.f. output (455 kc/s) is amplified and fed to the envelope detector. The following bandpass filter, being of the electronic type, has a centre frequency of 54 c/s and a bandwidth of 20 c/s. Thus all other frequency components, especially 27 and 81 c/s, are sufficiently suppressed. As an error-voltage generator, a phase comparator is employed that uses complementary transistors in a temperature-compensation design. The reference frequency is generated in the doubler stage employing the same modulation method.

The error voltage is applied to a servo amplifier that in turn controls the motor of an integrating servo system. It represents a conventional-type mechanical servo with a potentiometer output. A rate generator (tachometer) is coupled with the servo motor to improve the stability of the system. The output potentiometer is of the continuous type exhibiting only a small gap between the extreme values. Maximum and minimum voltages are calibrated to adjust β angles of 360 deg and 0 deg, respectively, in the control system. The slider furnishes the voltage for the strip-chart recorder and the control voltage V_c to position the centre angle to the β pattern.

By feeding V_c to the function generator, and adding the scanning function at the control summing point, the servo loop is closed. Thus, pattern angle β follows any continuous rotation of the polarization angle α until 360 deg is reached. In case the polarization angle exceeds 360 deg, a special circuit is provided that returns V_c instantaneously to zero volts. Thus, loss of angular information caused by the small gap is avoided.

As discussed before, only one pattern-control and error-voltage generating instrumentation is employed for detecting the polarization information for two signal frequencies. The fast steering method of this system makes this possible. However, the relatively slow response time of the servo cannot take two different input signals without exhibiting prohibitive transient responses. Therefore, two servos are provided which are alternately switched into the servo loop. This switching is performed by a 1-second timer remotely controlled and synchronous with the oscillator switch in the double-branch converter. Even with separate servos, a stop would occur in the event when the other servo is in closed-loop operation. To prevent transients of this type, a storage network, called 'rate memory', keeps the rate of change of V_c constant for a 1-second period and with the magnitude that prevailed during the preceding period.

This concludes the explanation of the system's functioning. Considering a test recording with one satellite signal only, when the receiver is tuned in and the servo loop is closed, the pattern monitor will display a pattern with its zero to 45 deg to the right of the centre of the pattern monitor, and remain there. The β-pattern angle can be checked by observing the centre of the semicircle on the screen of the control function monitor. The ends of the semicircle indicate the scan amplitude of ± 90 deg. With angle β following the polarization angle α, the semicircle will slowly change its position. This is due to the change of the output voltage V_c, which can be recorded in a strip-chart recorder or measured in a digital voltmeter with a print-out device in order to produce a permanent record.

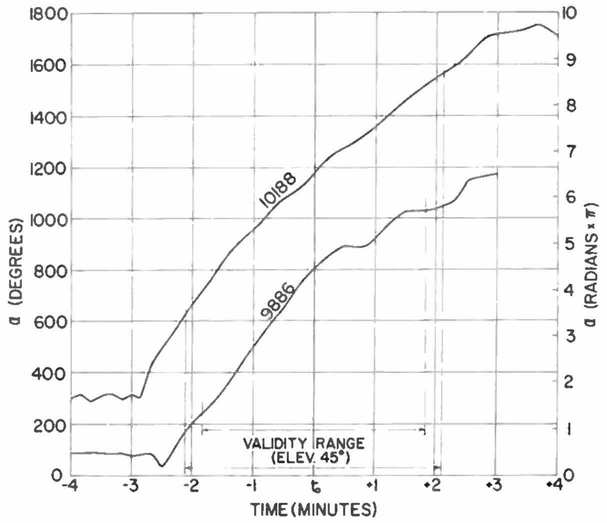
7. Continuous Faraday Rotation Recording

In Fig. 8, actual recordings of two passages of the satellite *Omicron 61* are shown. The upper trace displays the actual recording of the Faraday rotation. Zero degrees is indicated at the top, and 360 deg at the bottom of the chart. Directly below, the relative signal amplitude as received by a circularly polarized horizontal antenna is shown. The bottom trace of each orbit contains the time markers. Time advances from left to right. Since the rotation during one orbit is several times 2π, the recorder jumps three or four times respectively from 360 to 0 deg. These jumps can be added as shown in the graph of Fig. 8. The time t_0 indicates the inflection point of the orbit that occurs at the inflection of the Doppler frequency (closest approach). For the polarization a validity range is indicated from approximately $t_0 \pm 2$ minutes. This means that the method employed in this system produces correct measurements only when the satellite is in a near overhead position. In the case of horizontal incidence, the system loses its ability to distinguish polarization at all, for then it functions as a direction finder. In between the two extremes, a cone with a certain elevation angle can be described in which the satellite must be located in order to obtain a measurement exhibiting a tolerable error.

One interesting phenomenon emphasizes the ability of the system to record the fine structure of the Faraday rotation. In orbit 9886 at $T = +0{\cdot}5$ min, a complete halt of the Faraday rotation occurs. This will be true when the component of the earth magnetic field $\overline{B \cos \phi}$ (as mentioned in eqn. (1)), becomes zero. This takes place when the ray path angle ϕ with respect to the magnetic field vector becomes 90 deg.

8. Conclusion

The experimental receiving system of a polarization follower in its application as a research instrumentation has confirmed the feasibility of the described electronic scanning and orientation methods. It is evident that the performance of this system represents a detection method for linear polarization information.



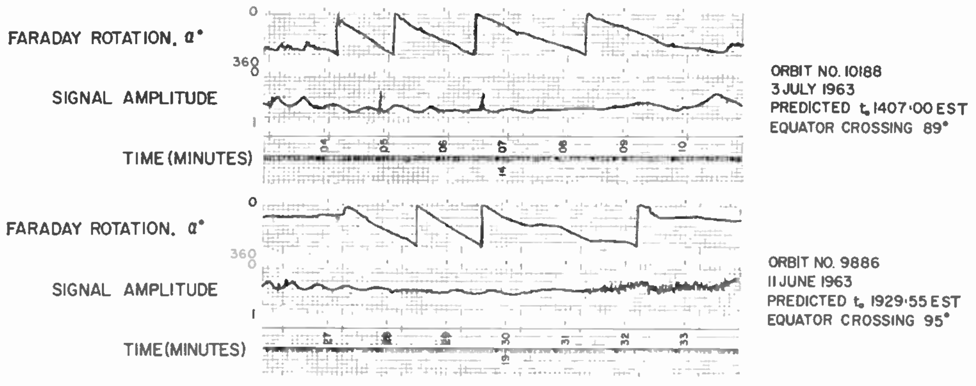


Fig. 8.

(above) Accumulated Faraday rotation.

(left) Faraday rotation, continuous recording.

Since the electronic scanning method can be extended to higher modulation frequencies and modified without difficulties, a communication system employing polarization modulation can be devised for the most general form of elliptical polarization. In all cases where the polarization distortion in the propagation medium is severe, and consequently debases the information contained in conventional modulation schemes, improvement in detection of such signals can be obtained. Thus the polarization follower described in this paper does not only support a rather special field of radio-propagation research, but also opens up many interesting areas for research investigations and applications in the field of polarization transmission and detection.

9. Acknowledgment

The author would like to acknowledge the assistance of Mr. J. Grau, Mr. W. Fischer and Mr. G. Strimple in the design, construction, and testing of the instrumentation. Thanks are also due to Dr. P. R. Arendt for his valuable advice relating to the physics of the ionosphere and for his support of this experiment.

10. References

1. T. Hatanaka, "The Faraday Effect in the Earth Ionosphere with Special Reference to Polarization Measurements of Solar Radio Emission", Science Report No. 5, School of Elect. Engrg (Cornell University, 30 August 1955).
2. I. C. Browne, J. V. Evans, J. K. Hargrave and W. A. S. Murray, "Radio echoes from the moon", *Proc. Phys. Soc.*, **B 69**, pp. 901–20, September 1956.
3. S. J. Bauer and F. B. Daniels, "Ionospheric parameters deduced from the Faraday rotation of lunar radio reflections", *J. Geophys. Res.*, **63**, p. 439, June 1958.
4. F. B. Daniels and S. J. Bauer, "The ionospheric Faraday effect and its applications", *J. Franklin Inst.*, **267**, No. 3, pp. 187–200, March 1959.
5. G. W. Swenson, jun.,"The utilization of ionosphere beacon satellites", X-250-62-32, NASA-GSFC.
6. F. De Mendonca, O. K. Garriot, "Ionospheric electron content calculated by a hybrid Faraday-Doppler technique" *J. Atmos. Terrestr. Phys.*, **24**, pp. 317–21, 1962.
7. P. R. Arendt, "Measurement of ionospheric electron content regardless of approximation", *Nature*, **197**, No. 4867, pp. 579–80, 9th February 1963.
8. H. V. Cottony and A. C. Wilson, "A high-resolution rapid-scan antenna", *J. Res. Nat. Bur. Stds*, **65D**, pp. 101–10, January-February 1961.
9. G. Vogt, "An electronic method for steering the beam and polarization of h.f. antennas", *Trans. Inst. Radio Engrs on Antennas and Propagation*, **AP–10**, No. 2, pp. 193–200, March 1962.

