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 41 No. 6 JUNE 1971

THE RADIO AND ELECTRONIC ENGINEER

The Journal of the Institution of Electronic and Radio Engineers

Electronic Control of Mechanical Handling

DURING the past two decades this Institution has organized several major conferences on what is usually termed 'industrial electronics'. The programmes have dealt with applications in which electronic techniques have been applied as a means of measurement of control. The application of electronic control to mechanical handling equipment has received comparatively little consideration from a systems point of view compared with devices and circuits developed to replace or improve other non-electronic techniques. However, over the past few years a number of applications using advanced electronic techniques have been put into practice and many more are being proposed. This makes a conference on this subject timely and the first of this year's I.E.R.E. conferences therefore has the theme of Electronic Control of Mechanical Handling and it will be held at the University of Nottingham from 6th to 8th July 1971.

In the application of electronics in industry it has always been necessary for the user to express his requirements in terms understood by the electronic engineer so that the best solution to his problem is produced. In many of the current applications of electronic control in the mechanical handling field there has been a breakdown in this communication with unfortunate results. Although this state of affairs is improving it is hoped that the present conference can provide a forum for discussion which will help to bridge the gap. To this end it is appropriate that the I.E.R.E. has been joined on the organizing committee for this conference by three of its sister Engineering Institutions—Electrical, Mechanical and Production—and by the Institute of Materials Handling and the National Materials Handling Centre, Cranfield.

The three days' conference has been arranged as five logically developed sessions: Components and Subsystems; Cranes, Stackers and Palletizers; Overall Management of Mechanical Handling Systems; Routing Systems; and Future Techniques and Robots. We are fortunate in having an opening address by Professor W. B. Heginbotham, Head of the Department of Production Engineering and Production Management at the University of Nottingham, who directs research projects into many of the problems of mechanical handling. Thirty papers make up the programme and these are being contributed by authors whose organizations are representative of research and development, manufacturers of systems and components, and of course users. The majority of the papers are from British organizations, but two originate from Belgium and the United States respectively. (A list of the papers was published in the April issue of *The Radio and Electronic Engineer*.)

The conference arrangements generally follow the well-established I.E.R.E. pattern of preprinted papers presented to bring out the most important points and thereby stimulate free discussion in which the organizing committee hopes that users (and potential users) will pose their special problems and enable electronic engineers to suggest means of solution. Another important feature is that the residential nature of the conference will bring added opportunities for informal discussion. The University of Nottingham has a self-contained campus which was much appreciated by those who took part in the conference on 'Integration of Design and Production' four years ago.

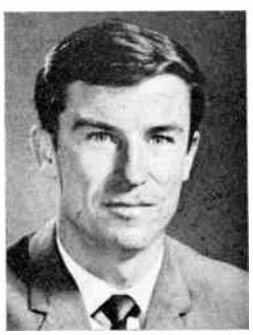
Next month's conference will be another demonstration of the wide-reaching applications of electronics. The bringing together of engineers from many industries will encourage the exchange of ideas in this important industrial role.

S. L. H. CLARKE

Contributors to this issue

Mr. John R. Brinkley (F. 1952, M. 1948) is now International Manager of mobile radio with the I.T.T. Corporation. He began his career in telecommunications with the British Post Office at Dollis Hill Research Station. During the war he was at the Home Office where he was responsible for the primary development of police, fire and civil defence mobile services. In 1958 Mr. Brinkley joined Pye Telecommunications and was successively chief engineer, technical director and managing director between 1956 and 1966. He has been responsible for many innovations in the mobile radio field and played a leading role in the introduction of 12·5 kHz channelling in the v.h.f. bands in the U.K.

Mr. Brinkley is a member of the Ministry of Posts and Telecommunications' Frequency Advisory Committee and the Mobile Radio Advisory Committee. He served as a member of the Institution's Council from 1963–66, and was a representative on a B.S.I. Technical Committee for several years. He has contributed several papers to the Institution on communications subjects and is the author of numerous articles in the technical press.



Mr. R. J. Todd (M. 1962, G. 1959) went to Australia in 1969 to take up an appointment as lecturer in the School of Electrical Engineering at the South Australian Institute of Technology. From 1952–58, he was a student apprentice with Siemens Brothers & Co. Ltd. (later A.E.I. (Woolwich) Ltd.) and gained his Graduate Membership after part-time study at the South East London Technical College. After 1958 he was employed as a telecommunications engineer at the A.E.I. Research Laboratories, Blackheath, working mainly on electronic telephone exchanges. In 1962 he joined Specto Avionics Ltd., as a senior engineer and led projects concerned with airborne navigational equipment and data recording. During 1965–66 he took a post-graduate course at the University of Aston in Birmingham and obtained an M.Sc. in Electrical Engineering. He returned to Specto Avionics Ltd. for one year before accepting an appointment as lecturer at Twickenham College of Technology where he was promoted to senior lecturer in 1968.

Mr. C. F. Ho (M. 1965, G. 1959, S. 1956) obtained his academic training and professional experience in Hong Kong, England and Canada. After graduating from Battersea College of Technology in 1959, he worked as a research engineer at the Automatic Telephone & Electric Co., Liverpool. Between 1962 and 1964 he was with the University of Manitoba, where he obtained his Master's degree. In 1965 he was employed as a senior applications engineer at Fairchild Semiconductor (H.K.) Ltd. until October 1967 when he took up an appointment as lecturer in electrical engineering in the University of Hong Kong. His current research interests are in solid-state device circuitry and network synthesis.

Dr. Jovan V. Surutka received the B.E. and D.E.E. degrees from the University of Belgrade in 1947 and 1957. From 1947 to 1951 he was a research assistant at the Institute for Telecommunications, Serbian Academy of Science, Belgrade, and he then joined the Faculty of Electrical Engineering at the University of Belgrade as a Teaching Assistant Professor until 1954, when he became an Assistant Professor. From 1951 to 1952 he held a National Education and Research Council of Yugoslavia Fellowship at the Laboratoire National de Radioélectricité, Paris, France. In 1959 he became an Associate Professor of Electromagnetics, and in 1968 a full Professor of Electrical Engineering at Belgrade and is currently Dean of the Faculty of Electrical Engineering. In addition, since 1956 he has been a consultant to the Radio-TV Broadcasting Corporation of Belgrade. He is the author or co-author of a number of technical papers and of two books, mostly on the theory and design of linear antennas and antenna arrays.

Mr. Božidar Djurich received his mathematical education at the Pedagogical Academy of Nish in 1956. After several years of teaching duties at various schools he continued his education at the Faculty of Engineering of the University of Nish where he graduated in electronic engineering in 1965. In 1966 he took up his present appointment of teaching assistant at the Department of Electronic Engineering of the University of Nish. He is at present studying for his Ph.D. in the same Department.

A note on the career of **Professor B. D. Rakovich,** who is head of the Department of Electronics in the Faculty of Electrical Engineering at the University of Belgrade, was published in the September 1970 issue of *The Radio and Electronic Engineer.*

U.D.C 535.215.6:546.289.443.23:621.383.53

The Electric and Photoresponse Characteristics of Ge/ZnSe Heterojunctions

By

J. T. CALOW, Ph.D., B.Sc.†

D. L. KIRK,
Ph.D., B.Sc., A.R.C.S., D.I.C.‡

S. J. T. OWEN,
Ph.D., B.Sc., C.Eng., M.I.E.E.‡

and

P. W. WEBB,
Ph.D., M.Sc., C.Eng., M.I.E.E.§

Wide band gap II–VI compounds deposited epitaxially upon semiconducting substrates have possible applications as solid-state infra-red detectors and imaging devices. The Ge/ZnSe heterojunction has been prepared by vacuum evaporation of epitaxial layers of zinc selenide on to orientated, single crystal, p-type, germanium substrates. Measurements have been made of the electrical characteristics, capacitance properties and photoresponse of these junctions. From these measurements a realistic band model has emerged involving intrinsic and extrinsic defects present in the bulk and interfacial region of the zinc selenide. The data presented suggest that a Mott-type barrier rather than a Schottky barrier is present at the germanium-zinc selenide interface. Techniques are described for reducing the magnitude of this Mott barrier and the resulting change in the physical properties and band structure are discussed.

1. Introduction

The results of an investigation into the electrical and photoresponse properties of the Ge/ZnSe heterojunction are presented in this paper. The system was chosen because of its possible application as a solid-state infra-red photocathode,[1] the mechanics of operation of such a device being envisaged as follows. Incident infra-red radiation falling upon heavily doped p-type germanium through an overlying epitaxial layer of n-type zinc selenide (Fig. 1), leads to the creation of a non-equilibrium concentration of electron-hole pairs at the germanium-zinc selenide interface. By an application of an electric field under reverse bias, a proportion of the electrons that were created within a diffusion length of the interface would be injected into the zinc selenide conduction band. These electrons would then be accelerated through the zinc selenide layer and emitted into vacuum, where they would be imaged on to a suitable phosphor screen.

The physical conditions involved in obtaining an epitaxial growth of single crystal zinc selenide on a germanium substrate, namely the degree of vacuum, the substrate orientation and the source and substrate temperature, have been evaluated and discussed in a previous paper.[2]

Germanium-zinc selenide heterojunctions have also been prepared by Yamoto, using an iodine disproportionation reaction,[3] but no details of their electrical characteristics have been reported. Hovel and Milnes, by using an HCl closed-volume process, have grown epitaxial layers of zinc selenide upon diffusion- and volume-doped p-type germanium.[4] They have related the electrical characteristics to impurity levels within the zinc selenide and then constructed transistor devices from such heterojunctions, with current gains as high as 0·70.[5] A recent paper by Hovel[6] has also confirmed the existence of a switching and memory action in this particular heterojunction system.[7]

However, two problems encountered in these studies are: first, obtaining reproducibility in the electric and photoresponse properties of the junction and, secondly, in explaining these properties in terms of a realistic band model that involves intrinsic or extrinsic defects that are present either in the bulk of the zinc selenide or at the heterojunction interface. An important feature of the present work is the evidence presented for the existence of a Mott[8] rather than a Schottky[9] barrier in the interface region, this barrier being formed even when the epitaxial deposition is performed from the vapour phase under ultra-high vacuum (u.h.v.) conditions. The magnitude of this potential barrier is a prime factor in controlling electron transfer across the junction and hence the relative efficiency of the device as a detector of infra-red radiation. Its formation is believed to be largely dependent upon the techniques utilized in the preparation of the germanium substrate surface. The barrier arises from adsorbed impurity atoms or ions, initially present upon the substrate surface, that are incorporated into the layers of zinc selenide deposited in the early stages of growth. The Mott barrier thus formed will be sensitive

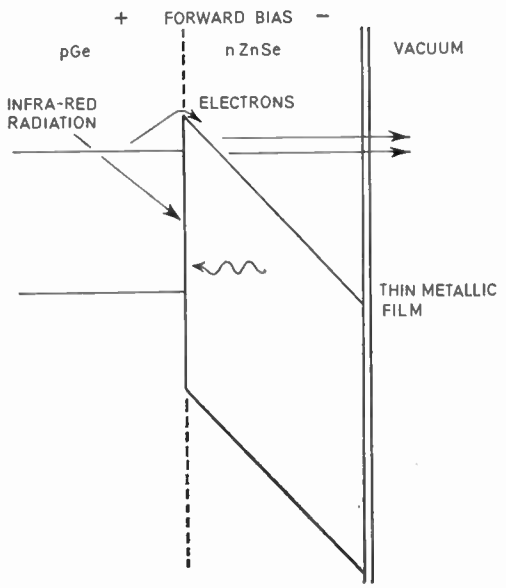


Fig. 1. Operation of the Ge/ZnSe photocathode.

† Formerly at the University of Nottingham; now with N.V. Philips, Aachen, Germany.

‡ Department of Electrical and Electronic Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD.

§ Formerly at the University of Nottingham; now with the Department of Electrical Engineering, University of Birmingham.

to both the concentration and type of impurity atom or ion incorporated into the interface. An energy band diagram based upon these general considerations has been evolved for the heterojunctions grown.

The problems of interfacial contamination and device reproducibility have been overcome by growing epitaxial layers of high-purity zinc selenide upon ultra-clean germanium substrates, the substrates having been prepared in u.h.v. conditions free from contamination by air or any form of chemical etchants. This has resulted in a layer of zinc selenide that is semi-insulating. Preliminary results are included of attempts to control the electronic properties of this layer by the controlled incorporation of an excess of zinc atoms during the epitaxial growth process.

2. Experimental

Single crystals of p-type germanium (0·1 to 0·25 ohm.cm) in a disk form approximately 2 cm in diameter and 0·5 mm thick were cut to within 2° of any specific crystallographic orientation. A plane mirror-like surface was achieved by electropolishing with a 2% sodium hydroxide electrolyte, the final surface being washed with both distilled and de-ionized water.

Growth of the zinc selenide on the germanium was performed with epitaxial deposition from the vapour phase in both conventional and u.h.v. systems. The growth source was similar to that described by Zulegg and Senkovits[10] for the evaporation of cadmium sulphide and the pressure regions in which the epitaxial growth of ZnSe was achieved may be classified as pressure regions A, B and C:

region A, 10^{-5}–10^{-6} torr (unbaked conventional systems);
region B, 10^{-6}–10^{-7} torr (baked conventional systems); and
region C, 10^{-7}–10^{-8} torr (u.h.v. system).

Base pressures as low as 1×10^{-10} torr were achieved in the u.h.v. evaporation chamber after suitable baking procedures. During growth, however, the pressure rose to 1×10^{-8} torr. The degree of mechanical perfection of the epitaxial layers was assessed by examining the zinc selenide layers using standard Laue method X-ray diffraction techniques, observations being made in both transmission and reflexion. A wholly spot-like pattern with no asterism† corresponded to an epitaxial layer approaching a high degree of order, whilst a ring pattern was taken to represent a disordered layer.

Ultra-clean substrates were prepared by having a second u.h.v. chamber which was separate from the main evaporation chamber and mounted vertically above it. This provided an environment free of zinc selenide. The two chambers were linked by a circular orifice of about 3 cm in diameter; this could be sealed off by a circular plate operated from outside the evaporation chamber. The second chamber was pumped independently by an 8 litres s^{-1} ion pump. It proved possible to maintain a base pressure of 1×10^{-10} torr in the main chamber with a pressure of 1×10^{-6} torr in the cleaning chamber. A cylindrical oven for the thermal cleaning of the substrate was housed vertically in the second chamber and enabled heat treatments up to 900°C to be achieved. The sample was suspended in the oven by hooks from a rod attached to a 20 cm linear motion drive. The linear drive enabled the sample to be moved from the cleaning oven into the main evaporation chamber prior to an epitaxial growth.

Two distinct sources of zinc selenide were utilized in the present study; zinc selenide powder being supplied by both Koch Light and Semi-Element Laboratories. The Koch Light material had a nominal purity of 99·99% and was analysed using a Debye-Scherer X-ray diffraction method. Lines other than those attributable to ZnSe were observed and by an application of Fink's index[11] were identified as ZnO. A chemical quantitative analysis of the proportion of ZnO is difficult, but the line intensities suggest that it may be as high as 5%. The Semi-Elements material was specified as 99·999% pure and was found to be free of any ZnO contamination.

Electrical and photoresponse currents were measured with a commercially available vibrating-reed electrometer (E.I.L. Model No. 33B–2). Light from a 100 W quartz iodine lamp was collected and focused using a front silvered mirror and selectively monochromated with a Hilger & Watts spectrometer (Model No. D285). Alternating photocurrents were observed by chopping the incident radiation at predetermined frequencies and detecting the photocurrent produced with a phase-sensitive detector (Brookdeal Electronics, Model No. PM332) and low noise amplifier, Model No. LA350. The photon fluxes quoted were determined with a calibrated Hilger & Watts thermopile, Model No. 2.6.1521, with a spectral range from 0·6 to 50 μm.

3. Results and Discussions

3.1 *Electrical Characteristics of Ge/ZnSe Heterojunctions*

No difficulty was encountered in making low resistance contacts to the p-type germanium. The germanium surface was sand-blasted, to increase the surface recombination velocity, and then either an indium dot was fused into the surface or contact was made directly with a DAG suspension of silver paste. A consequence of the sand-blasting was that any of the minority carriers injected across the heterojunction interface which had not recombined in the bulk would do so at the surface.

Two methods were used to obtain ohmic contacts to the ZnSe surface. The first involved the fusing of an indium dot on to the surface, under an atmosphere of hydrogen. This had the disadvantage that some of the indium evaporated over a large area. The other contacting method involved heating the heterojunction to approximately 150°C and then vacuum-depositing the indium with the contact area being delimited by suitable masks.

The electrical measurements were performed at room temperature and were restricted to devices where the ZnSe had been grown epitaxially upon the (100) germanium face. The choice of this particular orientation was dictated by the wider range of epitaxial growth

† Asterism is the radial or non-radial streaking appearing on X-ray photographs due to distortion in the crystal.

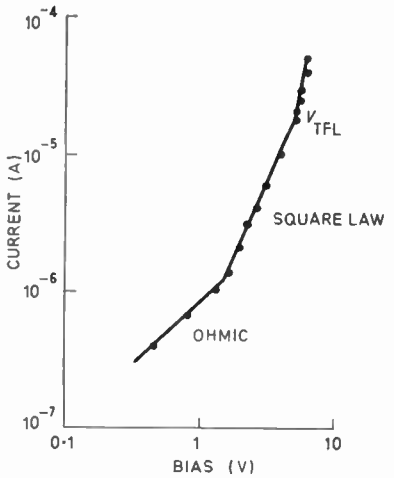


Fig. 2. Forward bias characteristic (Region B).

conditions available compared with those for the (111) and (110) faces.[2]

Two distinct types of current-voltage characteristic have been observed, depending upon the magnitude of the forward bias that had to be applied in order to cause a current of some 10^{-7} amperes to flow. Junctions grown at high substrate temperatures under conditions of poor vacuum (pressure region A) needed about 100 V applied in forward bias to produce such a current. In contrast, a few of the junctions grown at low substrate temperatures under a pressure in region B and the majority of those grown in region C only needed an applied forward bias of about 1 V. Only these latter heterojunctions will be discussed, as they have yielded the most quantitative information concerning the band diagram of this particular system. As well as observations of the I/V characteristics of such junctions, the heterojunction's capacitance as a function of the frequency of an applied a.c. electric field for different biasing voltages was also observed.

3.1.1 Forward bias (Fig. 2)

The biasing of the heterojunction is taken in the conventional sense with the forward bias having the p-type germanium held at a positive potential. Under forward bias, the behaviour of the junction is best considered in terms of a dielectric diode containing discrete shallow traps in the selenide layer. Such a situation has previously been discussed by Lampert,[12] but he makes no specification as to the nature of any potential barrier that may exist in the interface region.