Manuscript received by the Institution on 30th March 1964. (Paper No. 936/RNA33.)

DISCUSSION

Under the chairmanship of Mr. W. K. Grimley, O.B.E.

Dr. G. Ziehm: How does modulation of the incoming signal influence the servo-system?

Mr. Vogt (*in reply*): This system processes the two antenna outputs in quadrature, since both output voltages are influenced by signal modulation (a.m. or f.m.) simultaneously. In the same manner only the pattern amplitudes or the time phases change but the polarization angle remains constant. Therefore, there is basically no influence by signal modulation to the servo function. However, since the entire system is linear, the loop gain may change for slow signal amplitude changes or if the signal modulation frequency would be close to the scanning frequency the error voltage phase could be influenced.

Mr. W. D. Worthy: What is the reason for the servo-following for measuring the Faraday rotation? Since extremely good multipliers have been used for multiplying the signal by sin β and cos β, information on the angle is available from the electrical measurement.

Mr. Vogt (*in reply*): The best use of the multipliers, for making only electrical measurements, is made by an indicator system as described in Fig. 2. We should not forget that the Faraday rotation measuring device is incorporated in a receiving system using only one receiver for both multiplier outputs. This is done to insure thai the system measures accurately in the presence of noise and fadings.

The servo-following is necessary when the rate of change of the polarization rotation is too high for manual evaluation and permanent record of the time-varying polarization angle is required for statistical purposes.

U.D.C. 621.396.677.3

Theoretical and Experimental Studies of the Resolution Performance of Multiplicative and Additive Aerial Arrays

By

E. SHAW, B.Sc.†

AND

D. E. N. DAVIES, Ph.D.
(*Associate Member*)†

Reprinted from the Proceedings of the Symposium on "Signal Processing in Radar and Sonar Directional Systems", held in Birmingham from 6th–9th July, 1964.

Summary: This paper describes some interim results of a theoretical and experimental study of multiplicative signal processing for receiving aerial arrays, such as those used in centimetric radar systems. Theoretical comparison is made between multiplication and other forms of demodulation such as linear and square-law rectification, in terms of the effect of demodulation on the ability of the directional array to resolve two closely spaced signal sources (or targets in a radar system). It is shown that the multiplicative system is superior to the other two systems for targets of approximately the same signal strength. It is also shown that improvements in resolution can result from the use of integration if the signals from the two sources are incoherent or partially incoherent. The effect of noise on resolution is briefly discussed and attention is drawn to the effect of the different directional responses of the different signal and noise products of demodulation.

An experimental 8-element multiplicative array operating in the 3-cm band is described; this array is also capable of fast electronic scanning. Experimental directional responses of this array are presented for both single-target and multiple-target excitation, for static measurements and at electronic scanning rates of 2 kc/s.

1. Introduction

Multiplicative signal processing represents an alternative method of demodulating the output of directional receiving arrays and compared with the more familiar processes of linear-law, or square-law rectification possesses several advantages. One advantage of this form of processing is that the directional properties of the array may be varied by modifications to the signal processing and this can lead to the production of more optimum forms of directional characteristic for given applications. Three principal features of this type of processing may be listed as follows:

(1) The production of directional patterns with narrower beamwidths than the conventional additive array.

(2) Directional patterns whose side-lobes are of negative polarity relative to the main lobe over part or all of the range of angles. This produces properties similar to that of side-lobe suppression systems.

(3) The production of pencil-beam directional patterns from the combined output of two linear arrays arranged at right angles to one another.

2. The Nature of Multiplicative Signal Processing

Historically this form of signal processing has arisen in the fields of sonar[1,2] and radio astronomy[3,4] and little attention has been given to this topic by radar engineers. Furthermore the literature on the possible radar applications is almost exclusively concerned with idealized systems.[5,6] On the other hand the technique has proved operationally very successful in both sonar[7] and radio astronomy.

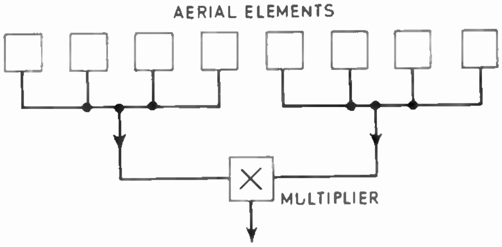


Fig. 1. A 4×4 multiplicative array.

As an example of a multiplicative pattern we can consider the product of the two equal halves of a linear array as shown in Fig. 1. The output of the multiplier contains two principal modulation products which represent directional characteristics of the array. Furthermore one such product is a demodulated signal retaining the relative phase information of the two

† Electronic and Electrical Engineering Department, University of Birmingham.

halves of the array in the form of the polarity of the output. Figure 2 shows the computed multiplicative directional pattern of an 8-element array divided into two equal halves before multiplication. It will be noticed that it possesses half the beamwidth of the rectified additive array. The first side-lobe of the multiplicative pattern is larger than that of the rectified additive pattern but all succeeding side-lobes are smaller.

It is important to compare the performance of multiplicative processed arrays with those of rectified additive arrays, and as can be seen from Fig. 2 the fact that the first side-lobes are of negative polarity represents a considerable advantage since for intensity modulated displays the negative side-lobe is effectively absent except when two closely spaced targets have to be resolved. The resolution performance of such arrays will be considered later.

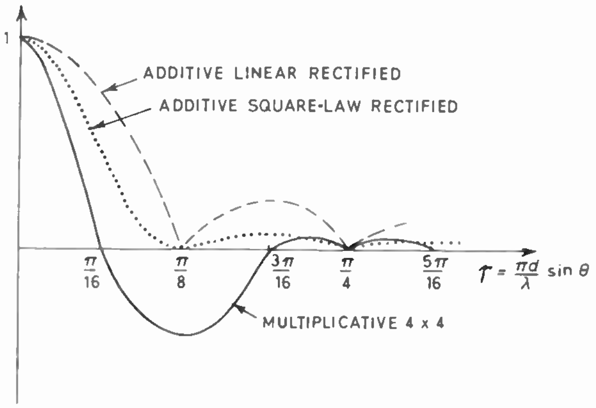


Fig. 2. Directional responses.

A further important point is that the directional pattern of a multiplicative array exhibits a square-law characteristic, since it is related to the product of the two input signals. For this reason when comparing directional patterns it is more reasonable to take a square-law-rectified additive array as a comparison standard, and this pattern is also shown in Fig. 2. It can also be seen that the multiplicative beamwidth is about 70% of the beamwidth of the square-law-rectified additive pattern.

The process of multiplication is irreversible and non-linear, therefore the laws of superposition do not apply except under certain circumstances which will be considered later. Consequently a proportion of the literature on this topic has attempted to distinguish between additive arrays and multiplicative arrays by calling the former linear processing and the latter non-linear processing. This approach is quite erroneous since it discounts the non-linear demodulation or rectification that must be used with an additive array. It is possible to consider a demodulation process that is linear such as the use of synchronous demodulation when used for single target working, but for any directional system required to cope with two or more targets the processing must be non-linear.

When a directional receiver is illuminated by two separate signal sources situated in the far field, the computed multiple target output is not given merely by the superposition of the single target directional responses at the appropriate angles and amplitudes, it is also necessary to consider the phase difference between the two sources. Therefore at this stage superposition is valid in the general sense taking account of both amplitude and phase. However when this multiple target response is demodulated, the demodulation process is non-linear and superposition does not apply to the overall system including demodulation. Thus the demodulated multiple target response cannot be calculated from the demodulated single target responses from which the phase information has been destroyed.

It has been shown by Berman and Clay[2] that the directional pattern of a linear multiplicative array (the term 'linear' in this instance referring to an array of elements geometrically arranged in a straight line) is given by the product of three factors:

$$D(p) = D_1(p)D_2(p)I(p) \qquad \ldots\ldots(1)$$

where $p = \dfrac{\pi d}{\lambda}\sin\theta \qquad \ldots\ldots(2)$

d = element spacing in array

λ = wavelength

θ = angle of direction relative to array normal.

$D_1(p)$ and $D_2(p)$ are the additive directional responses of the two parts of the array multiplied together. $I(p)$ is the interferometer directional pattern of a two-element array with its elements situated at the phase centres of the two parts of the array. If we take an 8-element array the additive directional response is given by:

$$D_a(p) = \frac{\sin 8p}{8\sin p} \qquad \ldots\ldots(3)$$

If this array is divided into two equal halves and the signals from these two halves are multiplied together the resultant directional response will be

$$D_m(p) = \left[\frac{\sin 4p}{4\sin p}\right]^2 \cos 8p \qquad \ldots\ldots(4)$$

where the cosine term represents the interferometer term.