(i) The characteristics of junctions grown in region B of vacuum conditions suggest that the impurity traps are the predominant feature controlling electron flow. Under a very low applied forward bias, the injected carrier concentration into the zinc selenide layer is less than the concentration of thermal carriers, so that Ohm's Law is obeyed. Upon increasing the bias voltage, the electron injection level is also raised and a square-law characteristic attributable to space-charge-limited current flow is then observed. The current and voltage are then related by the Mott-Gurney Law:[13]

$$J = \frac{9\mu\theta . \varepsilon_r \varepsilon_o V^2}{8(1+\theta)l^3} \qquad \text{.....(1)}$$

with l being the thickness of the selenide layer, μ the electronic mobility and θ the ratio of free to trapped electrons. From the observed characteristics, $\mu\theta$ was found to be 0·1 $cm^2\ V^{-1}\ s^{-1}$ and, assuming[14] an electronic mobility of 500 $cm^2\ V^{-1}\ s^{-1}$, θ was $2{\cdot}0 \times 10^{-4}$.

With a further increase in the applied forward bias, the quasi-Fermi level of the system rises until it passes through the trap level, after which the current rises rapidly to the trap-free space-charge-limited value predicted by the Mott-Gurney Law for trap-free insulators. The bias voltage at which the trap is filled ($V_{T.F.L.}$) can be related to the capacitance of the heterojunction, the total injected charge and the trap density by

$$V_{T.F.L.} = \frac{\text{total charge}}{\text{capacitance}} = \frac{eN_t . l^2}{\varepsilon_r \varepsilon_o} \qquad \text{.....(2)}$$

where N_t is the total number of traps within unit volume of the selenide. The trap-filled limit voltage suggested a trap density of $4 \times 10^{14}\ cm^{-3}$ within the bulk of the selenide. By then utilizing the relationship

$$\theta = \frac{N_c}{N_t} \exp (E_t - E_c)/kT \qquad \text{.....(3)}$$

where E_c is the conduction band edge, N_c the conduction band density of states and E_t the energy location of the trap, the trap depth was estimated to be 0·5 eV below the conduction band edge.

(ii) Devices grown under pressure region C displayed under low forward bias, a significantly different form of characteristic (Fig. 3). The

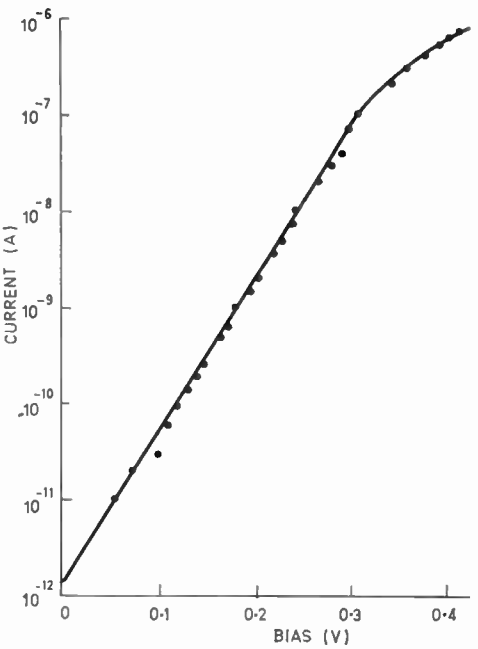


Fig. 3. Forward bias characteristic (Region C).

change in characteristic suggested that impurity traps were not now the principal factor controlling electron transport across the junction, but rather that some form of interfacial barrier had become dominant in controlling electron movement from the selenide layer into the bulk germanium. In such a situation, a simple model that invokes the kinetic emission of electrons over a step barrier[15] predicts a current-voltage relationship of the form

$$J = ne\left(\frac{kT}{2\pi m^*}\right)^{\frac{1}{2}} \exp\left[-e\left(\frac{V_D - V}{kT}\right)\right] \quad \ldots(4)$$

where n is the free carrier density, m^* the electronic effective mass and V_D the barrier height. The predicted gradient of a $\log J$ against V plot is (e/kT) which is in reasonable agreement with the experimentally observed value of $(e/1{\cdot}1kT)$. Such a feature, under sufficiently low bias, may have also been present in the electrical characteristics of the devices grown under vacuum condition B.

By assuming the resistance of the zinc selenide to be uniform up to the interfacial barrier, it was possible to estimate the resistivity (ρ) from the relationship

$$R = \frac{l.\rho}{A} \quad \ldots\ldots(5)$$

Using values for the thickness l of 12 μm and for the area A of 6×10^{-7} m^2 the resistivity was deduced to be 10^6 ohm cm, with a corresponding free carrier concentration of 10^{10} cm^{-3}. From this value and equation (4) an estimated value for V_D of 0·46 V was obtained. This was compatible with the observed current-voltage characteristic of the junction, which departed from a logarithmic dependence at approximately 0·4 V and assumed a square law from at around 0·6 V.

3.1.2 Reverse bias

The reverse bias measurements for junctions grown under vacuum conditions C are presented in Fig. 4. Electron flow will be controlled by the magnitude of the potential barrier that exists between the germanium Fermi level and the zinc selenide conduction band. A model that invokes the thermionic emission of electrons over a potential barrier may be applied to such a situation. A derivation of the characteristic consists of three stages.

(i) Calculating the number of electrons that have sufficient thermal energy to surmount the energy barrier ϕ that exists between the germanium Fermi level and the zinc selenide conduction band. For such a process, a predicted current density of the form

$$J = \frac{4\pi m^* e(kT)^2}{h^3} \cdot \exp\left[-\frac{(\phi)}{kT}\right] \quad \ldots\ldots(6)$$

is obtained.[16]

(ii) Making an allowance for the Schottky effect,[17] which leads to a lowering in the magnitude of the energy barrier ϕ. A lowering that is a result of the combined effects of the image force potential and applied electric field acting upon an electron moving from the germanium into the bulk of the zinc selenide.

(iii) Including in the calculation any potential barrier that is present in the region of zinc selenide lying close to the interface. Such a feature may play a significant role in determining the effective magnitude of the potential barrier that controls electron flow under a reverse bias situation.

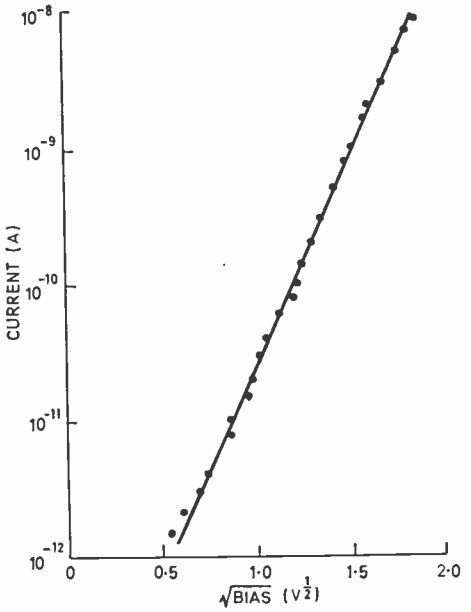


Fig. 4. Reverse bias characteristic (Region C).

If the combined effects of (i) and (ii) are considered, then the predicted current flow becomes

$$J = \frac{4\pi m^* e(kT)^2}{h^3} \cdot \exp\left[-\left(\frac{\phi}{kT}\right)\right] \times \exp\left\{\left[\frac{E_L \cdot e}{4\pi\varepsilon_o\varepsilon_r}\right]^{\frac{1}{2}} \cdot \frac{e}{kT}\right\} \quad \ldots\ldots(7)$$

where E_L represents the local electric field in the interfacial zinc selenide. If the applied electric field was then uniform across the selenide layer, E_L may be replaced by V/l (where V and l are respectively the applied voltage and thickness of the material) and J would be proportional to $\exp V^{\frac{1}{2}}$. The reverse bias characteristics are indeed of this form. Such an implied electric field distribution rules out the possible existence of a Schottky barrier in the interfacial selenide the electric field in a Schottky barrier being of the general form

$$E = \frac{eN_D(x-\delta_p)}{\varepsilon_r\varepsilon_o} \quad \ldots\ldots(8)$$

where x is the displacement from the Ge/ZnSe boundary and δ_p the depletion width. Further, the gradient of a graph of $\log_{10}J$ against $V^{\frac{1}{2}}$ is, from equation (7),

$$\left|\frac{e}{4\pi\varepsilon_o\varepsilon_r l}\right|^{\frac{1}{2}} \frac{e}{kT} \cdot \frac{1}{2{\cdot}303} \quad \ldots\ldots(9)$$

At room temperature, this equals $(1{\cdot}9\times10^{-4})/l^{\frac{1}{2}}$. The gradient of Fig. 4 is 3·07, with a calculated thickness

of 4×10^{-3} μm. The measured thickness of the zinc selenide layer was 12 μm. This suggests that there is a high resistivity, high field, layer of selenide across which most of the applied voltage is dropped and in which the electric field is uniform up to the germanium. Such a feature is characteristic of the formation of a Mott[8] rather than a Schottky potential barrier in the interfacial region of zinc selenide. This differs from the ideal Schottky barrier in that it arises from having an abrupt change in donor concentration going from a high to a low value within the bulk of a semiconductor.

It may be concluded that equation (7) adequately describes the electrical characteristic of the Ge/ZnSe heterojunction under an applied reverse bias. By extrapolation of the characteristic to zero bias, the barrier height ϕ between the Fermi level in the germanium and the top of the barrier in the selenide was estimated to be 1·0 eV.

3.1.3 Capacitance measurements

For devices grown in region C, the variation of the Ge/ZnSe heterojunction's capacitance with d.c. biasing level as a function of the frequency of a superimposed a.c. field is presented in Fig. 5. No variation in the device capacitance with changing frequency was observed in the reverse bias situation, although the capacitance did drop with increasing frequency under forward bias. If the barrier had been of a Schottky type, a variation of device capacitance with reverse bias of the form $(1/C^2)$ proportional to V would have been expected. The observed behaviour indicates that the thickness of the barrier region has remained independent of bias voltage. This is most readily explained by again postulating the existence of a Mott-type potential barrier.

The behaviour of the device with respect to a.c. electric fields is best considered in terms of an equivalent circuit, Fig. 5.

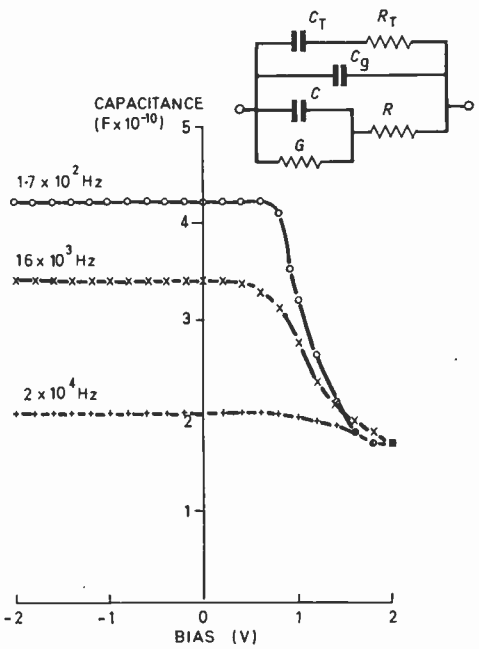


Fig. 5. Heterojunction capacitance as a function of d.c. bias voltage and as a function of the frequency of a superimposed a.c. voltage.

There are two distinct contributions to the complex impedance of this particular junction; namely, one component attributable to trap relaxation and the other arising from the potential barrier existing at the interface. Trap effects can be represented by a capacitance (C_T) in series with a resistance (R_T), whilst the effect of a potential barrier may be represented by a barrier capacitance (C), conductance (G) and a bulk series resistance (R). A geometrical capacitance due to the bulk dielectric properties of the ZnSe (C_g) is also included to complete the equivalent circuit. Under forward bias, the potential barrier is effectively removed and with it the capacitive effects of the barrier. Ancillary results,[7] have shown that the barrier effects are predominant at frequencies below 2×10^4 Hz, whilst trap relaxation effects become significant above 10^5 Hz. The best fit to the experimental results for a particular heterojunction was obtained with:

$$C = 2{\cdot}5 \times 10^{-10}\ \text{F} \qquad G = 10^{-10}\ \Omega^{-1}$$
$$C_T = 1{\cdot}5 \times 10^{-10}\ \text{F} \qquad R_T = 1{\cdot}25 \times 10^{3}\ \Omega$$
$$R = 2 \times 10^{5}\ \Omega \qquad C_g = 2 \times 10^{-11}\ \text{F}$$

The trap relaxation time $\tau = C_T . R_T$ was used to estimate the depth of the trap below the zinc selenide conduction band edge by utilizing the relationship:

$$\tau = \tau_o \exp E_t/kT \qquad \text{.....(10)}$$

where E_t is the trap depth and τ_o is approximately 10^{-9} s.[18] With an experimental value of $1{\cdot}9 \times 10^{-7}$ s for τ, a trap depth of 0·13 eV was deduced. Utilizing the previous value of θ, an approximate estimate for the trap concentration of $N_t = 2 \times 10^{19}$ cm^{-3} was obtained from the expression

$$\theta = \frac{N_c}{N_t} \exp - \frac{(E_t - E_c)}{kT} \qquad \text{.....(11)}$$

From the values obtained for the barrier capacitance, the thickness of the barrier region was estimated to be between 0·1 and 1 μm.

The values of E_t and N_t obtained from the electrical characteristics and from the capacitance measurements are significantly different. Group II–VI semiconducting compounds are often characterized by having appreciable concentrations of donor or acceptor like centres present in the forbidden gap. This is either a consequence of the poor chemical quality of the initial starting material and/or of the subsequent growth conditions. The electrical current-voltage characteristics will be governed by the impurity levels appearing immediately above the Fermi level of the selenide. In contrast, over the range of frequencies in which capacitive effects are observed, only shallow donor impurity levels associated with their characteristic relaxation times will make any contribution to the junction's capacitance. For devices grown under vacuum conditions C, the electrical characteristics suggested that the deeper trap had been removed by the improvement in vacuum conditions and had been replaced by a different impurity centre.

The current-voltage characteristics and capacitance measurements now give evidence for a basic band diagram of the form presented in Fig. 6. A potential barrier of about 0·50 eV in magnitude appears to be located between 0·1 and 1 μm from the germanium surface in the selenide layer. The precise shape of the

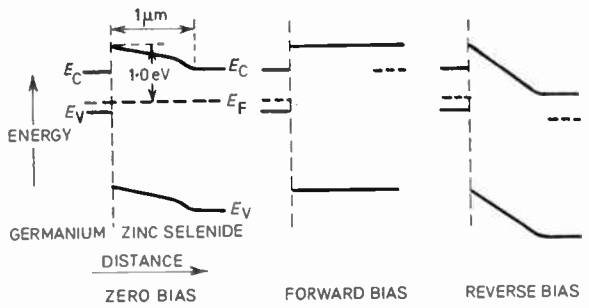


Fig. 6. The basic band diagram of the Ge/ZnSe heterojunction.

barrier profile is unknown, but probably follows that suggested by Mott. This form of potential barrier was first invoked to explain the electrical characteristics of copper oxide rectifiers. A Mott potential barrier requires an abrupt change in donor concentration within the bulk of the zinc selenide. The way that this occurs in practice still has to be considered.

Because the evaporation source was normally outgassed ten minutes prior to a growth, it is not expected that the composition of the vapour species impinging upon the germanium substrate will suddenly change. That is the donor concentration in the vapour beam should remain constant. An effective change in donor concentration could arise from a change in the trap density. If the selenide near the interface contained a large concentration of trap states, these could remove free electrons from the conduction band leading to a profile approximating to that of a Mott barrier. There are three ways in which this could occur during the epitaxial growth process:

(i) A temporary increase in substrate temperature upon opening the shutter could lead to a preferential evaporation of zinc from the substrate. This is unlikely, since selenium has the higher vapour pressure.

(ii) A chemical reaction between germanium and zinc selenide could produce a compound layer. Whilst this may occur over the first few atomic layers, it is not expected to extend as far as 0·1 μm into the bulk zinc selenide.

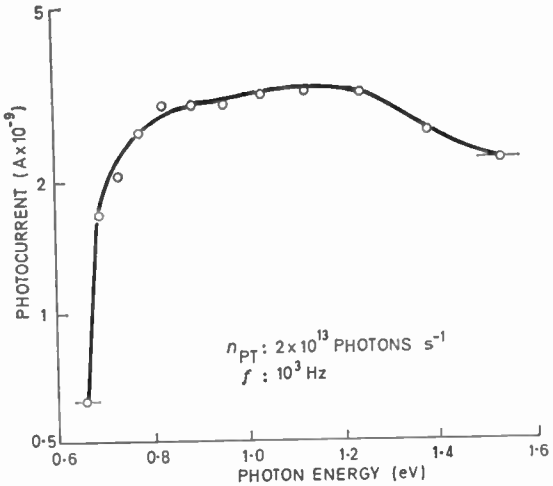


Fig. 7. A.c. photoresponse.

(iii) For the junctions discussed previously and for reported work on this heterojunction system,[3-6] no attempt has been made to produce ultra-clean germanium surfaces on which to deposit epitaxial semiconducting layers. It is possible that surface-absorbed atoms or ions present on the substrates are incorporated into the growing films, producing defect centres. The effect could extend a considerable distance into the bulk if the diffusivity of the atoms were sufficiently high at the growth temperature. A reasonable concentration of surface absorbed atoms would be about 10^{14} cm^{-2}. If these were incorporated within the first μm of zinc selenide, then the defect concentration would be 10^{18} cm^{-3}.

3.2 *Photoresponse Measurements of the Ge/ZnSe Heterojunction*

The a.c. and d.c. photoresponse of those junctions grown upon substrates that had been electropolished with the epitaxial deposition occurring under vacuum conditions C are presented in Fig. 7 and Fig. 9 (2). The principal features and conclusions arising from these investigations were:

3.2.1 A.c. photoresponse (Fig. 7)

(i) A photoresponse was first observed at a photon energy corresponding to the direct band gap of germanium. This remained constant up to a photon energy of 1·25 eV, after which it fell to approximately two-thirds of its maximum value at 1·6 eV. The effect thus originates primarily in germanium.

(ii) The photoresponse varied linearly with chopping frequency, suggesting that the photocurrent was a displacement current.

(iii) Under zero applied bias, there was an appreciable photoresponse. This increased sub-linearly with increasing reverse bias. That is, there was an associated photovoltage originating at the germanium surface. The photocurrent also increased sub-linearly with photon flux, tending towards saturation at a total photon flux of 7×10^{14} s^{-1}.

These effects may be explained in terms of band bending at the germanium surface. When the germanium surface is illuminated with photons whose energy is greater than the band gap, excess carriers are generated at the surface. These are separated by the built-in electric field, described by the built-in potential V_D. The separation of charge carriers induces a photovoltage V_p, which tends to compensate the diffusion potential. The maximum photovoltage is then equal to the built-in

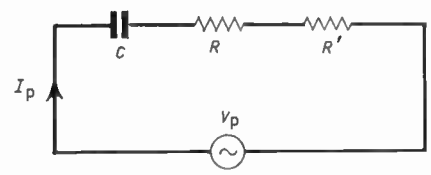


Fig. 8. Equivalent circuit for photocurrent.
R = bulk series resistance
R' = resistance across phase-sensitive detector input terminals.

voltage. Since the effect is observed across a capacitance, only the change in photovoltage can be measured. For a maximum value of I_p (the a.c. photocurrent) of 2×10^{-8} A at a total photon flux of 7×10^{14} s^{-1} (2×10^{16} $cm^{-2}\,s^{-1}$) and an equivalent circuit of the form shown in Fig. 8, it can be shown that:

$$I_p = \omega.C.V_p \qquad \text{.....(12)}$$

The corresponding photovoltage is then about $1{\cdot}3\times10^{-2}$ V. The direction of band bending was deduced as being downwards, since the germanium surface became negatively charged under illumination. That is, the internal field in the space charge region retained excess electrons at the surface and drove holes into the bulk. This conclusion is strengthened by an observed increase in photocurrent with increasing reverse bias, since the applied field at the germanium surface assists the band bending in charge separation.