As can be seen from Fig. 2 a multiplicative pattern can possess negative polarity side-lobes; it is also possible to synthesize such patterns with all the side-lobes negative. For single target applications this is equivalent to the complete suppression of side-lobes. However when such a response is used as a receiver for multiple signal sources with differing angular bearings, parts of the side-lobes may change polarity.

It is tempting to consider removal of the negative side-lobes by means of a rectifier, but again this is only of value for single target operation since the rectifier can only operate on the multiple-target response. The output, $S(p)$, of a scanned directional system can be represented as the convolution of the far-field distribution of targets $T(\sin\theta)$ with the directional response of the receiving system $D(p)$.

$$S(p) = \int_{-1}^{+1} T(\sin\theta)\, D(p-\sin\theta)\, \mathrm{d}(\sin\theta) \quad \ldots\ldots(5)$$

Therefore if a large target (or source) is situated in a position corresponding to a negative side-lobe the output of the receiver due to a small target situated in the main lobe may be cancelled out by the negative signal from the side-lobe. This is a problem which also arises in side-lobe suppression systems.[8] By making $T(\sin\theta)$ a complex expression, eqn. (5) takes account of the phase and amplitude distribution of targets.

3. Multiple Target Response of Directional Patterns

3.1. *Significance of Resolution*

The response of a directional system to simultaneous excitation with several sources is given by the convolution formula of eqn. (5). The main case to be considered here is one of particular practical significance; it is the ability of the system to distinguish two closely spaced sources (or targets in the case of a radar system). Now it can be argued that on an information theory basis this represents an irrelevant, or at least trivial case,[9] but it does represent an important engineering problem. The problem therefore, is to consider what conditions must be met in order that the scanned output of a direction system can resolve two separate targets.

It is clearly necessary to choose some criterion for resolution and the one chosen here is that the output corresponding to the main lobe of the multiple target response must contain a dip at the centre, of a specified level.

3.2. *The Effect of Phase Coherence*

Consider two sources x and y in the far field of a directional receiver with a directional response $D(p)$. Let the two sources radiate $\sin\omega t$ and $K\sin(\omega t+\psi)$ respectively.

Case I: Additive resolution with linear rectification.

Let

$$D(p) = D_a(p)$$

Then the combined output of the additive array due to the two sources, but before demodulation, would be:

$$D_a(p_x-p)\sin\omega t + K D_a(p_y-p)\sin(\omega t+\psi) \quad \ldots(6)$$

(where p_x and p_y are the angular positions of the two sources x and y on a 'p' scale). If this output were subject to linear rectification the output would become after filtration:

$$D_e(p_x, p_y) = \sqrt{D_a^2(p_x-p) + K^2 D_a^2(p_y-p) + 2K D_a(p_x-p) D_a(p_y-p)\cos\psi} \quad \ldots(7)$$

It can be seen from the above expression that it is the third term under the root sign that causes the dependence of the resolution properties upon the phase difference. Furthermore the mean value of the above expression as ψ varies through 2π is still dependent upon this cross-product term. It would therefore appear that simple integration of the above expression over a range of ψ will not remove the dependence of the output upon the relative phase or phase coherence of the two sources.

Case II: Additive resolution with square-law rectification.

If we replace the linear rectification of the previous section by a square-law rectifier the filtered output becomes:

$$D_s(p_x, p_y) = D_a^2(p_x-p) + K^2 D_a^2(p_y-p) + 2K D_a(p_x-p) D_a(p_y-p)\cos\psi \quad \ldots(8)$$

However the important difference between this case and the previous one is that the mean value of the above expression as ψ varies becomes

$$\overline{D_s(p_x, p_y)} = D_a^2(p_x-p) + K^2 D_a^2(p_y-p) \quad \ldots\ldots(9)$$

since the cross-product term integrates out. This result means that with suitable integration for incoherent target returns the multiple target response can be computed by the superposition of the squared single target responses without consideration of phase variations. Therefore since (with some qualifications) superposition applies to square-law rectification it may in this sense be considered a linear demodulation system but 'linear rectification' is necessarily a non-linear form of demodulation. However a linear process should also be reversible and no form of demodulation has this property except the trivial case of synchronous demodulation which has no meaning for multiple targets. This is not surprising since, as Woodward[10] pointed out, the process of demodulation destroys information.

Case III: Multiplicative resolution.

Let the array be divided into two parts having independent direction responses given by $D_1(p)$ and $D_2(p)$ and let the output of the two parts be fed to a multiplier. In the case of a 4×4 array the overall directional response would be

$$D(p) = D_1^2(p)\cos 8p = \frac{\sin^2 4p}{16\sin^2 p}\cos 8p \quad \ldots(10)$$

The additive output of each part in the presence of the two targets x and y is given by:

$$D_1(p_x-p)\sin\omega t + KD_1(p_y-p)\sin(\omega t+\psi) \quad \ldots\ldots(11)$$

and

$$D_2(p_x-p)\sin(\omega t+rp) + KD_2(p_y-p)\sin(\omega t+rp+\psi) \quad \ldots\ldots(12)$$

where rp is the phase difference between the signals arriving at the phase centres of the two parts of the array and corresponds to the interferometer term of eqn. (1). If these two signals are then fed to the inputs of a multiplier then the filtered output will be:

$$\begin{aligned} D_m(p_x, p_y) = {} & D_1(p_x-p)D_2(p_x-p)\cos rp + \\ & + K^2 D_1(p_y-p)D_2(p_y-p)\cos rp + \\ & + KD_1(p_x-p)D_2(p_y-p)\cos(rp+\psi) + \\ & + KD_1(p_y-p)D_2(p_x-p)\cos(rp-\psi) \end{aligned} \quad \ldots\ldots(13)$$

From the above it can at once be seen that the first two terms represent the directional response that would be computed on a superposition basis and the last two terms represent the unwanted cross products dependent upon the phase difference ψ. However, as with square-law rectification, the unwanted term may be integrated out provided there is a sufficient variation in ψ.

$$\overline{D_m(p_x, p_y)} = D_1(p_x-p)D_2(p_x-p)\cos rp + + K^2 D_1(p_y-p)D_2(p_y-p)\cos rp \ldots(14)$$

It may therefore be concluded that under such circumstances multiplicative processing can produce demodulated outputs from directional systems in which the multiple target response may be computed on a superposition basis, but that this is never true for the case of linear rectification.

One disadvantage of both square-law rectification and multiplication however is that they both possess square-law responses so that targets of relative strength $K:1$ produce outputs of relative strength $K^2:1$.

The conclusions drawn in this section are quite general and not dependent upon the form of the directional patterns, the way the multiplicative pattern is synthesized or the value of the interferometer term.

3.3. *An Exact Equivalence between Square-law Rectification and Multiplicative Signal Processing*

As a further demonstration of the fact that multiplicative signal processing is not a fundamentally different form of demodulation compared with square-law detection it is instructive to consider the relationship

$$\left[\sum_{i=1}^{I} a_i\right]^2 = \sum_{j=1}^{I}\left[a_j\sum_{i=1}^{I} a_i\right] \quad \ldots\ldots(15)$$

If a_i represents the output of the ith element of some general form of array, and a_j the output of the jth element, then the left-hand side of the above expression represents the square-law-rectified output of the array. The right-hand side represents a particular form of multiplicative processing of the same array shown in schematic form in Fig. 3. The above equivalence is completely general and therefore applies to both signal and noise for either single or multiple targets. The above equivalence is implicit in Ryle's aperture synthesis technique.[4]

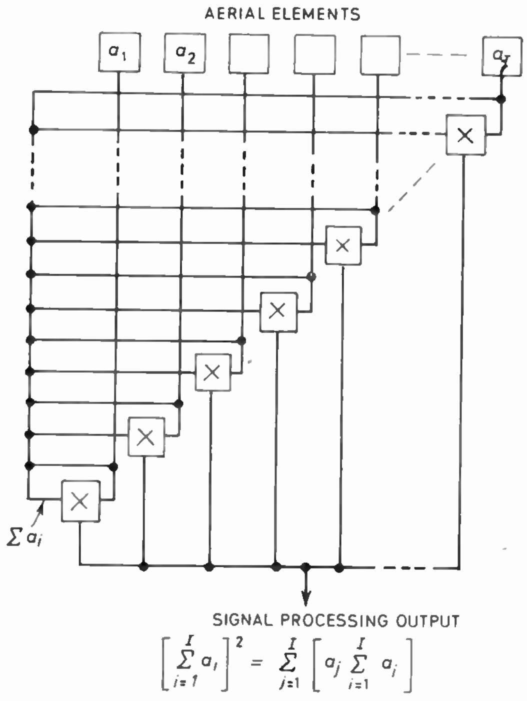

AERIAL ELEMENTS
a_1
a_2
a_I
Σa_i
SIGNAL PROCESSING OUTPUT
$\left[\sum_{i=1}^{I} a_i\right]^2 = \sum_{j=1}^{I}\left[a_j \sum_{i=1}^{I} a_i\right]$

Fig. 3. A form of multiplicative signal processing equivalent to square-law rectification.