It is possible that the photovoltage could have arisen because of the Dember effect.[19] When the surface of a semiconductor is illuminated with strongly absorbed photons, a high concentration of electron-hole pairs is produced at the surface. The carriers then diffuse away from the surface under the influence of the concentration gradient. In the case of germanium, electrons have the higher mobility and diffuse more rapidly than holes, setting up a negative charge upon the unilluminated surface. The corresponding potential opposes further electron flow into the bulk and is in the opposite sense to the observed barrier photovoltage.

A correction for the Dember potential may be applied by using:

$$V = \frac{kT}{e}\frac{b-1}{b+1}\log_e\frac{\sigma_o+d\sigma_{(o)}}{\sigma_o+d\sigma_{(d)}} \qquad \text{.....(13)}$$

where $b = \mu_n/\mu_p$,

σ_o = extrinsic conductivity,

$d\sigma_{(o)}$ and $d\sigma_{(d)}$ are the excess conductivity at the illuminated and unilluminated surfaces respectively.

The maximum value of the Dember potential may be estimated by setting $d\sigma_{(d)}$ equal to zero. The excess conductivity at the illuminated surface is then given by:

$$d\sigma_{(o)} = e(\mu_n+\mu_p)n_p.a.\tau' \qquad \text{.....(14)}$$

For germanium, the absorption coefficient, a, is about 10^4 cm^{-1}, the excess carrier life-time, τ', is around 10^{-4} s, and $(\mu_n+\mu_p)$ is of the order of 5×10^3 $cm^2\,V^{-1}\,s^{-1}$. The extrinsic conductivity of the germanium was about 7 $(ohm.cm)^{-1}$ and the photon flux 2×10^{16} $cm^2\,s^{-1}$. The correction for the Dember effect amounted to 9×10^{-3} V, which should be added to the observed photovoltage; this results in a barrier photovoltage of about 2×10^{-2} V. Although complete saturation of the photovoltage has not been observed, it is reasonable to take the above value as an estimate of the minimum built-in potential at the germanium surface.

3.2.2 D.c. photoresponse (Fig. 9(2))

In the idealized situation of Figs. 1 and 6 the d.c. photoresponse under reverse bias should be attributable to the direct excitation of electrons from the germanium valence band to states high in the germanium conduction band. Some photo-excited carriers would then cross the interface into the zinc selenide conduction band, causing a photocurrent. Physically the situation is analogous to photoelectric emission from a semiconductor into vacuum. The photocurrent I_p is related to a threshold energy E_T[20] by

$$I_p \propto (h\nu - E_T)^r \qquad \text{.....(15)}$$

where r takes integral or $\frac{1}{2}$ integral values between 1 and 5/2 depending upon where the excitation occurs from and whether the electron suffers phonon scattering; E_T is a function of the band gap, electron affinity and the Fermi level of the emitting material. Clearly the photocurrent of curve 2 in Fig. 9 rises too rapidly with increasing photon energy to be described by the above expression. Also the d.c. photocurrent increases linearly with reverse bias (Fig. 10), whereas the current would be expected to saturate when all the photo-excited electrons were collected by the applied electric field.

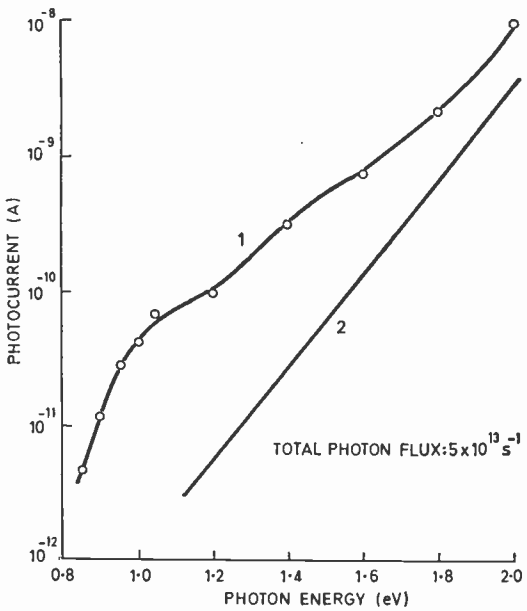


Fig. 9. D.c. photoresponse.

Such a behaviour is only explicable in terms of the release of electrons from deep lying trap states that help form the Mott barrier in the region of zinc selenide close to the interface. Undoubtedly photo-excitation from the germanium valence band and trap excitation will both occur, but the observed photoresponse suggests that the latter is the dominant mechanism yielding photocarriers, at least with photon energies ranging from 1·1–1·87 eV. The linear photocharacteristic also suggested that within the Mott barrier the trap density increased exponentially towards the zinc selenide valence band edge. However, the range of energy of the exciting light used represents a fraction of the total band to band energy transition of the wider gap material and it would be unwise to make any assertion as to the energy distribution or density of traps lying below the Fermi level of the zinc selenide.

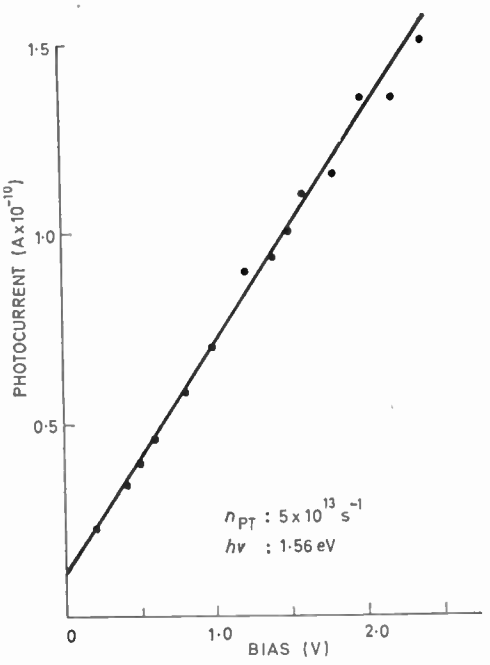


Fig. 10. D.c. photoresponse as a function of reverse bias.

3.3 *Heterojunctions Grown with High Purity ZnSe Deposited upon Ultra-Clean Germanium Surfaces*

For the epitaxial deposition of zinc selenide on to germanium substrates that had been thermally cleaned under u.h.v. conditions, the growth material was changed from Koch Light 99·99% pure zinc selenide to Semi-Elements 99·999% pure material, the latter being free of any ZnO contamination. The use of a higher purity starting material did not change the epitaxial growth condition in terms of source and substrate temperatures, but it did result in an increase in the observed growth rate. It also changed the surface morphology of the deposited layers; layers that were grown from the Koch Light material exhibited surfaces with heavy faceting. This was in contrast to the smooth layers grown from the Semi-Elements material. A possible explanation for this effect is that in the appropriate samples, trace amounts of ZnO are coming over in the vapour beam and that these are then acting as preferential sites for nucleation of the growth of the layer. If ZnO is appearing within the selenide layer it could play a significant role in controlling the electronic transport properties either across the interface or within the bulk of the zinc selenide. A similar growth feature has been observed by other workers for the auto-epitaxial deposition of silicon upon silicon[21, 22] where surface contamination by SiC (as detected with Auger spectroscopy) encouraged heavy faceting of the deposited layers of silicon. This was in marked contrast to the uncontaminated surfaces which allowed facet-free growths to be achieved.

Thermal cleaning of the germanium surface was achieved by annealing the substrate for 1 hour at 800°C in the ancillary chamber under u.h.v. conditions. This resulted in thermal etching of the germanium surface with the appearance of deep square etch pits. These pits did not affect the epitaxial overgrowth, for microscopic examination failed to reveal any correlation between surface features appearing upon the selenide layer with those present on the germanium surface that were visible through the selenide layer.

Heterojunctions grown using the higher purity starting material on the pre-cleaned substrates exhibited no detectable photo-electric response. The epitaxial selenide layers were also characterized by having a very high bulk resistivity, estimated as being as high as 10^{10}–10^{11} ohm.cm. This suggested that the material was either well compensated, with the free carrier concentration approaching that expected in an intrinsic material or that the Semi-Elements zinc selenide contained far fewer donor-like impurities than the Koch Light powder. Because of the high resistivity, it proved impossible to make an efficient electron injecting/extracting contact to the zinc selenide layer, thus preventing any study of the electrical characteristics of such devices.

A successful attempt to lower the resistivity of the selenide layers was achieved by incorporating into the layer a non-stoichiometric concentration of excess zinc during the growth process. This was achieved by adding 10 mg of high purity zinc to 2 g of zinc selenide placed in the evaporation tube. The source was then run at the evaporation temperature for a short period prior to the growth in order that the zinc evaporation would obtain equilibrium with the evaporation of the zinc selenide. This technique resulted in the deposited layers having a significantly lower resistivity (approximately 10^9 ohm.cm) and an enhanced photoresponse, curve 1 of Fig. 9, and quantum efficiency in the infra-red shown in Table 1.

Table 1. Comparison of quantum efficiencies

Photon energy eV	Device 1†	Device 2‡
0·825	$5{\cdot}5 \times 10^{-7}$	—
1·00	$5{\cdot}0 \times 10^{-6}$	$1{\cdot}6 \times 10^{-7}$
1·20	$1{\cdot}5 \times 10^{-5}$	$7{\cdot}5 \times 10^{-7}$
1·40	$3{\cdot}7 \times 10^{-5}$	$3{\cdot}7 \times 10^{-6}$
1·60	$9{\cdot}0 \times 10^{-5}$	$1{\cdot}7 \times 10^{-5}$
1·80	$7{\cdot}5 \times 10^{-4}$	$7{\cdot}5 \times 10^{-5}$
2·00	$1{\cdot}2 \times 10^{-3}$	$3{\cdot}7 \times 10^{-4}$

† Device 1: Epitaxial ZnSe deposited from Semi-Elements material (99·999% pure) upon ultra-clean germanium substrates.
‡ Device 2: Epitaxial ZnSe deposited from Koch Light material (99·99% pure) upon electropolished germanium substrates.

This method of doping is considered to be superior to that of diffusion doping in one of two ways. First, the incorporation of the zinc atoms occurs under u.h.v. conditions without any zinc impurity able to diffuse into the germanium substrate. Secondly, no diffusion annealing of the device is required after the growth process. This avoids any change in the heterojunction's properties that may arise from such a heat treatment.

The photoresponse as well as being enhanced does show some structure on the low-energy side, suggesting that photoexcitation of electrons from the germanium Fermi level and from trap states in the Mott barrier may both be occurring.

The shift of the photoresponse characteristic to lower energies also implies that there has been a lowering in the barrier height ϕ between the germanium Fermi level and conduction band states in the zinc selenide layer.

4. Conclusions

Detailed electrical and photoelectric measurements have been made upon devices where the ZnSe layer has been grown epitaxially upon the (100) face of p-type germanium, under varying degrees of vacuum. Figure 11 presents the band diagram of the Ge/ZnSe heterojunction. This essentially summarizes the interpretation of the results obtained from the heterojunctions grown. The band structure is a composition of four regions each of which plays an important part in controlling the electrical and photoresponse characteristics of the heterojunction

(i) *The Zinc Selenide Bulk*

This is of a high resistivity, about 10^6 ohm.cm, and contains a large number of shallow trap states. The concentration and energy location of these traps appears to be sensitive to the choice of growth material and to the degree of residual vacuum under which the epitaxial deposition occurs. The properties of the bulk zinc selenide dominate the forward bias characteristic at high bias, and lead to a space-charge-limited current flow.

(ii) *The Zinc Selenide Present at the Interface*

The resistivity of the material at the interface is very much higher than the bulk resistivity due to the presence of deep-lying trap states. This leads to the formation of a Mott-type potential barrier of about 0·5 eV height in the zinc selenide conduction band. Its effect has been observed under low forward bias and under reverse bias. It is believed to occur because adsorbed gases on the germanium surface are incorporated within the first micrometre of the zinc selenide growth. Photoresponse measurements give evidence for the presence of deep-lying traps within this region.

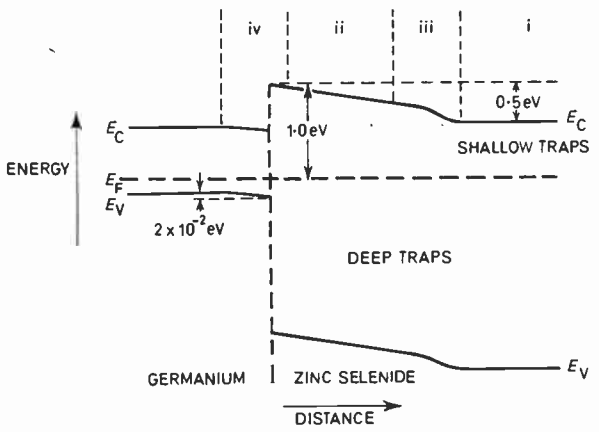


Fig. 11. The band diagram of the Ge/ZnSe heterojunction.

(iii) *The Transition between the Two Zinc Selenide Regions*

This is characterized by a redistribution of charge around a transition region. The width of the transition region x, is related to the barrier height V_D and to the ionized donor density N_D by:

$$x^2 = \frac{2\varepsilon_r \varepsilon_o V_D}{eN_D} \qquad \text{.....(16)}$$

Since the total width of the barrier region is less than 1 μm, the width of the transition region will be somewhat less. The lower limit of the ionized donor density may be estimated by taking x equal to 1 μm and V_D equal to 0·5 V, and is about 4×10^{14} cm^{-3}. This is considerably greater than the number of ionized donors in the bulk and is probably due to the ionization of traps within the transition region. In the bulk, these would normally be below the Fermi level, but because of the potential step some may be raised above the Fermi level and ionized.

(iv) *The Interface between the Germanium and the Zinc Selenide*

Reverse bias measurements have suggested a barrier height of about 1·0 eV between the germanium Fermi level and the zinc selenide conduction band. The photoresponse measurements have shown that the germanium energy bands are bent downwards at the interface, the amount of bending being approximately 2×10^{-2} eV.

It now appears possible to modify the magnitude of the Mott barrier occurring at the Ge/ZnSe interface by growing the epitaxial selenide layers upon ultra-clean germanium substrates. A method has also been evolved for controlling the electrical properties of the bulk zinc selenide by the incorporation of an excess of zinc into the selenide layers during the growth procedure. The specific treatment described has resulted in a heterojunction with a greatly increased quantum of efficiency in the 1·2–0·8 eV region of the electromagnetic spectrum. However, the exact mechanism for this improvement, as yet, is not fully understood.

5. Acknowledgment

The authors would like to thank the Ministry of Defence (Naval Department) for the generous financial support of this research and for permission to present this paper.

6. References

1. Calow, J. T., Owen, S. J. T. and Webb, P. W., 'The Ge/ZnSe Heterojunction'. Report No. H.2.66 (Univ. Nottingham, 1967).
2. Calow, J. T., Owen, S. J. T. and Webb, P. W., 'The growth and electrical characteristics of epitaxial layers of zinc selenide on germanium', *Phys. Stat. Sol.*, **28**, pp. 295–303, 1968.
3. Yamoto, T., 'Growth of the Ge/ZnSe heterojunction with an iodine disproportionation technique', *Japanese J. Appl. Phys.*, **4**, p. 541, 1964.
4. Hovel, H. J. and Milnes, A. G., 'The electrical characteristics of *n* ZnSe -*p*Ge heterodiodes', *Int. J. Electronics*, **25**, pp. 201–18, 1968.
5. Hovel, H. J. and Milnes, A. G., 'The ZnSe-Ge heterojunction-transistor', *I.E.E.E. Transactions on Electron Devices*, **ED-16**, pp. 766–74, 1969.

6. Hovel, H. J., 'Switching and memory in ZnSe-Ge heterojunctions', I.E.E.E. Device Research Conference, April 1970, Abstract No. RC 2864.
7. Calow, J. T., 'The ZnSe-Ge Heterojunction', Ph.D. Thesis, (University of Nottingham, 1970).
8. Mott, N. F., 'The theory of crystal rectifiers', *Proc. Roy. Soc.*, A, **171**, pp. 27–38, 1939.
9. Schottky, W., *Z. Physik*, **113**, p. 367, 1939. As discussed by Dekker, A. J., 'Solid State Physics'. (Macmillan, London, 1962).
10. Zulegg, R., and Senkovits, E. J., Abstract 95, Spring Meeting Electrochem. Soc., 1963.
11. Fink, A. J., 'The Inorganic Index to the Powder Diffraction File'. American Soc. Testing Materials, 1963.
12. Lampert, M. A., 'Simplified theory of space charge limited currents in an insulator with traps', *Phys. Rev.*, **103**, pp. 1648–56, 1956.
13. Mott, N. F. and Gurney, R. W., 'Electronic Processes in Ionic Crystals'. (Clarendon Press, Oxford, 1942.)
14. Aven, M. and Prener, J. S., 'The Physics and Chemistry of II–VI Compounds'. (North Holland, Amsterdam, 1967.)
15. Henisch, H. K., 'Rectifying Semiconducting Contacts'. (Oxford University Press, 1957.)
16. Kittel, C., 'Introduction to Solid State Physics'. (Wiley, New York, 1961.)
17. Schottky, W., *Z. Physik*, as discussed by Solymar, L. and Walsh, D., 'Lectures on the Electrical Properties of Materials'. (Clarendon Press, Oxford, 1970.)
18. Wang, S., 'Solid State Electronics'. (McGraw-Hill, New York, 1966.)
19. Many, A., Goldstein, Y. and Grover, N. B., 'Semiconductor Surfaces'. (North Holland, Amsterdam, 1965.)
20. Kane, E. O., 'Theory of photoelectric emission from semiconductors', *Phys. Rev.*, **127**, pp. 131–41, 1962.
21. Joyce, B. A., Bradley, R. R. and Booker, G. R., 'A study of nucleation in chemically grown silicon films using molecular beam techniques', *Phil. Mag.*, **15**, pp. 1163–67, 1967.
22. Watts, B. E., Bradley, R. R., Joyce, B. A. and Booker, G.R., 'Nucleation rate measurements and the effect of oxygen on epitaxial growth behaviour', *Phil. Mag.*, **17**, pp. 1167–87, 1967.

Manuscript received by the Institution on 27th October 1970. (Paper No. 1386/CC101.)

The Authors

Dr. J. T. Calow obtained his B.Sc. degree in physics in 1965 at the University of Nottingham. He then undertook research in the field of heterojunction phenomena within the Electrical and Electronic Engineering Department of the same University. He obtained his Ph.D. degree in 1970, and is now with the Philips organization at Aachen in Germany.

Dr. D. L. Kirk obtained his B.Sc. and A.R.C.S. degrees in 1964 in the Physics Department of Imperial College, London. His associateship was obtained by specializing in metal physics. Doctorate work was undertaken within the Metallurgy Department of Imperial College, and resulted in the award of Ph.D. and D.I.C. degrees in 1967–68. Two years were then spent as a postdoctoral fellow at the Clarendon Laboratory, Oxford, working on radiation induced defects in ceramic solids. He was appointed a lecturer in the Materials Group of the Department of Electrical and Electronic Engineering, Nottingham University, in 1969.

Dr. S. J. T. Owen graduated with a B.Sc. in physics in 1957 and was awarded the degree of Ph.D. in 1961, both at the University of Nottingham. He then joined the Department of Electrical and Electronic Engineering and was responsible for initiating teaching and research work in solid-state electronics. The development of the research has concentrated on semiconductors and their application in solid-state devices and in particular studies have been made of epitaxial gallium arsenide and semiconductor heterojunctions. Dr. Owen spent 1968 and 1969 as Visiting Professor of Electrical Engineering at the University of Alabama and was awarded a contract to do research at the N.A.S.A. George C. Marshall Space Flight Center, Huntsville.

Dr. P. W. Webb graduated from the Department of Electronic and Electrical Engineering at the University of Birmingham with a B.Sc. in 1959, an M.Sc. in 1961, and a Ph.D. in 1964. He then joined the Department of Electrical and Electronic Engineering at the University of Nottingham, where his researches concerned the electrical properties of epitaxial films of II–VI compounds on germanium. In 1969, Dr. Webb returned to Birmingham University where he has been concerned in the development of solid-state electronics and the design of integrated circuit devices.

U.D.C. 621.521

Non-linear Control System Stability Investigation using the Circle Criterion

By

C. F. HO, D.C.T.(Batt.), M.Sc., C.Eng., M.I.E.E., M.I.E.R.E.†

The circle criterion provides a convenient method of testing stability in non-linear feedback systems. The need for computer computation of the relative stability conditions for the Nyquist plot can be avoided by using a logarithmic gain-phase plot on the Nichols chart. The results of this technique agree within about 1% with those obtained from a computer solution of the Nyquist plot.

1. Introduction

While the relevant theories on automatic control systems had been developed well before the Second World War, widespread applications of linear automatic control techniques came only after that period. In 1950, with the discovery of the root-locus method by Evens, the development of linear control theory for single-input, single-output, time-invariant systems was essentially complete.

Due to mathematical complexity of even the simplest non-linear systems, non-linear control theory did not receive much attention until the early fifties. Before 1950, only the phase-plane method, suitable for analysing second-order systems was available. Then, there were independent attempts to adapt approximation methods such as those of Krylov and Bogoliubov to non-linear system analysis. The describing function method was the result. It remains to this day one of the most versatile approximation methods in control engineering.

In the late 1950s the work of Lyapunov was re-discovered and in particular, his so-called second method attracted widespread attention. In the meantime, two important approaches were revolutionizing the field of optimal control. These were the method of dynamic programming of Bellman and the maximum principle of Pontryagin. Both were advanced around 1956.

In 1959 the Romanian V. M. Popov discovered an exact frequency domain condition for the stability of a class of non-linear systems. This achievement along with later extensions by other investigators brought the spot-light back to the frequency-domain approaches. The circle criterion is an extension of Popov's work. It provides an easy method for the design of a large class of non-linear systems because the concepts and techniques are as simple and essentially the same as those used in the design of linear systems.

2. The Concept of Stability

The concept of stability must be defined prior to our discussing its criterion in a system. For linear systems, a general definition for stability might be given as follows:

Definition 1 A linear system is stable if its output is bounded for any bounded input.

In other words, let $c(t)$ be the output and $r(t)$ the input of a linear system. Then, if

$$|r(t)| \leqslant N < \infty \qquad \text{for } t \geqslant t_0,$$
$$|c(t)| \leqslant M < \infty \qquad \text{for } t \geqslant t_0.$$

Thus, the distinction between stability and instability of a linear system is defined by an abrupt transition as the loop gain is increased. Such an abrupt transition generally is not found in non-linear systems. The transition from stability to instability in non-linear systems usually exhibits a loss of continuity and jump resonance phenomena. This leads to the following two definitions for stability of non-linear systems.

Definition 2 A non-linear system is stable in the bounded sense if any bounded set of inputs produces a bounded set of outputs.

In other words, given a system described by the functional equation $y = H(x)$ the system is stable in the bounded sense if for all x it is such that

$$\|x\| < N, \qquad \|y\| = \|H(x)\| < M$$

Definition 3 A non-linear system is stable in the continuous sense if any bounded set of inputs produces a continuous set of outputs.

The term continuous means that for any functions $x_1(t)$ and $x_2(t)$ belonging to X, there exists $\delta > 0$ for any given $\varepsilon > 0$ such that

$$\|H[x_1]-H[x_2]\| < \varepsilon, \qquad \|x_1-x_2\| < \delta$$

These definitions of stability of non-linear system have been discussed in detail in Reference 1.

3. The Circle Criterion

The circle criterion represents an extension of linear techniques into the analysis of non-linear systems. It yields sufficient conditions for the stability of a large class of non-linear systems. The conditions are dependent upon the sector which bounds the non-linearity and not upon the actual non-linearity.

Consider the basic feedback system shown in Fig. 1 with r, e, u and y real-valued measurable functions of t for $t \geqslant 0$.

It is assumed that $r(t)$ and $y(t)$ are square integrable functions of time. That is, it defines the set of input and output functions in the definitions of stability for non-linear systems as that set of functions which are bounded and approach zero as $t \to \infty$.

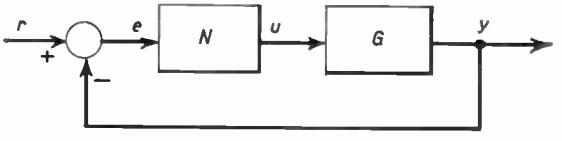


Fig. 1. Basic feedback system.

† Electrical Engineering Department, University of Hong Kong, Pokfulam Road, Hong Kong.

The block labelled N represents a memoryless, either time-invariant or time-varying non-linearity which satisfies the conditions illustrated in Fig. 5 that is:

$$u(t) = N[e(t), t]$$

$$N[0, t] = 0 \qquad \text{for } t \geqslant 0$$

and there exists a positive constant β and a real constant α such that

$$\alpha \leqslant \frac{N[e(t), t]}{e(t)} \leqslant \beta \qquad t \geqslant 0 \text{ for all real } e \neq 0. \quad \ldots\ldots(1)$$

The block labelled G is a linear time-invariant element with impulse response $y(\tau)$ so that

$$y(t) = \int_0^t g(t-\tau)u(\tau)\,d\tau - g_0(t)$$

where g and g_0 are real valued functions such that

$$\int_0^\infty |g(t)|\,dt < \infty \text{ and } \int_0^\infty |g_0(t)|^2\,dt < \infty \quad \ldots\ldots(2)$$

The function g_0 takes into account the initial conditions at $t = 0$.

The circle criterion thus defines sufficient conditions for stability as follows:

Theorem 1 A basic feedback system of Fig. 1 is stable in the bounded sense if it has a non-linearity of the form stated and satisfies the conditions illustrated in Fig. 2 and stated mathematically as:

(i) $\alpha > 0$, the Nyquist plot of $G(j\omega)$ lies outside the circle C_1 of radius $\frac{1}{2}\left[\frac{1}{\alpha} - \frac{1}{\beta}\right]$ centered on the real axis of the complex plane at $\left[-\frac{1}{2}\left(\frac{1}{\alpha} + \frac{1}{\beta}\right), 0\right]$ and does not encircle C_1.

(ii) $\alpha = 0$, the Nyquist plot of $G(j\omega)$ lies to the right of the line given by $\text{Re}\,[G(j\omega)] = -\frac{1}{\beta}$.

(iii) $\alpha > 0$, the Nyquist plot of $G(j\omega)$ is contained within the circle C_2 of radius $\frac{1}{2}\left[\frac{1}{\beta} - \frac{1}{\alpha}\right]$ centered on the real axis of the complex plane at $\left[-\frac{1}{2}\left(\frac{1}{\alpha} + \frac{1}{\beta}\right), 0\right]$

Theorem 2 A basic feedback system of Fig. 1 is stable in the continuous sense if it satisfies Theorem 1 and conditions of the type shown in Fig. 5 for slope of its non-linearity N. Mathematically this is:

(i) $u(t) = N[e(t), t]$

(ii) $N[0, t] = 0, \qquad \text{for } t \geqslant 0.$

(iii) There exist real constants α and β such that

$$\alpha \leqslant \frac{N[e_1(t), t] - N[e_2(t), t]}{e_1(t) - e_2(t)} \leqslant \beta$$

for all real $e \neq 0$ and $t \geqslant 0$.

(iv) $r(t)$ and $y(t)$ are square integrable functions of time; and

(v) one of the conditions given in Fig. 2.

Theorem 2 describes the incremental gain

$$\frac{u_1(t) - u_2(t)}{e_1(t) - e_2(t)} = \frac{N[e_1(t), t] - N_2[e_2(t), t]}{e_1(t) - e_2(t)}$$

being bounded in a sector. The slopes α and β of this sector, however, are not necessarily the same as those obtained by the application of Theorem 1 to the same non-linearity. Generally Theorem 2 will yield larger, or at least equal, critical regions in the complex plane.

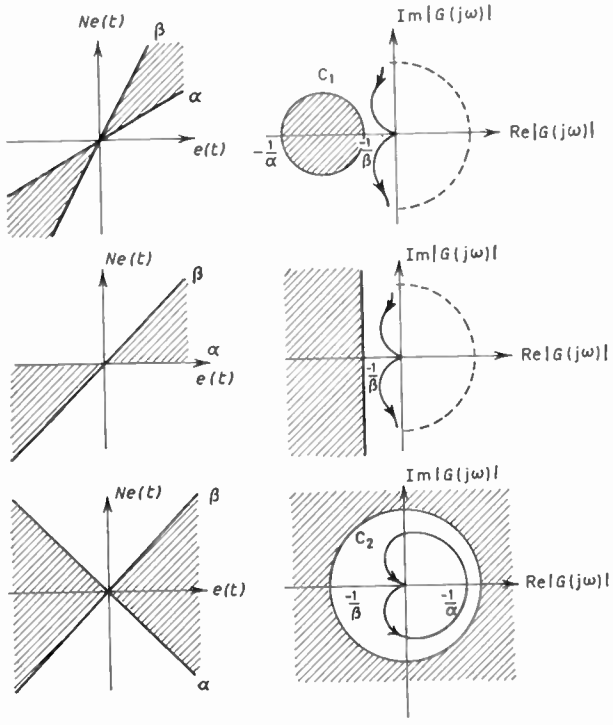


Fig. 2. Sectors of non-linearities and their corresponding critical regions in the $G(j\omega)$ plane.

4. Application of the Circle Criterion

Let us consider the basic feedback system of Fig. 1 with the linear plant $G(s)$ given by

$$G(s) = \frac{K}{\left(1 + \frac{s}{2}\right)\left(1 + \frac{s}{4}\right)\left(1 + \frac{s}{6}\right)}$$

and the non-linearity characteristic shown in Fig. 3.

It is assumed that all inputs meet the basic conditions of the circle criterion. Theorem 1 assures the stability of the system if the Nyquist plot of the linear plant $G(j\omega)$ lies outside the circle C of radius $\frac{5}{4}$ centered on the real axis of the complex plane at $[-\frac{5}{2}, 0]$ and does not encircle C. However, for the design of a feedback system concepts such as gain margin, phase margin etc. are needed. The values of K and ω for which the $G(j\omega)$ becomes a tangent to the critical circle in the Nyquist plot yield the required information. Unfortunately, finding this tangent condition in the circle criterion plot is not a simple matter and it may require computer solution. This paper suggests an alternative graphical approach

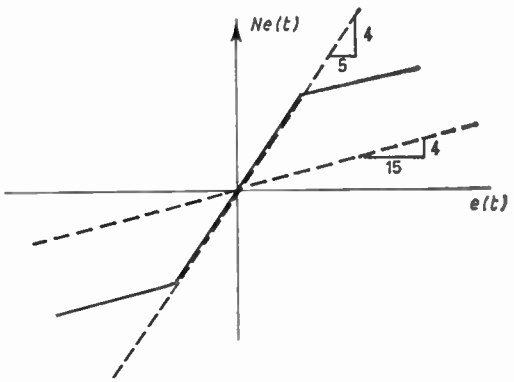


Fig. 3. Non-linearity characteristic.

within the easy reach of an engineer when he applies the circle criterion method of analysis.

5. Pole Shifting

The feedback system of Fig. 4 consists of a linear time-invariant element $KG(s)$, where K is a positive gain constant, and a memoryless time-invariant non-linearity N bounded in a sector α, β, as shown in Fig. 5.

By putting $N_1(x) = N(x) - n_0 x$ .....(3)

where $N_1(x)$ is another non-linearity and

$$n_0 = \frac{(\alpha+\beta)}{2} \qquad .....(4)$$

the sector $$\frac{N(x)}{x} \in [\alpha, \beta] \qquad .....(5)$$

of the original element N is then transformed into the sector

$$\frac{N_1(x)}{x} \in [\alpha_1, \beta_1] \qquad .....(6)$$

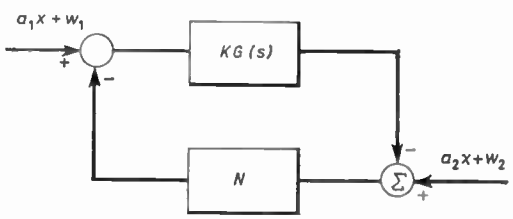


Fig. 4. A feedback system.

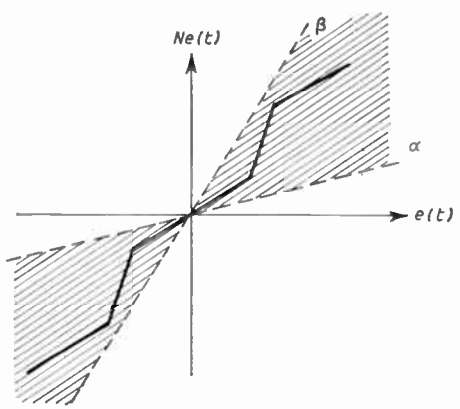


Fig. 5. Non-linearity bounded by shaded region.

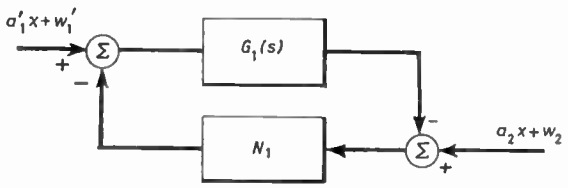


Fig. 6. Equivalent system of Fig. 4.

of an equivalent non-linearity N_1. By the relations of eqns. (3) to (6), it is apparent that

$$\alpha_1 = \frac{(\alpha-\beta)}{2} \qquad .....(7)$$

$$\beta_1 = \frac{(\beta-\alpha)}{2} = -\alpha \qquad .....(8)$$

Hence, a system of Fig. 5 is transformed into an equivalent system of Fig. 6 in which the non-linearity is N_1 and the linear plant

$$G_1(s) = \frac{KG(s)}{1+n_0KG(s)} \qquad .....(9)$$

Further, the input a_1x+w_1 of the original system will be transformed into an input $a_1'x+w_1'$ where

$$a_1' = a_1 - n_0a_2 \text{ and } w_1' = w_1 - n_0w_2$$

Since a_1, a_2, w_1 and w_2 satisfy the conditions of the circle criterion and n_0 is a positive constant a_1' and w_1' also satisfy these conditions. The transformation maps the Nyquist plot into the inside of a circle in the $G_1(j\omega)$ plane centered at the origin and of radius $\frac{2}{\beta-\alpha}$. This is illustrated in Fig. 7. Transformation eqn. (8) therefore maps constant $|G_1(j\omega)|$ circles into appropriate circles in the original $KG(j\omega)$ plane. Such a mapping is analogous to that used in the relation between open- and closed-loop transfer functions of linear time-invariant systems and can be expressed in the logarithmic gain vs. phase plane. Specifically, this mapping can be represented in the form of a Nichols chart.

6. Worked Example

To demonstrate the use of a Nichols chart for determining the relative stability for the circle criterion, let us consider again the example cited above. The original system and its equivalent are shown in Figs. 8(a) and (b) respectively.

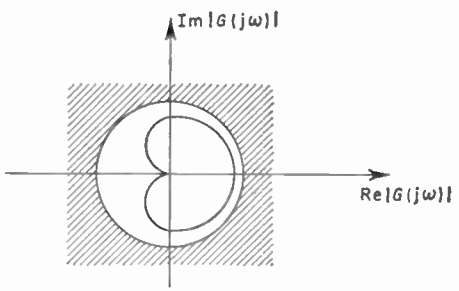


Fig. 7. The transformation map $G_1(j\omega)$ into inside of the admissible region.

The linear plant $G(s) = \dfrac{K}{\left(1+\frac{s}{2}\right)\left(1+\frac{s}{4}\right)\left(1+\frac{s}{6}\right)}$

and its non-linearity N has the characteristic described in Fig. 3. It is bounded in a sector with $\alpha = \frac{4}{15}$ and $\beta = \frac{4}{5}$. By means of eqns. (4) to (9), the equivalent system has

$$G_1(s) = \frac{KG(s)}{1+n_0KG(s)} = \frac{K}{\left(1+\frac{s}{2}\right)\left(1+\frac{s}{4}\right)\left(1+\frac{s}{6}\right)+n_0K}$$

N_1 is bounded in a sector having $\alpha_1 = -\frac{4}{15}$ and $\beta_1 = \frac{4}{15}$.

We can write $n_0G_1(s) = \dfrac{n_0KG(s)}{1+n_0KG(s)}$

and plot $n_0KG(j\omega)$ on the Nichols chart as a function of ω with $n_0K = 1$. This curve is then moved vertically upwards until it makes a tangent to the critical M circle which is given by

$$M_p = |n_0G_1(j\omega)| = \left|\frac{n_0KG(j\omega)}{1+n_0KGj(\omega)}\right|$$

$$= 20\log_{10}\frac{n_0}{\beta_1} = 6 \text{ dB in this example.}$$

The plot is shown in Fig. 9. The change in decibels when the curve is shifted along the vertical axis represents the value of n_0K in decibels and the tangential point yields the frequency. We interpret these values from the plot of Fig. 9, namely $n_0K = 12{\cdot}8$ dB

or $K = 8{\cdot}2$ and $\omega_c = 4{\cdot}6$ rad/s

Application of the circle criterion without transformation to this example, the values of K and ω_c for which the Nyquist plot $G(j\omega)$ becomes tangential to the critical region are found with the aid of a digital computer to be

$$K = 7{\cdot}94 \text{ and } \omega_c = 4{\cdot}72.$$

These are in close agreement with the above method.