The analytical value of such forms of multiplicative array as Fig. 3 is that by weighting the relative outputs of the different multipliers it is possible to produce different directional patterns. The required pattern may be subjected to Fourier analysis and expressed as a summation of terms representing a spatial correlation of the sampling points along the array.

3.4. *Directional Responses due to Two Targets*

Figure 4 shows the calculated output of an 8-element array when the additive output is square-law rectified and when the array is illuminated by two signal sources situated in the far field of the array. In the case shown the angular separation between the sources is $\pi/8$ rad (on a p scale where $p = (\pi d/\lambda) \sin \theta$). It can be seen that the resultant output is dependent upon the phase difference between the two sources and the outputs are computed for three different values of phase difference $\psi = 0, \pi/2$, and π.

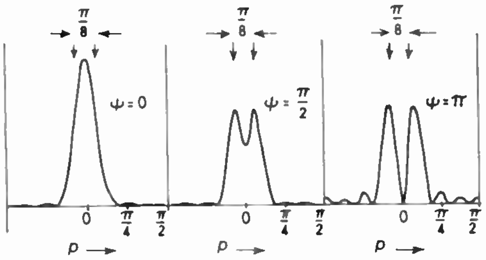


Fig. 4. Square-law-detected output for two targets.

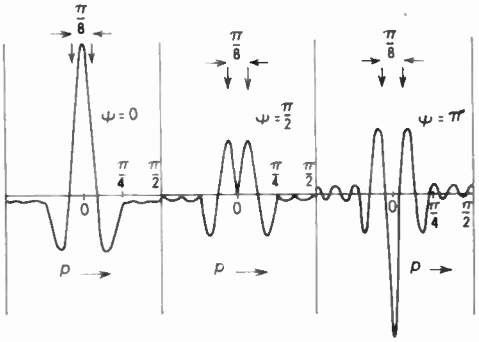


Fig. 5. Multiplicative output for two targets.

Figure 5 shows the corresponding two target patterns for the same array with 4 × 4 multiplicative processing. By comparing Figs. 4 and 5 it is possible to note that the multiplicative output provides better resolution in the sense that the dip between the two peaks is more pronounced for the case of $\psi = \pi/2$ and $\psi = \pi$, though the corresponding amplitude of the output is rather less for multiplicative processing.

It can be understood from the previous section that the resolving properties of a directional system are very dependent on the relative phase of the sources involved. If the value of ψ changes over a significant proportion of a cycle during the time that the targets are being observed then theoretically the resolution can be improved by integrating the resultant output over that period. If there is no appreciable variation of ψ over the period concerned then resolution has to be calculated on a basis of the probability of getting the correct phase. This is the reason that resolution is so dependent upon the 'relative phase coherence' of the returns from the two targets.

The significant time interval corresponding to the integration time will vary considerably depending upon the system under consideration. It may represent one pulse length, or one pulse repetition period, or the time taken for the beam to scan through the target. The limits of integration may be set by practical limitations on integrating circuits or operational restrictions on the behaviour of targets. In a real radar environment there may be variations of both amplitude and phase in the returns from targets due to such factors as target velocity, change of aspect and transmission variations due to an inhomogeneous medium. The well-known phenomena of target glint, scintillation and target fading rates are due to such causes.

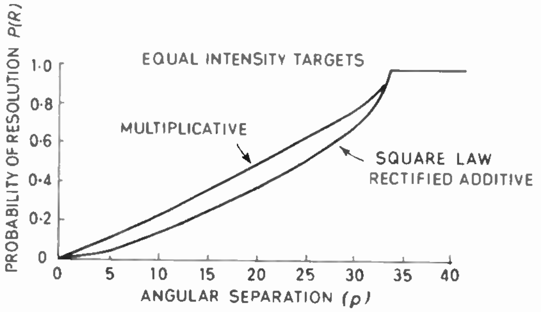
3.5. *Two Target Resolution in the Absence of Integration*

It is reasonable to assume that in a conventional radar system the value of ψ is not likely to change a significant amount in one pulse duration, since such a change corresponds to an absurd relative velocity between two targets moving close together. Therefore if the system is designed to afford little or no within-pulse integration the probability of resolution can be specified entirely in terms of the probability of obtaining a suitable value of ψ to produce resolution for a given target separation.

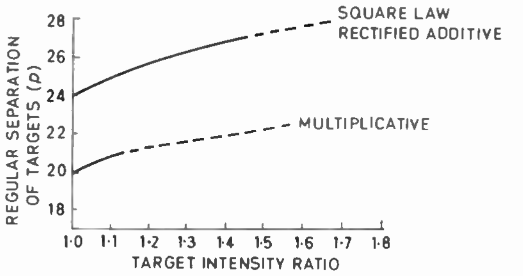
As a practical criterion of resolution it is proposed to stipulate that the receiver output must contain a dip in the region between the two peaks due to the two targets, of a specified level relative to the peak value. For the purpose of these calculations we shall specify a dip of 3 dB. In the case of targets of unequal signal strength the dip must be relative to the smaller of the two target peaks. To simplify the calculations it seems reasonable to make the approximation that the dip occurs midway between the two targets and that the smaller peak is given by the magnitude of the two target response in the direction of the smaller target.[11]

The above criterion is quite suitable for comparing square-law rectification with multiplicative processing, since they both exhibit square-law characteristics. But to compare these processes with linear-law rectification it would be necessary to stipulate a dip of 1·5 dB for the latter case. However in the absence of integration the square-law rectified multiple target response is just the square of the linear-law rectified multiple target response. Therefore in the absence of integration there is clearly no difference between the resolution performances of the two forms of rectification in terms of the above criterion.

Since for a given target environment all values of ψ are equally probable it is possible to define a probability of resolution $P(R)$, which may be defined as the probability of obtaining a value of ψ which satisfies the resolution criterion. One disadvantage of this criterion is that no account is taken of the degree of resolution provided that the criterion is satisfied; this should result in the calculated resolutions favouring the additive system.



(a) Probability of resolution against angular separation of targets on a p scale.



(b) Separation of targets against target intensity ratio.

Fig. 6. Resolution in absence of integration.

Figure 6(a) shows a plot of the value of this probability of resolution $P(R)$ against the angular separation of the targets on a 'p' scale for both rectification and multiplication. It can be seen from this graph that the multiplicative system always gives superior resolution to the additive system by a significant amount, but the improvement is less than the corresponding change in beamwidth. Figure 6(a) indicates that the target separation may be reduced by 25% for a constant probability of resolution of 0·4.

The above calculations are based upon the assumption that the two targets have the same signal strength. One disadvantage of any square-law-response system is the strong signal capturing effect, but in the form of comparison adopted here this effect appears for both forms of processing. Figure 6(b) shows how the resolution performance falls off as the ratio of the target strengths depart from unity. It is here that we find that the multiplicative system although providing superior resolution cannot maintain reasonable resolution over a wide range of target ratios. The dotted portion of the graphs indicate the stage when the weaker response falls below 50% of the single target main-lobe height. This occurs at a target ratio of 1·1 for multiplication and 1·4 for rectification. The graphs are plotted for $P(R) = 0{\cdot}5$.

The main reason for the worse performance of the multiplicative system in this context is the large value of the first side-lobe of the multiplicative directional pattern. This indicates the value in reducing these side-lobe levels even when they are negative.

The previous calculations of $P(R)$ depend upon the condition that ψ does not change in one pulse length. However the only way that it is possible to resolve two or more targets for a single pulse is with a within-pulse scanning system; the previous calculations therefore apply to such a system neglecting any integration. For the case of a mechanically scanned radar the above calculations are valid if the value of ψ does not change in the time taken for the beam to scan past the targets.