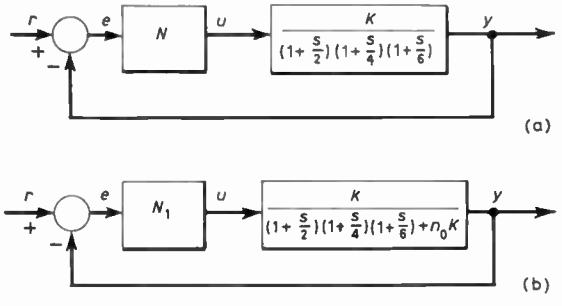


Fig. 8. (a) The non-linear feedback system.
(b) Its equivalent.

7. Conclusion

The circle criterion provides a useful method of testing stability in non-linear feedback systems. However, it has a minor disadvantage in that computer computation may be required to evaluate the relative stability conditions for the Nyquist plot. Where computer facilities are not available and repeated trial-and-error methods are not desired, the use of a logarithmic gain-phase plot as discussed in this paper represents a very useful extension of the circle criterion method of analysis.

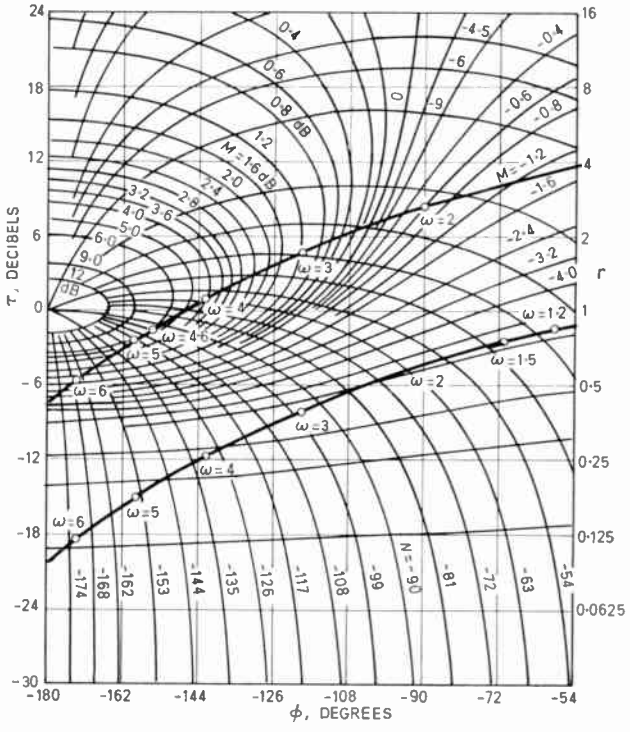


Fig. 9. The Nichols chart.

8. References

1. Zames, G., 'On the input-output stability of time-varying non-linear feedback systems—Part I: Conditions derived using concepts of loop gain, conicity and positivity', *I.E.E.E.E. Trans. on Automatic Control*, **AC–11**, pp. 228–39, April 1966.
2. Zames, G., *ibid.*, Part II: Conditions involving circles in the frequency plane and sector non-linearities', *loc. cit* pp. 465–76, July 1966.
3. Sandberg, I. W., 'A frequency-domain condition for the stability of feedback systems containing a single time-varying non-linear element', *Bell Syst. Tech. J.*, **43**, pp. 1601–8, July 1964.
4. Murphy, G. J., 'A frequency-domain stability chart for non-linear feedback systems', *I.E.E.E. Trans. on Automatic Control*, **AC–12**, pp. 740–3, December 1967.
5. Moore, J. B., 'A circle criterion generalization for relative stability', *I.E.E.E. Trans. on Automatic Control*, **AC–13**, pp. 127–8, February 1968.
6. Narendra, K. S., 'A geometrical criterion for the stability of certain non-linear non-autonomous systems', *I.E.E.E. Trans. on Circuit Theory*, **CT–17**, pp. 406–8, September 1964.
7. Gibson, J. E., 'Non-linear Automatic Control', (McGraw-Hill, New York, 1963).

Manuscript first received by the Institution on 9th November 1970 and in final form on 10th March 1971. (Paper No. 1387/IC 45.)

U.D.C. 621.396.677.3

Self and Mutual Impedances of Two Parallel Staggered Dipoles by Variational Method

By

J. V. SURUTKA, D.Sc.†

The problem of the self- and mutual-impedances of two arbitrarily located parallel dipoles is solved by using the variational method and a two-term trial function for current distribution. The impedances are calculated for half-wave and full-wave dipoles in non-staggered, echelon and collinear arrangements. In all three cases the results are in excellent agreement with those obtained by the Chang-King five-term theory. In the non-staggered case agreement with experimental results is also very satisfactory.

1. Introduction

The problem of two arbitrarily-located parallel dipoles has been treated by Chang and King[1] using an integral equation technique. They solved the integral equations by an approximate method first made available by King and Wu in 1965[2] in dealing with a single dipole antenna. Recently, Popović[8] analysed the problem of two arbitrarily-located identical parallel asymmetrical antennas. The purpose of the present paper is to give a variational solution to the same problem, using a two-term trial function for current distribution. Only symmetrical parallel staggered dipoles will be considered, but the method can be generalized to asymmetrical staggered dipoles, combining the procedure described in reference 8 with the method of analysing asymmetrical dipoles described in reference 9.

The variational method for determining the impedance of a thin cylindrical antenna was presented for the first time by Storer.[3, 4] Some years later Levis and Tai[5] proposed an extension of the application of this method to the problem of two coupled parallel linear antennas of unequal sizes. However, their approach was quite general and no definite final formulas were derived. In a recent paper Surutka and Popović[6] further developed the Levis and Tai proposal and derived the explicit variational formulas for the self- and mutual-impedances of two parallel, non-staggered dipoles of unequal size. The two-term trial functions for currents were those utilized by Storer.[3]

In this paper a similar technique was used in order to obtain the impedances of two identical parallel dipoles, the centres of which are located arbitrarily. The non-staggered and collinear arrays of two elements are special cases of the general one presented here. For the non-staggered case the theoretical results agree very well with experimental data measured by Mack.[7] Unfortunately, for other arrangements no experimental data are available. On the other hand, the theoretical results, obtained by variational method, favourably agree with those obtained by Chang and King,[1] and Popović.[8]

2. Integral Equations and Variational Expressions for Impedances

Let us consider two identical parallel dipoles of half length h and radius a, arranged as in Fig. 1. The distance between the axes of the dipoles is b, and the centres of the dipoles are displaced by a distance d parallel to the axes. The non-staggered and collinear arrangements are special cases of the one treated here, putting $d = 0$ and $b = a$, respectively.

As the whole system is point symmetric about the point M, bisecting the distance between the centres O_1 and O_2 of the two dipoles, the axes of the dipoles are oriented oppositely, so that the positive direction of the current in the antenna 1 is upward and downward in the antenna 2.

As usual, it is assumed that the two dipoles are driven by the slice or delta function generators having voltages V_1 and V_2, which are due to impressed fields $E^i_{z1} = V_1\delta(z_1)$ and $E^i_{z2} = V_2\delta(z_2)$ located in the infinitely narrow belts at the centres of the dipoles. $\delta(z)$ is Dirac delta function defined at $z = 0$.

Assuming an infinite conductivity of the antenna conductors the basic equation expressing the satisfied boundary conditions on the surface of the dipole 1 can

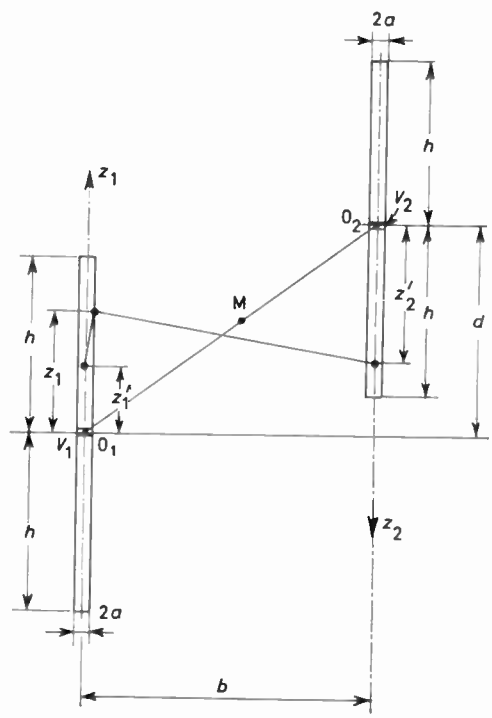


Fig. 1. Two parallel staggered dipoles.

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

be put in the form

$$V_1\,\delta(z_1) = \mathrm{j}\frac{\omega}{k^2}\left(k^2+\frac{\partial^2}{\partial z_1^2}\right)A_{z_1}(z_1) \qquad \ldots\ldots(1)$$

where

$$A_{z_1} = \frac{\mu_0}{4\pi}\left\{\int_{-h}^{h} I_1(z_1')\frac{\exp(-\mathrm{j}kr_{11})}{r_{11}}\,\mathrm{d}z_1' - \int_{-h}^{h} I_2(z_2')\frac{\exp(-\mathrm{j}kr_{12})}{r_{12}}\,\mathrm{d}z_2'\right\} \qquad \ldots\ldots(2)$$

is the vector-potential in the point z_1 on the surface of the dipole 1 (since both currents $I_1(z_1)$ and $I_2(z_2)$ are in the z direction, the vector-potential has the z-component only),

$$r_{11} = [a^2+(z_1-z_1')^2]^{1/2} \qquad \ldots\ldots(3)$$

$$r_{12} = [b^2+(z_1+z_2'-d)^2]^{1/2} \qquad \ldots\ldots(4)$$

$$k = 2\pi/\lambda = \omega(\varepsilon_0\mu_0)^{1/2}. \qquad \ldots\ldots(5)$$

In evaluating the vector-potential it is assumed that the whole current is concentrated in the axes of the dipoles.

A similar equation for the dipole 2 can be obtained from equations (1)–(4) by interchanging the subscripts 1 and 2.

Regarding the two identical antennas as the mutually coupled circuits, we can write the following relationships:

$$V_1 = Z_s I_1(0) - Z_m I_2(0) \qquad \ldots\ldots(6)$$

$$V_2 = -Z_m I_1(0) + Z_s I_2(0), \qquad \ldots\ldots(7)$$

where Z_s and Z_m are self- and mutual-impedances of two antennas, respectively. (The self-impedance of dipole 1 is its input impedance when dipole 2 is disconnected from the transmission line and vice versa.)

In order to obtain the fundamental variational expressions for the impedances, we first multiply equation (1) by $I_1(z_1)\mathrm{d}z_1$ and then integrate from $-h$ to $+h$. Taking into account that

$$\int_{-h}^{h} V_1 I_1(z_1)\delta(z_1)\mathrm{d}z_1 = V_1 I_1(0), \qquad \ldots\ldots(8)$$

we can write

$$Z_s I_1^2(0) - Z_m I_1(0) I_2(0) = \frac{\mathrm{j}\eta}{4\pi}\int_{-h}^{h}\int_{-h}^{h} I_1(z_1')I_1(z_1)K_{11}(z_1, z_1')\mathrm{d}z_1'\,\mathrm{d}z_1 - \frac{\mathrm{j}\eta}{4\pi}\int_{-h}^{h}\int_{-h}^{h} I_2(z_2')I_1(z_1)K_{12}(z_1, z_2')\mathrm{d}z_2'\,\mathrm{d}z_1, \qquad \ldots\ldots(9)$$

where $\eta = (\mu_0/\varepsilon_0)^{1/2} = 120\pi$ ohms, $K_{11}(z_1, z_1')$ and $K_{12}(z_1, z_2')$ are the kernels defined as follows:

$$K_{11}(z_1, z_1') = k\left(1+\frac{1}{k^2}\frac{\partial^2}{\partial z_1^2}\right)\frac{\exp(-\mathrm{j}kr_{11})}{r_{11}}, \qquad \ldots\ldots(10)$$

$$K_{12}(z_1, z_2') = k\left(1+\frac{1}{k^2}\frac{\partial^2}{\partial z_1^2}\right)\frac{\exp(-\mathrm{j}kr_{12})}{r_{12}}. \qquad \ldots\ldots(11)$$

With subscripts 1 and 2 interchanged, we can write a similar equation for the second antenna, thus obtaining a pair of equations from which the two impedances can be determined.

The calculation can readily be simplified by decoupling the two basic equations using a sequence method.[4] Because of the point symmetry of the system, it is easy to see that in the case of symmetric excitation conditions (sequence I), corresponding to the driving voltages $V_1 = V_2 = V_s$, we have $I_1(z_1) = I_2(z_2) = I_s(z)$. In the antisymmetric case (sequence II) $V_1 = -V_2 = V_a$ and then $I_1(z_1) = -I_2(z_2) = I_a(z)$. It is convenient to introduce normalized current distribution functions (for symmetric and antisymmetric case), defined as follows:

$$g_s(z) = I_s(z)/I_s(0) \qquad g_a(z) = I_a(z)/I_a(0). \qquad \ldots\ldots(12)$$

In accordance with the sequence method equation (9) for the antenna 1, as well as the corresponding equation for the antenna 2 (which was not written), give two sequence equations:

$$Z_I = Z_s - Z_m = \frac{\mathrm{j}\eta}{4\pi}\int_{-h}^{h}\int_{-h}^{h} g_s(z')g_s(z)K_{11}(z, z')\mathrm{d}z'\mathrm{d}z - \frac{\mathrm{j}\eta}{4\pi}\int_{-h}^{h}\int_{-h}^{h} g_s(z')g_s(z)K_{12}(z, z')\mathrm{d}z'\mathrm{d}z, \qquad \ldots\ldots(13)$$

$$Z_{II} = Z_s + Z_m = \frac{\mathrm{j}\eta}{4\pi}\int_{-h}^{h}\int_{-h}^{h} g_a(z')g_a(z)K_{11}(z, z')\mathrm{d}z'\mathrm{d}z + \frac{\mathrm{j}\eta}{4\pi}\int_{-h}^{h}\int_{-h}^{h} g_a(z')g_a(z)K_{12}(z, z')\mathrm{d}z'\mathrm{d}z, \qquad \ldots\ldots(14)$$

where

$$K_{11}(z, z') = k\left(1+\frac{1}{k^2}\frac{\partial^2}{\partial z^2}\right)\frac{\exp(-\mathrm{j}kr_{11})}{r_{11}} \qquad \ldots\ldots(15)$$

$$K_{12}(z, z') = k\left(1+\frac{1}{k^2}\frac{\partial^2}{\partial z^2}\right)\frac{\exp(-k_{12}),}{r_{12}} \qquad \ldots\ldots(16)$$

and

$$r_{11} = [a^2+(z-z')^2]^{1/2} \qquad \ldots\ldots(17)$$

$$r_{12} = [b^2+(z+z'-d)^2]^{1/2}. \qquad \ldots\ldots(18)$$

Z_I and Z_{II} are driving input impedances of two dipoles corresponding to the symmetric and antisymmetric excitation conditions.

Since the two kernels, equations (15) and (16), remain unchanged when z and z' are substituted one for the other, it can be easily shown that the expressions (13) and (14) have the stationary property with respect to small changes in the relative distribution functions. It means that

$$\delta Z_I = 0 \quad \text{and} \quad \delta Z_{II} = 0. \qquad \ldots\ldots(19)$$

As shown by Storer[3, 4] in the case of a single symmetrical antenna, a two-term trial function of the type

$$I(z) = A\sin k(h-|z|)+B[1-\cos k(h-|z|)] \qquad \ldots\ldots(20)$$

leads to very satisfactory results in evaluating the impedance by the variational method. The same type of trial function for currents will be used in this paper, taking into account the unsymmetry in current distribution

functions. So, the normalized distribution functions for the symmetric and antisymmetric cases can be written as follows:

$$g_s(z) = \frac{I_s(z)}{I_s(0)} = \begin{cases} b_1 f_1(z) + b_2 f_2(z) & z \in [0, h] \\ b_3 f_3(z) + b_4 f_4(z) & z \in [-h, 0] \end{cases} \quad \ldots\ldots(21)$$

$$g_a(z) = \frac{I_a(z)}{I_a(0)} = \begin{cases} a_1 f_1(z) + a_2 f_2(z) & z \in [0, h] \\ a_3 f_3(z) + a_4 f_4(z) & z \in [-h, 0] \end{cases} \quad \ldots\ldots(22)$$

where

$$f_1(z) = \sin k(h-z) \qquad f_2(z) = 1 - \cos k(h-z)$$
$$f_3(z) = \sin k(h+z) \qquad f_4(z) = 1 - \cos k(h+z) \quad \ldots\ldots(23)$$

and

$$g_s(0) = 1$$
and $$g_a(0) = 1. \quad \ldots\ldots(24)$$

The continuity condition for the current at the feeding point, $g(0_+) = g(0_-)$, leads to

$$b_2 = [1 - b_1 f_1(0)]/f_2(0) \qquad b_4 = [1 - b_3 f_1(0)]/f_2(0)$$
$$a_2 = [1 - a_1 f_1(0)]/f_2(0) \qquad a_4 = [1 - a_3 f_1(0)]/f_2(0). \quad \ldots\ldots(25)$$

Introducing (21) and (22) into equations (13) and (14), the impedances Z_I and Z_{II} can be put in the following forms

$$Z_I = Z_s - Z_m = A_3 b_1^2 + A_4 b_3^2 + 2A_7 b_1 + 2A_8 b_3 + 2A_{10} b_1 b_3 + A_{11} - A_{12} \quad \ldots\ldots(26)$$

$$Z_{II} = Z_s + Z_m = A_1 a_1^2 + A_2 a_3^2 + 2A_5 a_1 + 2A_6 a_3 + 2A_9 a_1 a_3 + A_{11} + A_{12}, \quad \ldots\ldots(27)$$

where

$$A_1 = w_{11} + v_{11} - 2(w_{12} + v_{12}) f_1(0)/f_2(0) + (w_{22} + v_{22}) f_1^2(0)/f_2^2(0)$$
$$A_2 = w_{11} + v_{33} - 2(w_{12} + v_{34}) f_1(0)/f_2(0) + (w_{22} + v_{44}) f_1^2(0)/f_2^2(0)$$
$$A_3 = w_{11} - v_{11} - 2(w_{12} - v_{12}) f_1(0)/f_2(0) + (w_{22} - v_{22}) f_1^2(0)/f_2^2(0)$$
$$A_4 = w_{11} - v_{33} - 2(w_{12} - v_{34}) f_1(0)/f_2(0) + (w_{22} - v_{44}) f_1^2(0)/f_2^2(0)$$
$$A_5 = (w_{12} + w_{14} + v_{12} + v_{14})/f_2(0) - (w_{22} + w_{24} + v_{22} + v_{24}) f_1(0)/f_2^2(0)$$
$$A_6 = (w_{12} + w_{14} + v_{34} + v_{23})/f_2(0) - (w_{22} + w_{24} + v_{44} + v_{24}) f_1(0)/f_2^2(0)$$
$$A_7 = (w_{12} + w_{14} - v_{12} - v_{14})/f_2(0) - (w_{22} + w_{24} - v_{22} - v_{24}) f_1(0)/f_2^2(0)$$
$$A_8 = (w_{12} + w_{14} - v_{34} - v_{23})/f_2(0) - (w_{22} + w_{24} - v_{44} - v_{24}) f_1(0)/f_2^2(0)$$
$$A_9 = w_{13} + v_{13} - (2w_{14} + v_{14} + v_{23}) f_1(0)/f_2(0) + (w_{24} + v_{24}) f_1^2(0)/f_2^2(0)$$
$$A_{10} = w_{13} - v_{13} - (2w_{14} - v_{14} - v_{23}) f_1(0)/f_2(0) + (w_{24} - v_{24}) f_1^2(0)/f_2^2(0)$$
$$A_{11} = (w_{22} + w_{24}) 2/f_2^2(0)$$
$$A_{12} = (v_{22} + v_{44} + 2v_{24})/f_2^2(0). \quad \ldots\ldots(28)$$

The integrals w_{ik} and v_{ik} are defined as follows:

$$w_{ik} = (-1)^{m+n} \frac{j\eta}{4\pi} \int_0^{h_n} \left[\int_0^{h_m} f_i(z) K_{11}(z, z') dz \right] f_k(z') dz' \quad \ldots\ldots(29)$$

and

$$v_{ik} = (-1)^{m+n} \frac{j\eta}{4\pi} \int_0^{h_n} \left[\int_0^{h_m} f_i(z) K_{12}(z, z') dz \right] f_k(z') dz' \quad \ldots\ldots(30)$$

where

$$\begin{array}{lll} \text{for} & i = 1, 2 & m = \begin{cases} 1 \\ 2 \end{cases} \qquad h_m = \begin{cases} h \\ -h \end{cases} \\ \text{for} & i = 3, 4 & \\ \text{for} & k = 1, 2 & n = \begin{cases} 1 \\ 2 \end{cases} \qquad h_n = \begin{cases} h \\ -h. \end{cases} \\ \text{for} & k = 3, 4 & \end{array}$$

There are six different w_{ik} integrals: w_{11}, w_{21}, w_{13}, w_{41}, w_{22} and w_{24}. The remaining ten integrals can be expressed by the former ones: $w_{33} = w_{11}$, $w_{12} = w_{34} = w_{43} = w_{21}$, $w_{31} = w_{13}$, $w_{23} = w_{32} = w_{14} = w_{41}$, $w_{44} = w_{22}$, $w_{42} = w_{24}$. Ten v_{ik} integrals have different values: v_{11}, $v_{12} = v_{21}$, $v_{13} = v_{31}$, $v_{14} = v_{41}$, v_{22}, $v_{32} = v_{23}$, $v_{24} = v_{42}$, v_{33}, $v_{34} = v_{43}$ and v_{44}.

Owing to the stationary property of equations (13) and (14), the current coefficients a_1, a_3, b_1 and b_3 can be determined by requiring that

$$\frac{\partial Z_{I,II}}{\partial a_i} = 0 \qquad \frac{\partial Z_{I,II}}{\partial b_i} = 0 \qquad i = 1, 3. \quad \ldots\ldots(31)$$

From equation (31) we have

$$a_1 = \frac{A_6 A_9 - A_2 A_5}{A_1 A_2 - A_9^2} \qquad a_3 = \frac{A_5 A_9 - A_1 A_6}{A_1 A_2 - A_9^2}$$
$$b_1 = \frac{A_8 A_{10} - A_4 A_7}{A_3 A_4 - A_{10}^2} \qquad b_3 = \frac{A_7 A_{10} - A_3 A_8}{A_3 A_4 - A_{10}^2}. \quad \ldots\ldots(32)$$

3. Evaluation of Integrals w_{ik} and v_{ik}

The computing time for evaluation of the integrals w_{ik} and v_{ik} can be considerably reduced if the double integrals are transformed into single ones.

Using the method of partial integration and changing the variables it can be shown that the following identity is valid:[3]

$$\int_0^{h_m} f_i(z) K(z, z') dz = k \int_0^{h_m} f_i(z) \left(1 + \frac{1}{k^2} \frac{\partial^2}{\partial z^2}\right) \frac{\exp(-jkr)}{r} dz$$
$$= \int_0^{h_m} \frac{\exp(-jkr)}{r} \left(1 + \frac{1}{k^2} \frac{\partial^2}{\partial z^2}\right) \times f_i(z) dz \pm \frac{f_i(0)}{k} \times \frac{\partial}{\partial z'} \frac{\exp(-jkr_0)}{r_0} - \frac{f_i'(h_m)}{k} \times \frac{\exp(-jkr_{h_m})}{r_{h_m}} + \frac{f_i'(0)}{k} \frac{\exp(-jkr_0)}{r_0}, \quad \ldots\ldots(33)$$

where, in the case $K(z, z') = K_{11}(z, z')$

$$r = r_{11}, \qquad r_0 = r_{11}|_{z=0} \qquad \text{and} \qquad r_{h_m} = r_{11}|_{z=h_m} \qquad \text{.....(34)}$$

and, in the case $K(z, z') = K_{12}(z, z')$

$$r = r_{12}, \qquad r_0 = r_{12}|_{z=0} \qquad \text{and} \qquad r_{h_m} = r_{12}|_{z=h_m}. \qquad \text{.....(35)}$$

The minus sign in the second term on the right side of equation (33) refers to the latter case.

Using the identity (33) the integral (29) can be transformed as follows:

$$\begin{aligned}(-1)^{m+n}\frac{4\pi}{j\eta}w_{ik} = \int_0^{h_n}\Bigg[k\int_0^{h_m}\frac{\exp(-jkr)}{r}\times\\ \times\left(1+\frac{1}{k^2}\frac{\partial^2}{\partial z^2}\right)f_i(z)dz\Bigg]\times\\ \times f_k(z')dz' - f_k(0)f_i(0)\times\\ \times\frac{\exp(-jk)\sqrt{a^2+d^2}}{k\sqrt{a^2+d^2}} -\\ -f_t(0)\int_0^{h_n}\exp\frac{(-jkr_0)}{kr_0}\times\\ \times f'_k(z')dz' - f'_i(h_m)\times\\ \int_0^{h_n}\frac{\exp(-jkr_{h_m})}{kr_{h_m}}f_k(z')dz' + f'_t(0)\times\\ \times\int_0^{h_n}\frac{\exp(-jkr_0)}{kr_0}f_k(z')dz', \qquad \text{.....(36)}\end{aligned}$$

where r, r_0 and r_{h_m} are defined by (34).

A similar expression can be written for v_{ik} by changing the algebraic sign of the second and third term on the right-hand side of equation (36) and substituting b for a. Equation (35) defines r, r_0 and r_{h_m}.

In the case $i = 1, 3$, $f_i(z)$ are the sine-functions defined by (23), and hence

$$\left(1+\frac{1}{k^2}\frac{\partial^2}{\partial z^2}\right)f_i(z) = 0.$$

Therefore, the first term on the right-hand side of equation (36) disappears and the evaluation of the w_{ik} and v_{ik} reduces to the evaluation of the single integrals. There are only two w_{ik} integrals (w_{22} and w_{24}) and three v_{ik} integrals (v_{22}, v_{24} and v_{44}) wherein the first term on the right-hand side does not disappear ($i = 2, 4$). But, also, in all of them the first term can be transformed into the single integrals by the method of partial integration and by appropriate changes of variables.

4. Numerical Results and Conclusion

The variational method presented here has been used to calculate the self- and mutual-impedances of two dipoles for the following three different arrangements:

(1) non-staggered parallel dipoles,

(2) dipoles in echelon, i.e. $b = d$, and

(3) collinear dipoles.

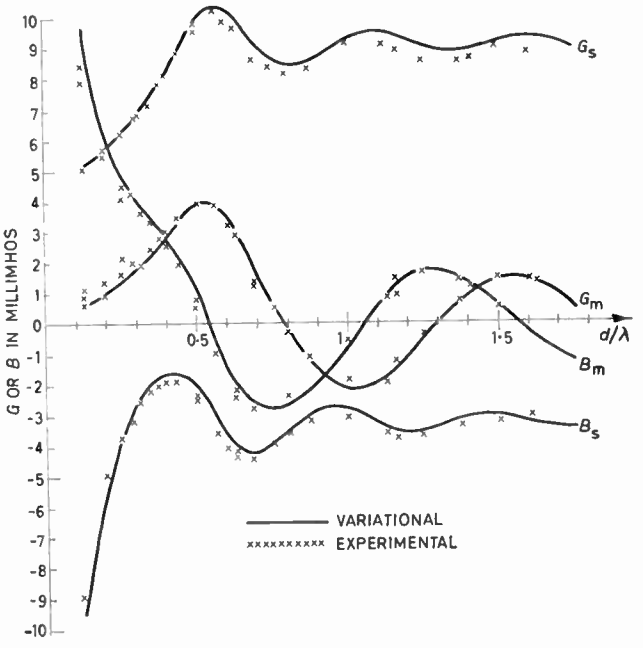


Fig. 2. Theoretical and experimental (Mack) self- and mutual-admittances of two parallel non-staggered half-wave dipoles, $a/\lambda = 0{\cdot}007022$

Both half-wave and full-wave dipoles have been treated and in both cases the radii of $0{\cdot}007022\,\lambda$ have been assumed.

The aforementioned arrangements and dimensions have been chosen from Chang and King's work[1] so as to enable a direct comparison of the two methods. In addition, Popović[8] analysed the same cases using the polynomial approximation for current.

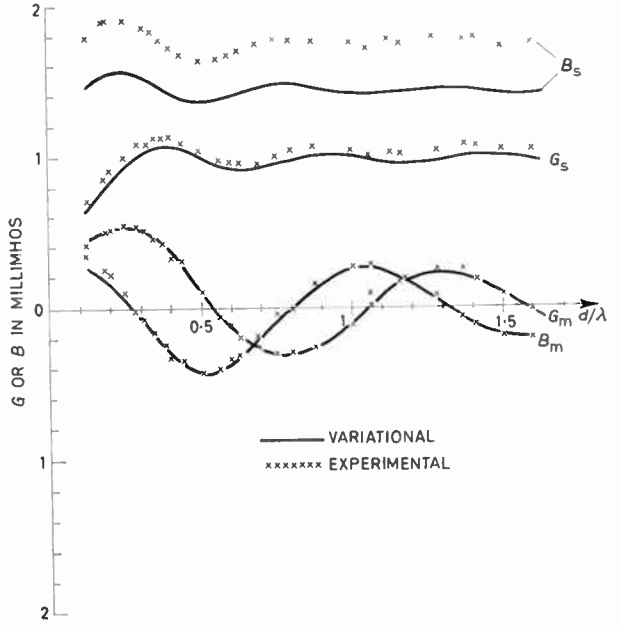


Fig. 3. Theoretical and experimental (Mack) self- and mutual-admittances of two parallel non-staggered full-wave dipoles, $a/\lambda = 0{\cdot}007022$.

Table 1

Self- and mutual-impedances (in ohms) of non-staggered ($d = 0$) array of two elements, $a/\lambda = 0{\cdot}007022$

$h = 0{\cdot}25\,\lambda$

	Variational		Chang and King	
b/λ	$Z_s = R_s + jX_s$	$Z_m = R_m + jX_m$	$Z_s = R_s + jX_s$	$Z_m = R_m + jX_m$
0·050	104·0+j 22·2	102·4+j 4·2	99·3+j 25·6	97·7+j 7·2
0·100	96·7+j 21·4	89·3−j 18·2	93·1+j 24·7	85·9−j 15·4
0·150	91·4+j 25·8	73·7−j 34·1	88·4+j 28·6	71·3−j 31·5
0·250	90·3+j 35·8	39·4−j 53·4	87·4+j 37·9	38·6−j 50·9
0·375	97·6+j 39·9	−2·5−j 53·0	94·0+j 41·7	−1·8−j 51·1
0·500	101·2+j 35·7	−29·8−j 31·4	97·4+j 37·9	−28·4−j 30·7
0·625	98·6+j 32·7	−36·1−j 3·2	95·2+j 35·1	−34·8−j 3·6
0·750	96·2+j 34·2	−24·9+j 18·8	92·9+j 36·5	−24·4+j 17·9
0·875	97·2+j 36·2	−4·7+j 27·1	93·8+j 38·3	−4·9+j 26·1
1·000	98·8+j 35·6	13·5+j 20·3	95·3+j 37·7	12·8+j 19·8
1·125	98·4+j 34·2	21·3+j 4·8	94·9+j 36·5	20·5+j 4·9
1·250	97·3+j 34·5	16·9−j 10·3	93·9+j 36·8	16·5−j 9·7
1·375	97·5+j 35·5	4·6−j 17·6	94·0+j 37·6	4·7−j 16·9
1·500	98·3+j 35·3	−8·3−j 14·5	94·8+j 37·5	−7·8−j 14·2
1·750	97·6+j 34·7	−12·7+j 6·8	94·1+j 36·9	−12·4+j 6·4
2·000	98·1+j 35·2	5·8+j 11·2	94·6+j 37·4	5·5+j 11·0
2·250	97·7+j 34·8	10·1−j 5·1	94·3+j 37·0	9·8−j 4·8
2·500	98·0+j 35·1	−4·5−j 9·1	94·5+j 37·3	−4·2−j 8·9
2·750	97·8+j 34·9	−8·4+j 4·0	94·3+j 37·1	−8·2+j 3·8
3·000	98·0+j 35·1	3·6+j 7·7	94·5+j 37·3	3·4+j 7·5
		$h = 0{\cdot}50\,\lambda$		
0·050	200·4−j 786·7	121·1+j 254·6	195·5−j 795·7	115·7+j 261·0
0·100	195·2−j 648·5	81·6+j 216·0	190·9−j 655·5	76·9+j 221·0
0·150	204·3−j 565·3	69·7+j 178·7	200·5−j 571·5	65·7+j 182·0
0·250	237·0−j 471·2	73·4+j 132·2	233·9−j 476·4	71·4+j 133·4
0·375	302·6−j 417·3	109·2+j 91·0	299·8−j 423·0	108·5+j 90·9
0·500	380·2−j 450·7	155·6−j 1·0	376·4−j 456·9	154·4−j 1·5
0·625	342·3−j 522·6	79·7−j 113·0	338·1−j 527·6	78·4−j 111·9
0·750	297·0−j 490·5	−16·3−j 105·1	293·5−j 495·6	−16·6−j 104·4
0·875	313·7−j 456·2	−68·6−j 65·3	310·3−j 461·6	−68·0−j 64·6
1·000	346·8−j 464·8	−95·2−j 2·4	343·0−j 470·3	−94·2−j 2·3
1·125	338·5−j 496·6	−61·8+j 66·6	334·7−j 501·6	−61·4+j 65·9
1·250	312·6−j 488·3	3·5+j 79·0	309·5−j 493·4	3·2+j 78·4
1·375	317·6−j 467·3	49·6+j 52·6	314·1−j 472·9	48·7+j 52·1
1·500	336·8−j 469·3	71·0+j 5·2	333·0−j 474·8	70·1+j 5·4
1·750	319·2−j 486·0	2·5−j 62·8	316·0−j 491·0	3·0−j 61·8
2·000	332·5−j 472·1	−56·6−j 6·8	328·6−j 477·5	−55·6−j 6·9
2·250	322·4−j 484·1	−5·3+j 51·5	318·9−j 489·3	−5·5+j 50·5
2·500	330·3−j 473·8	46·8+j 7·4	326·8−j 479·4	45·6+j 7·7
2·750	324·1−j 482·8	6·4−j 43·3	320·9−j 487·9	6·7−j 42·4
3·000	329·1−j 475·0	−39·8−j 7·6	325·5−j 480·7	−39·0−j 7·2

Table 2

Self- and mutual-impedances (in ohms) of two parallel dipoles in echelon, $b = d$, $a/\lambda = 0{\cdot}007022$

$h = 0{\cdot}25\,\lambda$

	Variational		Chang and King	
$b/\lambda = d/\lambda$	$Z_s = R_s + jX_s$	$Z_m = R_m + jX_m$	$Z_s = R_s + jX_s$	$Z_m = R_m + jX_m$
0·050	108·6+j 20·1	106·5+j 8·2	104·7+j 23·5	102·7+j 10·7
0·100	98·9+j 20·3	89·2−j 10·1	95·6+j 23·6	86·3−j 7·9
0·150	92·9+j 26·2	69·4−j 23·3	90·1+j 28·9	67·4−j 21·5
0·250	93·2+j 35·5	29·8−j 37·7	90·1+j 37·5	29·4−j 36·1
0·375	98·3+j 36·6	−7·9−j 28·0	94·8+j 38·7	−7·1−j 27·2
0·500	98·4+j 34·5	−18·9−j 4·4	94·9+j 36·8	−18·1−j 4·6
0·625	97·5+j 34·9	−9·0+j 11·0	94·1+j 37·1	−8·9+j 10·4
0·750	98·0+j 35·2	4·8+j 10·1	94·5+j 37·3	4·5+j 9·8
0·875	97·9+j 34·9	9·2−j 0·1	94·5+j 37·0	8·9+j 0·1
1·000	97·8+j 35·0	3·3−j 7·1	94·3+j 37·2	3·3−j 6·8
1·125	97·9+j 35·0	−4·4−j 5·2	94·5+j 37·2	−4·1−j 5·1
1·250	97·9+j 34·9	−5·9+j 1·5	94·4+j 37·1	−5·7+j 1·4
1·375	97·8+j 35·0	−1·0+j 5·4	94·4+j 37·2	−1·1+j 5·1
1·500	97·9+j 35·0	4·0+j 3·0	94·4+j 37·2	3·8+j 2·9
1·750	97·9+j 35·0	−0·1−j 4·2	94·4+j 37·2	−0·1−j 4·0
2·000	97·8+j 35·0	−2·8+j 2·3	94·4+j 37·2	−2·7+j 2·2
2·250	97·9+j 35·0	3·1+j 0·7	94·4+j 37·2	3·0+j 0·7
2·500	97·9+j 35·0	−1·8−j 2·7	94·4+j 37·2	−1·2−j 2·5
2·750	97·9+j 35·0	−1·1+j 2·3	94·4+j 37·2	−1·1+j 2·2
3·000	97·9+j 35·0	2·3−j 0·5	94·4+j 37·2	2·2−j 0·4
		$h = 0{\cdot}50\,\lambda$		
0·050	529·2−j 1325·5	−44·6+j 761·6	284·7−j 1078·6	56·7+j 533·9
0·100	324·9−j 875·7	−5·9+j 445·1	250·2−j 824·9	7·2+j 399·1
0·150	265·4−j 680·0	9·1+j 299·3	239·4−j 672·7	9·8+j 288·7
0·250	252·6−j 508·9	43·6+j 156·4	246·5−j 511·3	44·4+j 152·2
0·375	303·2−j 438·2	85·2+j 70·0	300·3−j 442·3	85·1+j 67·4
0·500	357·3−j 459·2	97·5−j 15·5	353·9−j 465·1	95·4−j 16·8
0·625	336·4−j 499·1	30·6−j 72·7	331·8−j 504·4	27·7−j 71·0
0·750	315·2−j 481·7	−30·0−j 45·3	311·7−j 486·2	−30·2−j 42·3
0·875	326·8−j 471·7	−39·1−j 1·7	323·9−j 477·4	−37·4−j 0·0
1·000	331·1−j 480·1	−17·8+j 25·3	327·4−j 485·7	−15·3+j 25·0
1·125	324·8−j 480·7	11·0+j 21·9	321·1−j 485·7	11·8+j 20·1
1·250	326·4−j 476·7	19·3+j 0·6	323·1−j 481·9	18·6−j 0·8
1·375	328·3−j 479·0	8·1−j 13·8	324·7−j 484·6	6·8−j 13·8
1·500	326·1−j 479·4	−7·4−j 11·4	322·6−j 484·6	−8·0−j 10·3
1·750	327·5−j 478·9	−3·6+j 9·3	323·9−j 484·2	−2·4+j 9·4
2·000	327·0−j 478·2	7·4−j 2·2	323·4−j 483·7	6·8−j 3·2
2·250	326·6−j 478·7	−5·1−j 3·7	323·1−j 484·0	−5·3−j 3·0
2·500	327·0−j 478·9	0·2+j 5·2	323·4−j 484·2	1·0+j 5·1
2·750	327·0−j 478·5	3·4−j 2·8	323·6−j 483·9	3·1−j 3·1
3·000	326·8−j 478·6	−3·7−j 0·8	323·3−j 483·7	−3·8−j 0·3