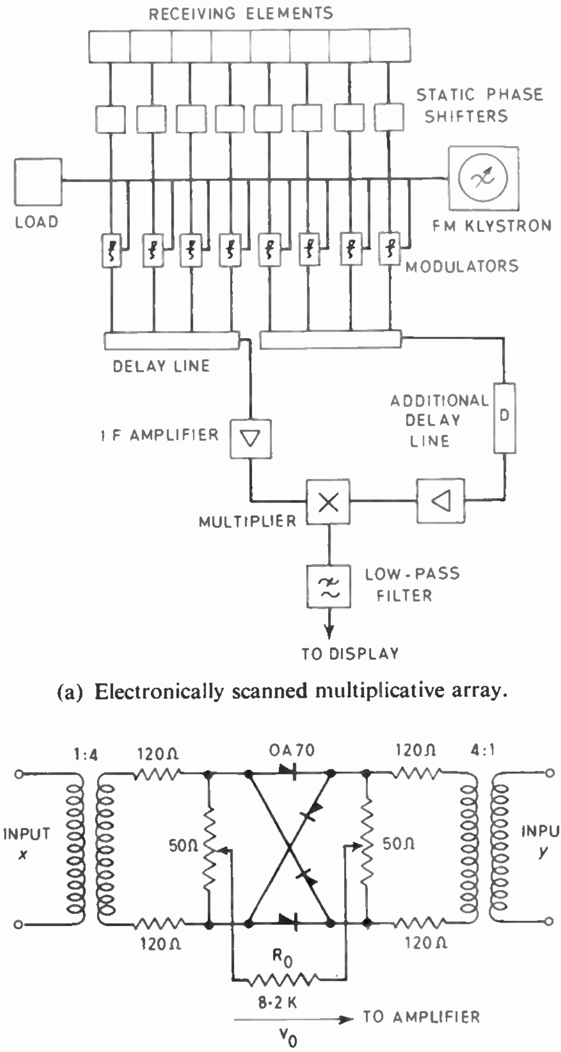
4. An Experimental Multiplicative Array and the Use of Electronic Scanning

4.1. *Description of Experimental System*

To examine the properties of multiplicative arrays an 8-element multiplicative array has been constructed in the Electronic and Electrical Engineering Department of the University of Birmingham. This array is also capable of fast electronic scanning. The latter facility has been introduced for two reasons: firstly, because multiplicative processing is particularly applicable to systems employing within-pulse electronic scanning, since the narrower beamwidth applies only to a receiving array and therefore a wider beam transmitter would waste energy unless within-pulse scanning were used[12]; secondly, the facility of electronic scanning provides a very convenient method of measuring the resultant directional responses.

Electronic scanning of the array is accomplished by the method described elsewhere by the present authors.[11, 13] The outputs of the receiving array elements are fed via mixers to tapping points on a delay line. The local oscillator, which is capable of frequency modulation also feeds all the mixers in phase. Thus as the frequency of the local oscillator varies so the frequency through the delay line varies. This provides a variable phase shift which deflects the beam.

Figure 7(a) shows how this scheme is used to scan the output of the multiplicative array. The two halves of the array are scanned separately using two delay lines with a common local oscillator, but the output



(a) Electronically scanned multiplicative array.

(b) Multiplier circuit.

Fig. 7. The experimental multiplicative array system.

of one delay line must be subjected to a further delay in order to scan the phase centres of the two halves of the array. For a beam deflection of n additive beamwidths (where n = the total number of elements in the array) the required phase modulation is 2π radians/section of delay line and it can therefore be shown that the relationship between the delay per section t_2 and the frequency sweep Δf is

$$\Delta f = \frac{1}{t_2} \qquad \ldots\ldots(16)$$

For the experimental arrangement shown in Fig. 7(a) the receiving array was illuminated by a source situated in the far field of the array and the local oscillator was frequency modulated in a saw-tooth manner. The output of the multiplier is then a time waveform representing the dynamic directional response of the scanning receiving array. Static directional characteristics may be measured in the usual way by rotating the array when the frequency modulation is removed from the klystron local oscillator.

4.2. *The Aerial Array and R.F. Assembly*

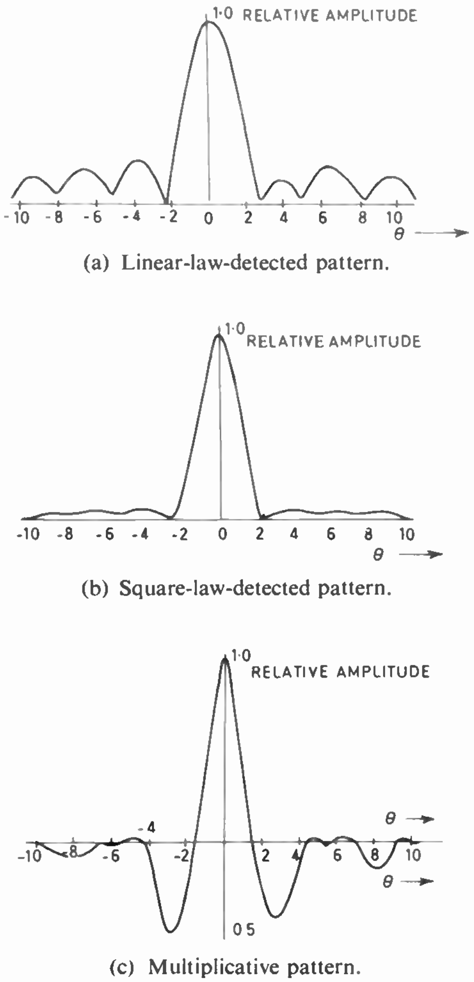
The receiving array consists of eight square horns, each of side 2·83λ, mounted contiguously with an overall aperture of 22·6λ, operating at a frequencv of 9415 Mc/s. This configuration gives additive beamwidths in the vertical and horizontal planes of 18 deg and 2·4 deg respectively. The multiplicative beamwidths are 13 deg and 1·2 deg in the corresponding planes. (The reduction of vertical beamwidth is due to the square-law response.) These beamwidths are based on the separation of the 3-dB points on the voltage pattern for a single point target and are expressed in terms of real angle, θ.

The horns feed the incident signals to crystal mixers via phase-shifters which compensate for static phase errors in the equipment. The mixers are fed with the frequency-swept carrier from the local oscillator, by means of directional couplers so spaced that each channel receives the carrier at the same phase. The frequency sweep of the klystron local oscillator is 20 Mc/s and the intermediate frequency is 40 Mc/s. Two i.f. amplifiers feed the signals to the two inputs to the multiplier. The demodulated signal at the multiplier output is used to measure the directional response. When the array undergoes electronic scanning this signal is displayed on an oscilloscope using the scanning saw-tooth wave-form as a time-base.

The multiplier used is a ring multiplier having the same form as a ring modulator circuit, but with the two inputs fed in to the transformers and the output taken from the two transformer centre-taps as shown in Fig. 7(b).

4.3. *Experimental Directional Patterns*

Figure 8 shows experimental results providing comparison between different forms of static directional pattern. Figure 8(a) is the additive pattern with linear-law rectification, this was obtained by observing the unrectified output of the receiver on a wide-band oscilloscope and plotting the modulus. The square-law-rectified pattern of Fig. 8(b) was obtained by feeding identical unrectified additive signals into the two inputs of the multiplier. Figure 8(c) represents the output of the array with 4 × 4 multiplicative processing using the ring multiplier shown in Fig. 7(b). The patterns, which were measured in the usual way by mechanical rotation of the array, show reasonable agreement with the theoretical pattern of Fig. 2.



(a) Linear-law-detected pattern.

(b) Square-law-detected pattern.

(c) Multiplicative pattern.

Fig. 8. Patterns obtained by mechanical rotation of aerial.

Figure 9 shows some corresponding results for the output of the array when it is undergoing continuous electronic scanning in a saw-tooth manner at a scanning rate of 2 kc/s. Figure 9(a) is the additive output of the 8-element array before demodulation. Figure 9(b) shows the output of the same array with 4 × 4 multiplicative processing.