Table 3

Self- and mutual-impedances (in ohms) of two parallel collinear dipoles, $d = 0$, $a/\lambda = 0{\cdot}007022$

	Variational		Chang and King	
b/λ	$Z_s = R_s + jX_s$	$Z_m = R_m + jX_m$	$Z_s = R_s + jX_s$	$Z_m = R_m + jX_m$
		$h = 0{\cdot}25\,\lambda$		
0·550	99·6+j 32·8	30·0−j 5·9	96·4+j 35·5	29·0−j 4·2
0·600	97·8+j 33·8	18·6−j 11·8	94·5+j 36·1	18·1−j 10·6
0·650	97·5+j 34·5	10·3−j 13·6	94·1+j 36·7	10·2−j 12·6
0·750	97·7+j 35·1	−0·4−j 11·4	94·2+j 37·3	−0·1−j 10·8
0·875	97·9+j 35·0	−5·9−j 5·0	94·4+j 37·2	−5·5−j 4·9
1·000	97·9+j 34·9	−5·6+j 0·7	94·4+j 37·2	−5·3+j 0·5
1·125	97·8+j 35·0	−2·6+j 3·5	94·4+j 37·2	−2·5+j 3·2
1·250	97·9+j 35·0	0·5+j 3·4	94·4+j 37·2	0·5+j 3·2
1·375	97·9+j 35·0	2·3+j 1·6	94·4+j 37·2	2·1+j 1·5
1·500	97·9+j 35·0	2·3−j 0·4	94·4+j 37·2	2·2−j 0·4
		$h = 0{\cdot}50\,\lambda$		
1·050	325·8−j 500·1	−16·7+j 62·4	321·0−j 498·5	−9·7+j 49·2
1·100	319·6−j 486·6	2·3+j 42·1	317·2−j 489·0	4·3+j 33·3
1·125	319·8−j 482·7	7·0+j 35·1	317·6−j 485·8	8·2+j 28·1
1·250	325·4−j 476·0	15·7+j 12·3	322·6−j 481·7	14·1+j 8·7
1·375	328·2−j 478·0	13·1−j 1·2	324·5−j 483·7	10·7−j 2·6
1·500	327·3−j 479·3	6·0−j 7·0	323·7−j 484·7	4·0−j 6·4
1·625	326·5−j 478·9	−0·3−j 6·7	323·1−j 484·1	−1·1−j 5·4
1·750	326·8−j 478·4	−3·6−j 3·5	323·6−j 483·8	−3·2−j 2·4
1·875	327·0−j 478·6	−4·0−j 0·1	323·5−j 484·2	−3·1+j 0·5
2·000	327·0−j 478·8	−2·4+j 2·2	323·3−j 483·7	−1·4+j 2·0

The calculated theoretical self- and mutual-impedances for the non-staggered array of two elements have been presented in Table 1. The distance between the dipoles varies from 0·05 λ to 3 λ. For the sake of comparison the results obtained by Chang and King, originally expressed as admittances, have been converted into impedances and included in Table 1. The agreement between the results obtained by these two methods is remarkable. Agreement with the results presented by Popović was also found to be good.

Also, the impedances for the non-staggered dipoles are compared with the experimental data measured by Mack[7] and presented diagramatically in Figs. 2 and 3. Very good agreement between theoretical and experimental results in the case of half-wave dipoles (Fig. 2) can easily be noticed. Similar conformity is seen even in the case of full-wave dipoles (Fig. 3) with the exception of the imaginary part of self-admittance, B_s. Moreover, the shape of the theoretical diagram for B_s is very similar to the diagram of experimentally obtained data, the only difference being in that the theoretical diagram shows somewhat lower values.

The results obtained by the variational method for the arrangements in echelon and for the case of collinear dipoles are presented in Tables 2 and 3. For comparison, the results of Chang and King are also included in these tables. Again the comparison of these results reveals an excellent agreement, and both sets of results are in good agreement with those reported in reference 8.

A comparison of the theoretical and experimental results for echelon and collinear arrangements could not be carried out, since experimental results for the latter were not available. However, in view of the fact that the agreement is quite satisfactory in the case of non-staggered dipoles, one might be justified in concluding that a similar degree of agreement could be expected in all cases. This conclusion may also be said to follow from the fact that the results obtained by the present method are in good agreement with those obtained by two other, different methods.

A note on the relative advantages of the present method and the polynomial approach[8] might be perhaps useful, for placing the present method in a proper perspective. The variational method is known to yield values of input impedances which are for an order of magnitude more accurate than the approximate current distribution used. Therefore, the results obtained by the present method with the two-term trial function for current should be regarded as approximately corresponding to the third-order polynomial approximation. On average, this was indeed found to be close to the truth.

On the other hand, the variational approach

represents an optimization of the driving-point current only, distribution of current along the entire antenna length not being so accurate. The polynomial approach therefore perhaps has advantages in this respect over the present approach, particularly if h is significantly greater than $\lambda/4$.

5. Acknowledgments

The author is greatly indebted to Mr. A. Jovanović for his aid in obtaining numerical results reported here.

6. References

1. Chang, V. W. H. and King, R. W. P., 'On two arbitrarily located identical parallel antennas', *I.E.E.E. Trans. on Antennas and Propagation*, **AP-16**, p. 309, May 1968.
2. King, R. W. and Wu, T. T., 'Currents, charges and near fields of cylindrical antennas', *Radio Science*, **69D**, p. 429, March 1965.
3. Storer, J. E., 'Solution to Thin Wire Antenna Problems by Variational Methods', Doctoral dissertation, Harvard University, June 1951.
4. King, R. W. P., 'The Theory of Linear Antennas' (Harvard University Press, Cambridge, Mass., 1956).
5. Levis, C. A. and Tai, C. T., 'A method of analyzing coupled antennas of unequal sizes', *I.E.E.E. Trans. on Antennas and Propagation*, **AP-4**, p. 128, April 1956.
6. Surutka, J. V. and Popović, B. D., 'A Variational Method of Evaluating Impedances of Two Coupled Antennas of Unequal Sizes', Publ. of the Electrotechnical Faculty of Belgrade, Series: Mathematics and Physics, No. 197, 1967.
7. Mack, R. B., 'A Study of Circular Array', Cruft Laboratory, Harvard University, Cambridge, Mass., Techn. Rept. 383.
8. Popović, B. D., 'Analysis of two identical parallel arbitrarily located thin asymmetrical antennas', *Proc. Instn Elect. Engrs*, **117**, No. 9, p. 1735, September 1970.
9. Popović, B. D. and Surutka, J. V., 'A variational solution to the problem of asymmetrical cylindrical dipole', *I.E.E.E. Trans. on Antennas and Propagation*, **AP-19**, No. 1, p. 17, January 1971.

Manuscript first received by the Institution on 7th September 1970 and in final form on 2nd November 1970. (Paper No. 1388/Com. 45.)

STANDARD FREQUENCY TRANSMISSIONS—May 1971

(Communication from the National Physical Laboratory)

May 1971	Deviation from nominal frequency in parts in 10^{10} (24-hour mean centred on 0300 UT)			Relative phase readings in microseconds N.P.L.—Station (Readings at 1500 UT)		May 1971	Deviation from nominal frequency in parts in 10^{10} (24-hour mean centred on 0300 UT)			Relative phase readings in microseconds N.P.L.—Station (Readings at 1500 UT)	
	GBR 16 kHz	MSF 60 kHz	Droitwich 200 kHz	*GBR 16 kHz	†MSF 60kHz		GBR 16 kHz	MSF 60 kHz	Droitwich 200 kHz	*GBR 16 kHz	†MSF 60 kHz
1	− 299·8	+ 0·2	+ 0·1	599	602·5	17	− 300·1	0	0	585	594·1
2	− 299·9	0	+ 0·1	598	602·6	18	− 300·0	0	+ 0·1	585	593·9
3	− 299·9	+ 0·1	+ 0·1	597	601·1	19	− 300·0	0	+ 0·1	585	594·0
4	− 299·9	− 0·2	+ 0·1	596	601·3	20	− 299·9	0	+ 0·1	584	594·0
5	− 299·9	+ 0·1	+ 0·1	595	600·4	21	− 300·0	0	+ 0·1	584	594·3
6	− 300·0	0	+ 0·1	595	604·4	22	− 300·0	0	+ 0·1	584	593·9
7	− 299·9	+ 0·1	+ 0·1	594	603·3	23	− 300·0	0	0	584	594·3
8	− 299·9	+ 0·1	+ 0·1	593	602·1	24	− 300·0	− 0·1	+ 0·1	584	595·0
9	− 299·8	+ 0·1	+ 0·1	591	600·9	25	− 300·0	+ 0·1	0	584	595·6
10	− 299·9	+ 0·2	+ 0·1	590	598·9	26	− 300·0	0	+ 0·1	584	594·8
11	− 299·8	+ 0·2	+ 0·1	588	597·2	27	− 300·0	0	0	584	595·2
12	− 299·8	+ 0·2	+ 0·1	586	595·4	28	− 300·0	0	+ 0·1	584	595·4
13	− 299·9	+ 0·1	+ 0·1	585	594·8	29	− 300·0	0	+ 0·1	584	595·6
14	− 300·1	+ 0·1	+ 0·1	586	594·2	30	− 299·9	+ 0·1	+ 0·1	583	594·9
15	− 299·8	+ 0·1	+ 0·1	584	593·1	31	− 299·9	0	+ 0·1	582	594·9
16	− 300·0	− 0·1	+ 0·1	584	594·1						

All measurements in terms of H.P. Caesium Standard No. 334, which agrees with the N.P.L. Caesium Standard to 1 part in 10^{11}.

* Relative to UTC Scale; $(UTC_{NPL} - Station) = + 500$ at 1500 UT 31st December 1968.

† Relative to AT Scale; $(AT_{NPL} - Station) = + 468{\cdot}6$ at 1500 UT 31st December 1968.

U.D.C. 621.396.931.029.63

12·5 kHz Channel Spacing for Mobile Communications in the U.H.F. Band between 420 and 512 MHz

By

J. R. BRINKLEY,
C.Eng., F.I.E.R.E.†

First read at a meeting of the Institution's Communications Group in London on 25th November 1970 and subsequently presented at the 1970 Conference of the I.E.E.E. Vehicular Technology Group held in Washington D.C. from 2nd to 4th December 1970.

The historical and technical background which lead to the development of 12 kHz channel spacing at u.h.f. is described. The advantages and disadvantages of the narrower channel spacing, including problems of achieving satisfactory adjacent channel protection and frequency stability, are discussed. A series of field tests within the London metropolitan area is reported which demonstrated that coverage and quality of reception is as good as with 25 kHz channel spacing.

1. Introduction: The Need for More Channels for Mobile Radio

Mobile radio is one of the fastest growing sectors of the telecommunications industry and it has a history of continuous growth in every country in the world. In the United States there are approximately 3 million radiotelephones licensed and the growth rate is over 10% per annum. In the United Kingdom there are approximately 150 000 mobile radiotelephones and the number of mobiles and growth rate is 15% per annum. In Germany the number of mobiles and growth rate is about the same as in the U.K. Scandinavian countries have highly developed mobile radio services with correspondingly high growth rates, as do Canada, Australia and New Zealand.

Although the numbers of vehicles fitted may seem large and the growth rates high, the percentage of vehicles fitted in each country is still quite small, generally lying between 0·5 and 3%. There is therefore no question of a market saturation if adequate frequencies can be made available and indeed the only tendency for slower growth to be manifest occurs when adequate channels become difficult to obtain in any given area or country.

There are further new factors coming into prominence which will cause accelerated growth and increased demands for channels. These include the introduction and widespread adoption of pocket or personal radiotelephone services for a wide variety of purposes. This market may well prove to be as big again as vehicle mobile radio. The projected use of data transmissions to and from vehicles will also add to the demand for new channels. Such services may require a standard of coverage and freedom from interference not ordinarily available on voice mobile channels.

2. Increasing the Frequency Space for Mobile Radio

There are only two practical methods of providing the additional channels needed for the future expansion of mobile radio. One is to increase the total frequency space available and the other is to decrease the spacing between channels, thereby making more channels per megahertz available.

The best opportunities for increasing the frequency space available to mobile radio lie in the u.h.f. band on either side of the existing mobile band, 450–470 MHz. This band has the important advantage of having international recognition and acceptance as a mobile band. It also has been found to have important operational attractions including very low general noise levels and in particular an almost complete lack of ignition interference. Excellent penetration of streets and buildings is a further desirable characteristic of the band. The very small antennas required also make it highly suitable for pocket radiotelephone services.

The extension of this band has taken divergent courses in the United States and Europe. In Europe a number of countries including France, Germany, Denmark, Sweden and Finland are extending the frequency allocation downwards to 420 MHz. In the U.S. proposals to extend the band upwards to 512 MHz on a shared basis with television services are now approaching implementation. In the U.K. at the time of writing, band extension is under consideration by the Government. Any proposal to reduce mobile channel spacing in the u.h.f. band should therefore take into consideration the technical factors affecting a band of frequencies lying between 420–512 MHz.

These band extensions will undoubtedly relieve the immediate pressure for channels. How much relief is obtained in Europe will depend upon how much of the 420–450 MHz band is released for mobile services. In the U.S. the degree of relief achieved will depend upon the success of somewhat complicated television band-sharing proposals.

It is clear however that mobile radio's long-term growth problems cannot be solved by band extension alone and that if further channel splitting is practicable at u.h.f. it will also need to be adopted. 12·5 kHz channelling at u.h.f. should therefore be examined carefully and if practicable it should be adopted.

The potential advantage which will accrue from a further splitting of u.h.f. channels is very substantial.

† ITT Europe Inc., 190 Strand, London WC2R 1DU

The number of double-frequency channels at 450–470 MHz would increase from 400 to 800, the number between 420–450 MHz would increase from 600 to 1200. The number in the band 470–512 MHz would go up from 840 to 1680 in those geographical areas where sharing with television is permitted.

As will be discussed later, the maximum benefit from channel splitting is only fully effective if the double-frequency principle of frequency allocation is adhered to. (See Appendix 1.) This is fortunately the case for the u.h.f. band in the United States and Great Britain and the majority of countries throughout the world.

3. Outline History of Mobile Radio Channel Spacing at V.H.F. and U.H.F.

When mobile radio began in the late 1940s the initial channel spacing in the U.S. in the v.h.f. 150 MHz band was 120 kHz. This spacing was successively reduced, first to 60 kHz and then to 30 kHz where it now stands. In the U.K. channels initially spaced at 100 kHz were split first to 50 kHz and then in 1963 to 25 kHz. In January 1968 the British Post Office introduced 12·5 kHz channelling both in the U.K. high-band (165–173 MHz) and in the low band (71·5–88 MHz).

The channel spacing in the u.h.f. band, initially 50 kHz, was recently reduced to 25 kHz both in the U.S. and the U.K. 25 kHz channelling at u.h.f. is the current standard throughout the world, except in Germany and Holland, where it is interesting to note that 20 kHz spacing has been adopted. (See also Appendix 2.)

4. 12·5 kHz Channel Spacing at V.H.F. in the United Kingdom

The introduction and subsequent history of 12·5 kHz channel spacing at v.h.f. in the U.K. in 1968 is interesting and relevant to our subject. Of the 150 000 mobile stations operating in the U.K. about 100 000 are already operating on 12·5 kHz channelling and the proportion remaining on 25 kHz at v.h.f. is rapidly diminishing. (The precise numbers are not known because of uncertainties in the licensing procedures.) The experience gained in the three years of operation by the allocating authority, by the industry and by the users has been extremely satisfactory. No special problems have arisen in relation to frequency stability. No serious complaints of adjacent channel interference have been registered and the number of available channels has been greatly increased to the benefit of users and manufacturers alike. Intermodulation problems have been few and have not led to any serious loss of frequency channels. They have been mainly confined to the type of main station problems which can be treated by the use of filters and circulators, and by adjusting antenna spacing.

The successful introduction of 12·5 kHz channelling at v.h.f. was confirmed by the British Government when in opening a new frequency band, the 'mid-band' (105–108 MHz and 138–141 MHz) in 1969, it decided to adopt 12·5 kHz spacing, a decision which was fully endorsed by users and manufacturers.

Careful comparison of the 'on-channel' performance of 12·5 kHz channelling systems with the 25 kHz channelling systems which they are replacing has shown that only a minor degradation of service is experienced. This takes the form of a small increase in ignition interference, chiefly in the mobile receiver, due to pulse lengthening caused by the narrower bandwidths of the 12·5 kHz receivers.

In the majority of systems the increased ignition interference has passed without comment by users and in no case has it given rise to serious complaint. A compensatory advantage of the narrower bandwidth employed is that the incidence of interference from diathermy and stray carriers such as those radiated by television receiver local oscillators has been correspondingly reduced. The change of channelling standard has not materially altered the limit range of systems but has led rather to some increase in noise levels near limit range.

5. Channel Spacing in the V.H.F. 150 MHz Band in the United States

Unlike the U.K., the U.S. has not split its channelling standards in the 150 MHz v.h.f. band and the present standard is 30 kHz. The reasons for this are probably as follows:

(1) The U.S. has extensive single-frequency allocations in the 150 MHz band. Such allocations are very prone to interference chiefly due to intermodulation between fixed stations. Channel splitting unfortunately causes the number of possible intermodulation products in single frequency allocations to rise sharply. Under these conditions splitting the channels does not necessarily give a large increase in channel availability and its introduction could cause severe administrative problems.

(2) In the U.S., fixed and mobile power levels are some 6–10 dB higher than in the U.K. and Europe. This aggravates intermodulation problems and adjacent channel interference problems.

(3) The U.S. uses f.m. exclusively in the v.h.f. bands. The deterioration due to ignition pulse lengthening may possibly appear to be worse in f.m. than in a.m. when channels are split. The majority of systems in the U.K. are a.m., although numerous f.m. 12·5 kHz v.h.f. systems are working successfully in that country. (See also Appendix 3.)

6. 12·5 kHz Channel Spacing at U.H.F.