There is an important difference between the static directional characteristics of Fig. 8 and the dynamic (or electronically scanned) patterns of Fig. 9. The first are the true directional responses of the array and this is given by the product of the array factor and the directional response of one array element $D_E(p)$. Thus the pattern of Fig. 8(a) should correspond to a response

$$D(p) = D_E(p)\frac{\sin 8p}{8\sin p} \qquad \text{......(17)}$$

For the experimental array each element was a horn with a response

$$D_E(\theta) = \frac{\sin(8\pi\sin\theta)}{8\pi\sin\theta} \qquad \text{......(18)}$$

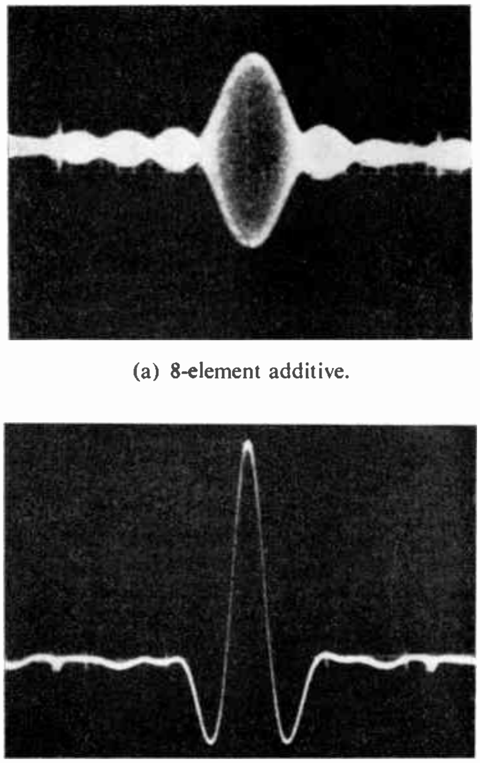
The output of the electronically scanned array however is merely the array factor times a constant multiplying factor. This is because the directional response of the individual elements is not being scanned. The multiplying factor will therefore be the response of the individual element in one fixed direction, (p_1), that is the direction of the signal source. Therefore the output of the electronically scanned array is

$$D(p) = D_E(p_1)\frac{\sin 8p}{8\sin p} \qquad \text{......(19)}$$

In the case of multiplicative processing the factor $D_E(p)$ appears at both inputs to the multiplier and the output is therefore proportional to the square of this factor. Thus for the 4 × 4 multiplicative array the static pattern is given by

$$D(p) = D_E^2(p)\left[\frac{\sin 4p}{4\sin p}\right]^2 \cos 8p \qquad \text{......(20)}$$

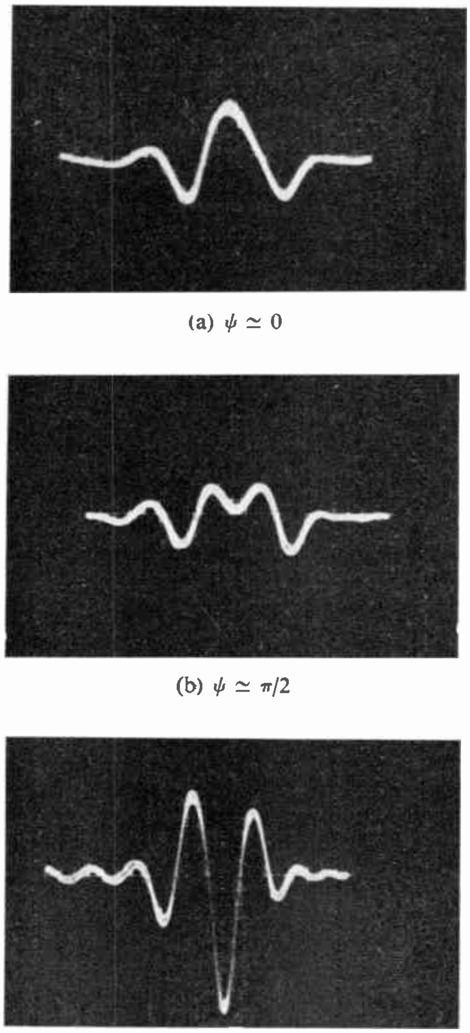
and the dynamic response will be the same with $D_E(p_1)$ in place of $D_E(p)$.



(a) 8-element additive.

(b) 4 × 4 multiplicative.

Fig. 9. Electronically-scanned directional patterns.



(a) $\psi \simeq 0$

(b) $\psi \simeq \pi/2$

(c) $\psi \simeq \pi$

Fig. 10. Two target scanned output for various values of the relative phase difference (ψ).

Figure 10 shows some experimental directional patterns for two target excitation. This is achieved by illuminating the receiving array by two small phase-locked sources. The two sources are waveguide horns fed from a common oscillator with facilities for varying the relative phase and amplitude of the two outputs. The results shown correspond to equal amplitudes and fixed phase differences. Work is at present under way to introduce controlled amount of phase variation between the two sources and to investigate the effects of given amounts of integration under such situations.

The experimental programme has provided valuable insight into the problems of designing and operating receiving arrays with multiplicative signal processing. Although it has successfully demonstrated that such a system can be made to operate there are still problems to be investigated before using such techniques for low signal levels with large dynamic ranges. Nevertheless the results so far are encouraging and if further investigations show that some target situations favour different forms of signal processing to others, it will be interesting to consider the development of forms of processing to optimize given conditions.

5. Two-target Resolution with Integration

It is intended to extend the previous experimental results to include cases with differing amounts of integration. However without entering into any further analysis it is reasonable to conclude that the use of such integration can provide further improvement in resolution. In the case of square-law rectification and multiplicative processing it can reach the limit corresponding to the superposition of single target responses. This limit cannot be reached with linear rectification though it is interesting to reflect that if we could square-law-rectify, integrate and then take the square root, the resulting processing would have the superposition advantage of square-law rectification without the square-law response.

A further interesting point relates to possible advantages to be gained from any system with a high power-law response together with a fading target situation. If the fading is not synchronous and in phase, then the strong signal capturing effect will pick up whichever is the stronger and if this alternates, the integrated result can improve the resolution.

It can therefore be concluded that integration can improve the resolution of a directional receiver with either form of demodulation provided that there is suitable variation of either relative phase ψ or relative target amplitude k.

6. Some Effects of Noise upon Resolution

6.1. *Signal/Noise Ratio*

A discussion of the comparative values of different forms of demodulation must necessarily include some mention of the signal/noise performance. Some early papers on multiplicative processing considered the practically unlimited degree of effective super-directivity that could be obtained by multiple cascaded multiplications. This knowledge is only of practical value if it is also known what effect this has upon the signal/noise ratio and multiple target cross-products.

If the signals at the two inputs to a multiplier are taken as correlated quantities and the two noise inputs taken as uncorrelated, then there are three types of product at the output of the multiplier:

(1) Signal-signal products. These are the wanted products, and in the case of unmodulated input signals consist of d.c. terms.

(2) Noise-noise products. These represent unwanted noise but the probability distribution of such noise will not be the same as the input distribution and for the case of Gaussian inputs takes the form of a Hankel function.[15]

(3) Signal-noise products. These cross-products will have a probability distribution dependent upon the probability distributions of both the input signals and noise and they may be classed either as signal or as noise depending upon application.

As a result of the above situation it is possible to define signal/noise ratio in several different ways[14-16] and some authors prefer not to introduce the concept at all but to calculate probability of detection or some such parameter. In this paper it is only intended to comment on a few relevant points and draw some general conclusions.

6.1.1. The detection of targets

If we consider the problem of detecting the presence of a target with a radar system we are looking for a change in the level of the mean value of the demodulated output of the receiver. A reasonable definition of signal/noise ratio in such a case might be

$$\frac{\text{change in r.m.s. value of waveform due to target}}{\text{r.m.s. value of waveform in absence of target}}$$

For such a case it is not difficult to appreciate that the presence of the signal-noise cross products improve the signal/noise ratio since they assist in indicating the presence of a target. Although detailed calculations on probability of detection must involve analysis of the probability distributions nevertheless for most first order comparisons it is useful to compare signal/noise ratios based on r.m.s. values.

6.1.2. Position accuracy and resolution

The accuracy of determination of the direction of a target, or the accuracy of resolution is often defined in terms of the rate of variation of the directional response at a given point on the characteristic relative to the r.m.s. noise level.[17,18] This approach is not restricted to split beam types of array. In such a case however the presence of signal-noise cross products increase the relative level of the noise background and therefore such products may be grouped with noise in the definition of a signal/noise ratio for bearing accuracy or resolution.

6.2. *The Directional Responses of Signal and Noise Demodulation Products*

Although the distinctions drawn between different forms of definition for signal/noise ratio have been described in terms of multiplicative demodulation, they apply equally to other forms. In the case of resolution performance in the presence of noise it is also necessary to consider other parameters of these demodulation products, namely their directional outputs.

If we first consider single target excitation then the directional output of the signal-signal product is the normal directional response of the array. The noise-noise output is independent of direction for both receiver and medium noise assuming an isotropic noise field in the medium. But the output of the signal-noise product will have a directional response which may well differ from that of the signal-signal product.

In the case of the multiplicative array the respective responses are given by:

$$\text{signal-signal product}\quad D(p) = D_1(p)D_2(p)\cos rp \quad \text{...(21)}$$

$$\text{noise-noise product}\quad D(p) = N_1 N_2 = \text{(non-directional)} \quad \text{...(22)}$$

$$\text{signal-noise product}\quad D(p) = \sqrt{N_2^2 D_1^2(p) + N_1^2 D_2^2(p)} \quad \text{...(23)}$$

(for uncorrelated noise inputs of r.m.s. level N_1 and N_2).

It can be seen from the above that the directional output of the signal-noise product is the r.m.s. sum of the directional responses of the two parts of the aerial used to produce the multiplicative system. For the product of two equal halves of the array, then $D_1(p) = D_2(p)$ so the signal-noise directional output is proportional to the directional response of half the array. A comparison between this response and the signal-signal response for an 8-element array is shown in Fig. 11.

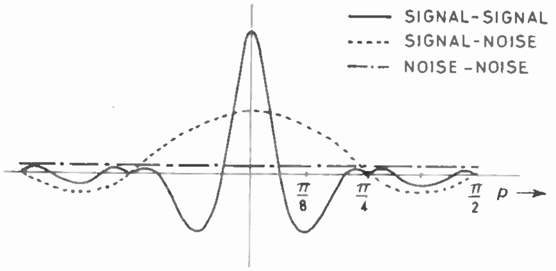


Fig. 11. Directional characteristics of the signal and noise products for multiplicative demodulation.

A similar approach can be made for multiple targets and the corresponding outputs for a multiplicative system with two targets (x and y) is given by

$$\text{s–s product} = D_1(p_x, p_y)D_2(p_x, p_y) \quad \text{......(24)}$$

$$\text{n–n product} = N_1 N_2 \quad \text{......(25)}$$

$$\text{s–n product} = \sqrt{N_2^2 D_1^2(p_x, p_y) + N_1^2 D_2^2(p_x, p_y)} \quad \text{...(26)}$$

where $D_1(p_x,p_y)$ and $D_2(p_x,p_y)$ represent the multiple target responses of the two parts of the aerial system, taking account of the phase difference and relative signal strengths of x and y.

It can be seen from the above that a detailed calculation of the resolution of two or more targets must take account of the signal/noise ratio and that the directional responses of the various signal and noise demodulation products must also be considered.

7. Some Differences in the Application of Signal Processing in the Fields of Radar, Sonar and Radio Astronomy

7.1. *Spatial Frequency Approach*

The principal differences between the application of multiplicative signal processing to the fields of radar and radio astronomy are twofold. In the first place radio astronomy is not concerned with the complex reflecting properties of targets. Secondly, the incoherent nature of the radio astronomy signal sources enables more advantage to be taken of multiplicative processing, though this feature may not apply to the same extent in the case of radar astronomy. It is clear from the previous sections that there will be little gain of resolution for a radar system due to signal incoherence, since this corresponds to targets with a high relative velocity, and obviously two targets with a high relative velocity do not remain close together, so that integration would not be available over a reasonable period.

Some papers in the radio astronomy field discuss the performance of receiving arrays in terms of spatial frequencies.[19] If the far-field source or target distribution is analysed the resultant Fourier components are termed the spatial frequency components of the target distribution $T(p)$. If some directional receiver is scanned across the far-field in order to determine the target distribution, the receiver output $S(p)$ which is given by the convolution integral of eqn. (5) represents an attempt to reproduce the far field distribution; but owing to the finite size of the aerial the accuracy of this reproduction will be limited. Therefore if a Fourier analysis is made of $S(p)$, some of the components of $T(p)$ will be of incorrect amplitude and others will be absent altogether. It is therefore convenient to think of the directional receiver as a spatial filter which has a given spatial frequency response which passes only limited components of the far-field distribution as given by the equation:

$$[S(p)] = [T(p)][D(p)] \qquad \ldots\ldots(27)$$

where the square brackets denote the Fourier transform. This equation may be derived directly from eqn. (5).

This approach is of particular interest to systems employing fast continuous scanning of the directional receiver. The output of the scanned receiver due to a point target is the directional pattern in the form of a repetitive time waveform. Therefore the frequency spectrum of the receiver output is in this case the same as the spatial frequency response of the directional receiver. The value of this result is that the directional pattern of such a scanning receiver may be varied or tapered merely by passing the output of that receiver through a frequency equalizing network. However the magnitude of the spectrum is such that in the case of a radar system this form of tapering is only practical for very fast scanning, such as within-pulse scanning.

A slight qualification is necessary regarding the use of equalizing networks to control directional patterns since there is a difference between applying the equalizing before and after demodulation. This is because in the latter case the output of the receiver also contains spectral components due to the multiple-target cross-products arising in the demodulation process.

Since most of this paper has been concerned with the differences between multiplicative and square-law additive processing, the spatial frequency responses corresponding to an 8-element array with these forms of processing is given in the appendix. The question of which directional pattern (or which spatial frequency response) most nearly approaches the optimum has no meaning without specifying the prior knowledge about the far field. It has been established that the best mean-square-error approximation to the far field is obtained by a $(\sin x)/x$ directional pattern, but for radar applications where it is required to detect small targets in the presence of large targets on other bearings, it is necessary to employ directional patterns with low side-lobes. Signal processing can therefore assist in varying the directional patterns to suit the target distribution.

7.2. *Resolution in Sonar and Radar*

A previous paper by Welsby[20] dealt with the relative resolution of multiplicative arrays and linear-law rectified additive arrays. There is a discrepancy between the results of this paper and the present analysis. The reason for this discrepancy is an ambiguity of sign of the phase difference between the two sources which arises in the geometrical approach of Welsby's Appendix 2 and this leads to a different mechanism of resolution whereby the two targets are resolved one at a time.

However, despite this difficulty, the sonar experimental results demonstrate that even in fairly calm conditions the form of the resolution closely resembles the explanation due to the incorrect analysis, the two separate targets appeared one at a time and produced a twinkling effect. Now it is clear from the analysis in the present paper together with the microwave experimental results that the multiple-target pattern is an even function of p for equal-strength targets and therefore cannot produce this effect. It is therefore reasonable to conclude that the experimental resolution in the sonar system was due principally to relative amplitude variations. This effect would be exaggerated by the square-law response.

However this effect may be of value in considering the resolution performance of a radar system employing a high power-law response. In a practical radar target situation the relative phase will not vary appreciably but it is quite reasonable to expect targets to provide significant changes in echoing area due to small changes in aspect. It therefore appears that the resolution of targets may be assisted due to target fading phenomena and it is hoped to study this matter further.

8. Conclusions

This paper has compared the angular resolving properties of linear arrays, subjected to rectification and to multiplication types of signal processing. It is shown that the resolving properties of such arrays are dependent upon the statistical properties of the relative phase difference ψ between the signals received from the two or more targets (or sources) to be resolved. If ψ does not vary with time it is shown that multiplicative processing is superior to rectification in terms of resolution of targets of equal strength but its performance is poor when the relative strength of the targets vary. There is no distinction between linear- or square-law rectification for this case.

If ψ varies with time and it is possible to integrate the resultant outputs, the resolution of all systems can be further improved, in most cases, up to a limit set by the resolution calculated on a basis of superposition of single target patterns. This limit is not attained for systems employing linear rectifiers. Attention has been drawn to the effect of relative amplitude variations between targets, it appears that in certain circumstances this may assist resolution.

An experimental multiplicative array operating in the 3-cm band has been described. This array is also capable of fast electronic scanning. Experimental static and scanned (dynamic) directional responses of this receiving system are presented for both single target and two-target excitation. Multiplicative signal processing has particular application to systems employing within-pulse scanning of a receiving beam.

The significance of signal/nóise ratio in terms of target detection and multiple-target angular resolution is briefly discussed and it is shown that in the latter case it is necessary to consider the directional output response of the separate signal-signal, noise-noise and signal-noise cross products from the demodulator.

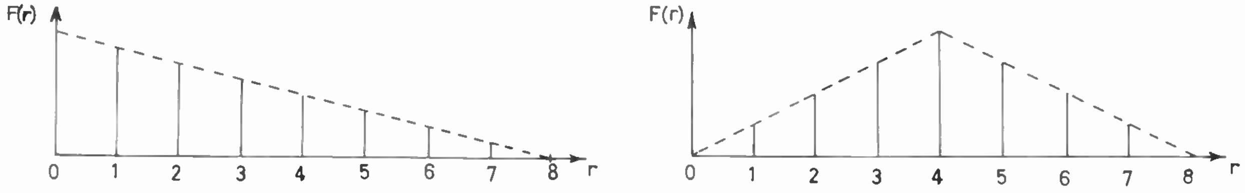
The work described in this paper has been restricted to one type of multiplicative array but since it has been shown that the first pair of negative side-lobes reduces the resolving properties of the array, it is clearly of value to develop forms of multiplicative pattern with reduced side-lobe levels. Such studies may also provide useful methods for controlling such parameters as side-lobe levels, so that these parameters may be easily varied in an operational environment to obtain an optimization of some portion of the overall performance of a directional system.

9. Acknowledgments

The authors wish to thank several colleagues for stimulating discussions on several topics related to this work, in particular Professor D. G. Tucker and Mr. A. Szabo.

10. References

1. J. J. Faran and R. Hills, "The Application of Correlation Techniques to Acoustic Receiving Systems", Acoustic Research Laboratories, Harvard University. Technical Memorandum, No. 28, November, 1952.
2. A. Berman and C. S. Clay, "Theory of time-averaged product arrays", *J. Acoust. Soc. Amer.*, **29**, p. 806, 1957.
3. B. Y. Mills and A. G. Little, "A high resolution aerial system of a new type", *Aust. J. Phys.*, **6**, p. 272, 1953.
4. M. Ryle, "A new radio interferometer and its application to the observation of weak radio stars", *Proc. Roy. Soc.*, **A211**, p. 351, 1952.
5. M. E. Pedinoff and A. A. Ksienski, "Multiple target response of data processing systems", *I.R.E. Trans. on Aerials and Propagation*, **AP-10**, No. 2, p. 112, March 1962. Also discussion on above paper by R. H. Macphie, loc. cit., No. 5, p. 642, September 1962.
6. I. W. Linder, "Application of Correlation Techniques to Antenna Systems", University of California, Electronics Research Laboratory Report Series No. 60, Issue 267, January 1960.
7. V. G. Welsby, J. H. S. Blaxter and C. J. Chapman, "Electrically scanned sonar in the investigation of fish behaviour", *Nature*, **199**, pp. 980–1, 7th September 1963.
8. J. Croney, "A New Proposal for Eliminating Side-lobe Echoes from Radar Displays", Admiralty Signal and Radar Establishment Technical Note No. AX–55–1, January 1955.
9. V. G. Welsby, "The angular resolution of a receiving aperture in the absence of noise", *The Radio and Electronic Engineer* (*J. Brit.I.R.E.*), **26**, p. 115, August 1963.
10. P. M. Woodward, "Probability and Information Theory with Applications to Radar" (Pergamon Press, London, 1955).
11. E. Shaw, "A Multiplicative Radar System with Electronic Scanning". Unpublished Report, Electrical Engineering Department, University of Birmingham, March 1963.
12. V. G. Welsby and D. G. Tucker, "Multiplicative receiving arrays", *J. Brit.I.R.E.*, **19**, p. 369, June 1959.
13. D. E. N. Davies, "A fast electronically scanned radar receiving system", *J. Brit.I.R.E.*, **21**, p. 305, April 1961.
14. D. G. Tucker, "Signal/noise performance of multiplier (or correlation) and addition (or integration) types of detector", *Proc. Instn Elect. Engrs*, **102**, Part C pp. 181–90, 1955. (I.E.E. Monograph No. 102R, February 1955.)
15. E. L. R. Webb, "Note on the product of random variables", *Canadian J. Phys.*, **40**, p. 1394, 1962.



(a) 8-element square detection.

(b) 8-element 4 × 4 multiplication.

Fig. 12. Spatial frequency characteristics of arrays with square-law and multiplicative detection.

16. M. J. Jacobson, "Output Signal/Noise Ratio of an Array Correlator with R C Averager". Rensselaer Polytechnic Institute, New York. Math. Report No. 15, June 1958.
17. M. Federici, "The precision of directional measurement of a sound source with directive receiving systems", *La Ricerca Scientifica*, No. 29, p. 2301, November 1959.
18. C. R. Fry and D. G. Tucker, "The effect of noise on the determination of direction in a multiplicative receiving system", *Proceedings of the Symposium on "Signal Processing in Radar and Sonar Directional Systems"*. Birmingham 1964, Paper No. 23.
19. R. N. Bracewell and J. A. Roberts, "Aerial smoothing in radio astronomy", *Aust. J. Phys.*, 7, pp. 615–640, December 1954.
20. V. G. Welsby, "Multiplicative receiving arrays", *J. Brit. I.R.E.*, **22**, No. 1, pp. 5–12, July 1961.

11. Appendix 1: The Spatial Frequency Characteristic of Arrays with Square-law Detection and Multiplication

The scanned output of a directional receiver is given by the convolution integral of eqn. (5). Now the function $T(p)$ representing the target or source distribution will be zero outside the range $p_0 = \pm\pi/2$ and can be expressed in the form of a Fourier series:

$$T(p_0) = a_0 + \sum_{m=1}^{\infty} a_m \cos 2mp_0 + b_m \sin 2mp_0 \ldots\ldots(28)$$

The directional pattern $D(p)$ may also be considered restricted to the range of real angles and taken as zero outside $|p| > \pi/2$. If we take $D(p)$ as an even function of p we may express its Fourier series as:

$$D(p) = \sum_{r=0}^{(n-1)} C_r \cos 2rp \qquad \ldots\ldots(29)$$

The convolution integral may then be written

$$S(p) = \sum_{m=1}^{\infty} \sum_{r=1}^{(n-1)} \int_{-\pi/2}^{\pi/2} (a_m \cos 2mp_0 + b_m \sin 2mp_0) C_r \times$$

$$\times \cos 2r(p-p_0)\,\mathrm{d}p_0 + \int_{-\pi/2}^{\pi/2} a_0 C_0\,\mathrm{d}p_0 \ldots(30)$$

The integral will yield non-zero solutions only when $m = r$ and the solution can be expressed in the form:

$$S(p) = \sum_{r=0}^{(n-1)} F(r) \frac{\pi}{2} (a_r \cos 2rp + b_r \sin 2rp) \ldots(31)$$

It is thus seen that any spatial frequencies of the target distribution higher than $m = (n-1)$ are completely lost at the receiver output. The filter-like characteristic of the array is given by the function $F(r)$ and this is plotted in Fig. 12 for the cases of 8-element arrays with multiplicative and square-law additive processing.

Manuscript received by the Institution on 16th January 1964. (Paper No. 937/RNA34.)

The discussion at the Symposium on this and two associated papers will be published in a forthcoming issue of *The Radio and Electronic Engineer*.

I.E.R.E. GRADUATESHIP EXAMINATION, MAY 1964

PASS LISTS

The following candidates who sat the May 1964 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 356 candidates. Of these 172 sat the examination at centres in Great Britain and Ireland and 184 sat the examination at centres overseas. The names of successful candidates resident in Great Britain and Ireland are published in the October/November issue of the *Proceedings* of the I.E.R.E.

	Candidates appearing	*Pass*	*Fail*	*Refer*
Section A				
Great Britain	91	37	47	7
Overseas	102	26	69	7
Section B				
Great Britain	81	27	46	8
Overseas	82	12	66	4

OVERSEAS

The following candidates have now completed the Graduateship Examination and thus qualify for transfer or election to Graduate or a higher grade of membership.

BHASIN, K. E. (S), *Bangalore, India.*

DEVGON, H. L. (S), *Bangalore, India.*

McGREAL, D. E. (S), *Winnipeg, Canada.*

RAMACHANDRAN, C. (S), *Colombo, Ceylon.*

REINER, J. (S), *Tel-Aviv, Israel.*

SAINI, B. (S), *Bangalore, India.*

SCHMITT, H. (S), *Cape Town, S. Africa.*

SUBRAMANIAN, H. (S), *New Delhi, India.*

WAKERLEY, P. A., *Bulawayo, S. Rhodesia.*

ZUGIC, V. (S), *Belgrade, Yugoslavia.*

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

BALAKRISHNAN, D., *Calcutta, India.*

CHAIKIN, J., *Haifa, Israel.*

CHERIYAN, M. C., *Calcutta, India.*

DHANARAJAN, J., *Bombay, India.*

FASHOLA, V. K. (S), *Lagos, Nigeria.*

HERATH, J. A., *Colombo, Ceylon.*

HUEN, J. C. C. (S), *Hong Kong.*

IROANYAH, C. (S), *Nigeria.*

JAGADISH, S. (S), *Bangalore, India.*

KALYANDRUG, N. M., *New Delhi, India.*

LANGDOWN, C. W. G., *B.F.P.O. 151.*

McSTAY, R. E. D. (S), *Christchurch, N. Zealand.*

MEHROTRA, M., *Lucknow, India.*

PARAMATHMA, P. (S), *Bangalore, India.*

RAMMANNA, H. S. (S), *Bangalore. India.*

ROBERTS, J., *Ndola, N. Rhodesia.*

SHAH, K. B. K. (S), *Bombay, India.*

SHAKESPEARE, G. (S), *Kirkuk, Iraq.*

SOONAWALA, N. M. (S), *New Delhi, India.*

SREEPERUMBODURI, R. R., *Calcutta. India.*

SRIDHARAN, R., *Bangalore, India.*

TOH, P. K., *Singapore.*

WESSON, P. M. (S), *B.F.P.O. 69.*

WHITE, G. P. J. (S), *Kaduna, Nigeria.*

WILLIAMS, J. W. (S), *Auckland, N. Zealand.*

YONG, L. T. (S), *Singapore.*

The question papers set in Section A of the May 1964 Graduateship Examination are published in the October/November issue of the *Proceedings* of the I.E.R.E., together with answers to numerical questions and examiners' comments. Parts 3 and 4 of Section B and Part 5 of Section B will be published in subsequent issues of *Proceedings.*

(S) denotes a Registered Student.