The introduction of 12·5 kHz channel spacing in the U.K. having been entirely satisfactory, it is natural to consider how these advantages can be obtained at u.h.f. since, as has been stated, most countries are using the double-frequency allocation system in this band.

There are in fact substantial reasons why 12·5 kHz channelling should be even more successful at u.h.f. than at v.h.f., as follows:

(a) Double-frequency working has been much more widely adopted throughout the world at u.h.f. than at v.h.f. Double-frequency working, by reducing the dynamic range of signals involved between systems, eases the problems of intermodulation and adjacent channel protection.

(b) Ignition noise is almost entirely absent at u.h.f. so that the problem of pulse lengthening of ignition impulses in the narrower 12·5 kHz i.f. filters does not arise as a significant issue. The improvement in stray carrier protection will still apply as at v.h.f.

(c) U.h.f. fixed stations typically employ high gain omni-directional antennas. These antennas reduce the field strength in the immediate vicinity of the fixed station and in consequence reduce the dynamic range of the signals to which both fixed and mobile receivers are exposed.

(d) Transmitter power levels used at u.h.f. are conventionally lower than at v.h.f. This trend is due to two causes. First, power has been more difficult to generate at u.h.f., particularly with transistors, and secondly, its effect on range is less rewarding than at v.h.f.

(e) Many typical u.h.f. applications are for short range systems, including lower power portable and pocket use. Such systems generally have a reduced dynamic range of signals and this is favourable to the narrower channel spacing condition.

Against these advantages have to be set the following considerations:

(f) There is a deterioration in signal/noise ratios at 12·5 kHz spacing due to the reduced deviation. This occurs chiefly in the good signal areas and amounts to 2·5–3 dB. This is true at v.h.f. and u.h.f. Comparative SINAD measurements are shown in Fig. 1.

(g) It is more difficult to obtain the required adjacent channel protection at 12·5 kHz than at 25 kHz. This is true at v.h.f. and u.h.f.

(h) Frequency stability requirements are more stringent at u.h.f.

6.1 *Adjacent Channel Protection*

Protection against adjacent channel interference becomes more difficult to achieve as channel spacings are narrowed. Apart from frequency stability, which is considered later, there are two main reasons for this.

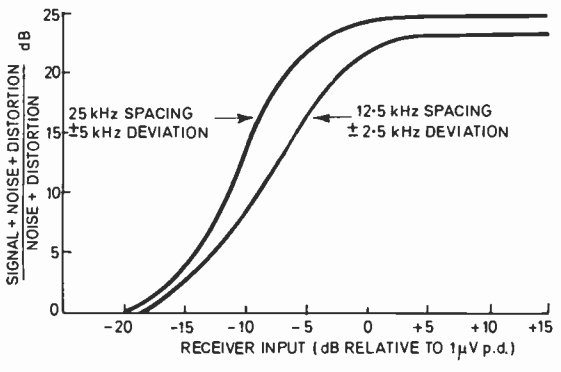


Fig. 1. SINAD vs. receiver input for 12·5 kHz and 25 kHz channel spacing.

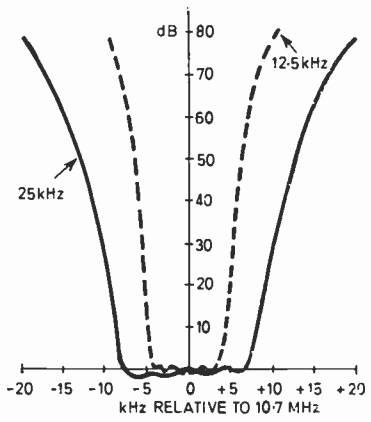


Fig. 2. Crystal filters for 12·5 kHz and 25 kHz channel spacing.

6.1.1 Receiver selectivity

The i.f. filter shape factor is somewhat poorer at 12·5 kHz as compared to the shape factor of 25 kHz filters. Comparative curves are shown in Fig. 2. The 12·5 kHz filter shape factor shown is that of the current state of the art and is probably capable of some improvement with further filter development. Since, however, the same 12·5 kHz filters are used in v.h.f. and u.h.f. systems, no new factor arises at u.h.f. in this respect.

6.1.2 Transmitter sideband radiation into the adjacent channel

As the channel spacings are narrowed, transmitter deviation must be reduced. At 12·5 kHz channelling at v.h.f. a maximum deviation of 2·5 kHz has been specified and found satisfactory and the same figure is recommended for u.h.f. The u.h.f. specification proposed allows a further 500 Hz frequency drift so that to this extent adjacent channel sideband radiation will be greater at u.h.f. in the worst drift condition. Alternative courses of reducing deviation to 2 kHz or slightly increasing frequency stability can be considered.

6.2 *Frequency Stability*

The assumptions made in proposing 12·5 kHz channelling at u.h.f. are that crystal oscillator stabilities of $\pm$ 5 parts in 10^6 will be adopted in mobile equipment and $\pm$ 2·5 parts in 10^6 in fixed station equipment. The mobile stability requirement is readily obtained over the temperature range of $-$ 10°C to + 50°C. Where temperatures down to $-$ 30°C are required this stability may be obtained either by crystal compensation or simple thermostat-controlled heater mats activated at 0°C. These temperature stabilities are currently achieved and in wide use in 25 kHz u.h.f. systems.

The higher temperature stability recommended at the fixed station is readily obtained and in fact surpassed by the use of a simple oven of the proportional control type fitted over the fixed station crystals. The development of the proportional control crystal oven has in fact made a useful contribution to the introduction of narrower channelling. One minor trouble experienced with the introduction of 12·5 kHz channelling at v.h.f. has been with the earlier bi-metallic strip thermostatic crystal

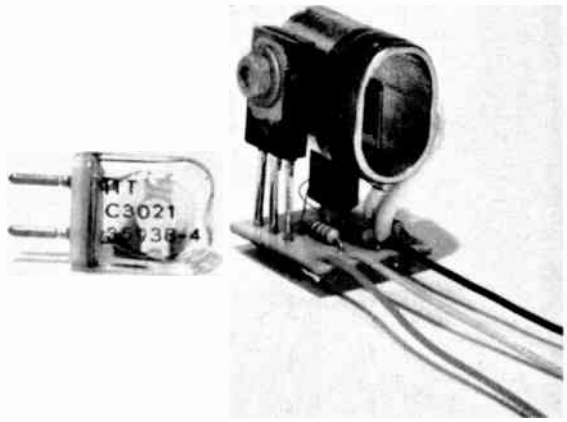


Fig. 3. Proportional control crystal oven.

ovens used at fixed stations. If the thermo-mechanical control elements fail then the crystal oscillator can drift into the adjacent channel causing temporary but serious interference.

This is avoided in the ITT STAR equipment in two ways. First, bi-metallic thermostats have been eliminated. Secondly, the crystals, although operated and adjusted to frequency in the equipment at 50°C, are cut for 20°C. An oven failure (which is very unlikely to take place and has not in fact been experienced) would only cause a minor drift in frequency amounting to less than 5 kHz even if the ambient temperature at the fixed station was as low as −10°C.

A further advantage of the proportional oven is that it does not cycle in the manner characteristic of bi-metallic thermostat-controlled ovens. This makes an even control of temperature possible to within 1 degC (Fig. 3).

A further 10% improvement in frequency stability will be required at 512 MHz. This can either be obtained by reducing the fixed station frequency tolerance from ± 1 kHz to ± 500 kHz or by reducing mobile station tolerances to ± 4 parts in 10^6 whichever is thought to be more convenient.

6.3 *Crystal Ageing*

Crystal ageing, that is change of frequency with age, is a factor which must be considered in mobile systems. At u.h.f., and particularly at 12·5 kHz channel spacing, it becomes significant.

All crystal ageing problems can of course be eliminated if one is prepared to readjust crystal frequencies often enough. Since however crystal adjustment takes time and costs money it is obviously desirable to minimize ageing and hence the number of adjustments carried out during the life of the equipment.

In this respect the ITT STAR equipment incorporates an important advance in technique over its v.h.f. 12·5 kHz predecessors—namely that it uses glass encapsulated crystals. These crystals have particularly good ageing characteristics due to the absence of impurities and partial leaks which are experienced with metal can

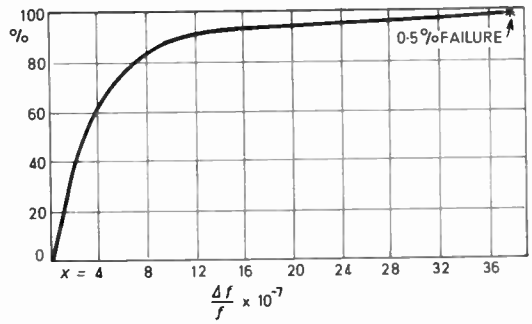


Fig. 4. Percentage of 200 crystals ageing less than $\pm \alpha \times 10^{-7}$ between day 30 and day 512.

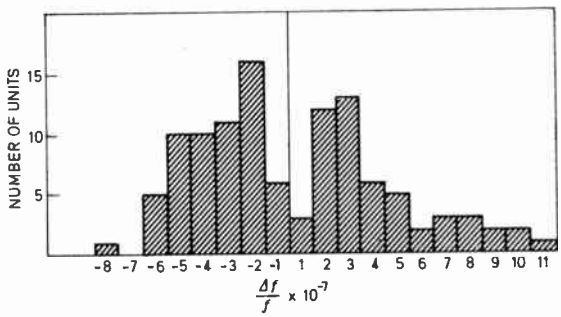
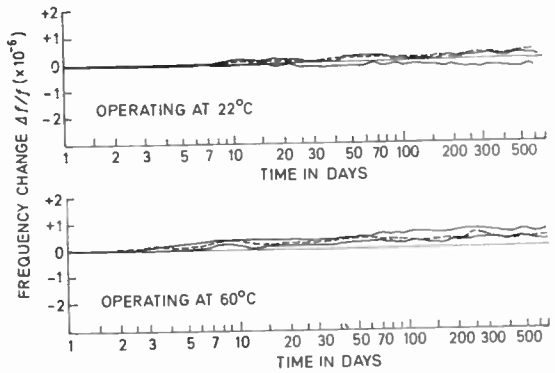
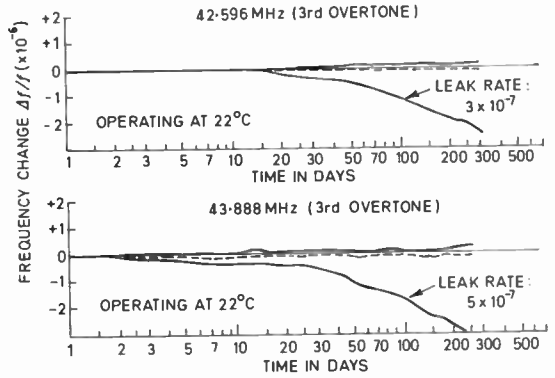


Fig. 5. Ageing of all-glass third-overtone crystals at 85°C for one year from date of manufacture.



(a) HC-6/U cold weld type. 3·2768 MHz.



(b) HC-18/U cold weld type.

Fig. 6. Frequency ageing characteristics.

crystals sealed by hot soldering methods. The results of ageing runs carried out on two sets of glass encapsulated crystals are shown in Figs. 4 and 5.

From these tests it can be seen that if the crystals are pre-aged then not more than one adjustment in the life of the equipment should suffice to keep the crystal error due to this cause within 1 part in 10^6. This represents a very acceptable performance and it can be related to the F.C.C. mandatory requirement which is to check and correct frequencies once per year.

Cold weld metal crystals appear to achieve a similar freedom from ageing and are an important alternative to glass encapsulation (Fig. 6).

6.4 *Field Tests*

In order to test the effectiveness of 12·5 kHz working at u.h.f. it was decided in 1969 to carry out laboratory and field tests on equipment meeting the standards described.

The equipment used was standard production equipment from the ITT STAR range of 25 kHz u.h.f. equipment shown in Fig. 7. Fixed stations, mobiles and pocket radiotelephones were modified to the new specification. The modification, which is very simple, consists simply of changing the 10·7 MHz i.f. crystal filter from a 25 kHz channelling filter to a 12·5 kHz channelling filter and reducing the deviation settings of transmitters from 5 kHz to 2·5 kHz. In addition at the fixed station the small proportionally-controlled oven was fitted over the transmitter and receiver crystals. This oven controls the temperature of the crystals within a degree of 50°C. A change of one resistor in the receiver squelch circuit is also necessary.

A 12·5 kHz channelling u.h.f. system using modified equipment of this type has been under test in London for over 12 months. The layout of the London test system is shown in Fig. 8. The metropolitan London area was chosen as giving severe and widely varying field conditions, including undulating terrain, high buildings, narrow streets, under-passes and heavy traffic.

Fig. 8. Coverage map for u.h.f. mobile system with 12·5 kHz channel spacing.

The site for the fixed station was chosen at Hampstead where the ground level is about 400 feet. A 10 dB gain omni-directional antenna is installed 70 feet above ground level on top of a block of flats. The fixed station equipment is remotely controlled by land line from Mobile Radio Laboratories at New Southgate, over an 8-mile telephone line. A permanent 25 kHz u.h.f. system which gives good coverage of London is operated from the same site, using the same antenna. This gives a valuable standard of performance comparison.

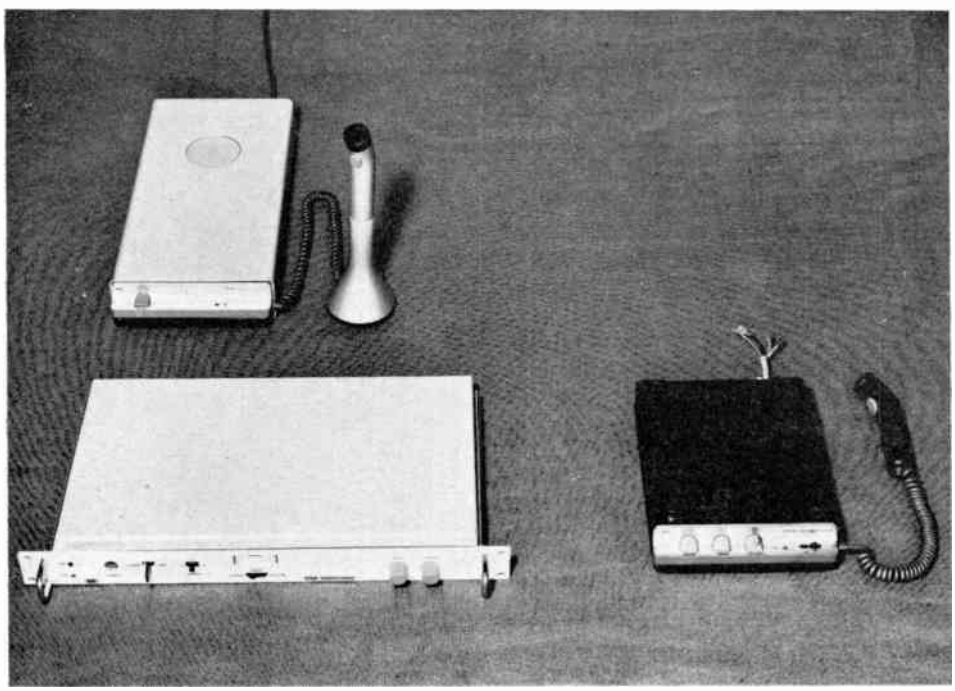


Fig. 7. U.h.f. fixed station and mobile equipment for 12·5 kHz channel spacing.

The results of tests and observations made over the 12-month period show that the coverage at 12·5 kHz is virtually indistinguishable from the coverage obtained from the 25 kHz system. In terms of absolute range, signal/noise, ignition levels and speech quality the difference between the two systems subjectively is barely perceptible.

In terms of adjacent channel performance the mobile receiver squelch does not open, no matter how close the mobile approached the fixed station on the adjacent channel. This is a better result than is obtained at 12·5 kHz v.h.f. where the adjacent channel can be heard intermittently for several hundred yards around the same site. The improvement in this respect at u.h.f. is no doubt due to the effect of the high-gain antenna described in Section 6 under (c).

7. Pocket Radiotelephones

U.h.f. pocket radiotelephone systems have made great strides in the past few years: some 25 000 equipments are in use by the police in the U.K., and as many again are used in commercial and industrial applications. A typical unit is shown in Fig. 9. It weighs 0·454 kg (16 oz) and measures 18·3 × 6·35 × 3·17 cm (7·3 × 2·5 × 1·25 in).

Fig. 9. U.h.f. pocket radiotelephone for 12·5 kHz channel spacing. In the interior view the aerial is at top, transmitter and receiver sections are to the left and right of the upper half of the chassis respectively, the press-to-transmit switch is on the left and the on/off and volume control is just above the loudspeaker; the microphone is at the bottom.

Advantage has been taken of the small antenna required at u.h.f. to enclose the antenna, thus greatly improving the convenience and appearance of the equipment.

This sector of mobile radio now has the fastest growth and in the long run may far exceed the vehicle market in numbers of units and monetary volume. U.h.f. channel splitting to 12·5 kHz is therefore of great interest and significance in these applications.

A main question arising in such small equipments is the provision of the i.f. filter. A 12·5 kHz channel spacing crystal filter measuring only 18·3 × 12 × 15·3 mm (0·73 × 0·47 × 0·6 in) and operating at 10·7 MHz is now available. This filter provides the necessary receiver adjacent channel selectivity. No other important difficulty has been experienced in designing 12·5 kHz versions of the 25 kHz equipments.

The conditions under which pocket systems using these equipments operate are particularly appropriate to 12·5 kHz for the following reasons:

(1) The low power used (generally a fraction of a watt) and lower mobile antenna efficiencies result in reduced exposure to high signal levels.

(2) The fixed stations are typically far more widely dispersed than vehicle mobile system fixed stations which tend to concentrate in the same areas. Pocket systems may be located for instance, at building sites, docks, harbours, railway tunnels, airports and factories. In general such installations are on geographically separate sites and multiple systems covering the same area are the exception rather than the rule. In this they tend to differ from mobile systems where many systems are set up to cover the same area.

(3) Where large numbers of pocket systems do congregate, as at airports, there is much more scope for changing fixed station antenna locations if interference problems arise.

8. Conclusions

The standards of frequency stability required for 12·5 kHz spacing in the u.h.f. band may be readily obtained by the methods recommended. These involve chiefly a change to glass encapsulated or cold weld crystals. The slight degradation in adjacent channel protection consequent upon the 500 Hz increase in the mobile frequency tolerance can be avoided if desired by reducing system deviation from 2·5 kHz to 2 kHz and accepting a 2 dB reduction in signal/noise ratio or by reducing frequency tolerances slightly.

The benefits which will accrue from the introduction of 12·5 kHz channelling are considered to be very great, provided double-frequency allocation is adopted. In view of the expansion pressures on all mobile radio frequency allocations it is strongly recommended that 12·5 kHz channelling at u.h.f. be introduced wherever possible, with as early a timescale as can be agreed by the parties concerned. It should be noted in this connexion that the benefits of 12·5 kHz spacing would be greater in the proposed new extensions of the u.h.f. band since no problems of interleaving with existing 25 kHz channels will occur.

9. Acknowledgments

The data on crystal ageing in Figs. 4 and 5 were supplied by ITT Crystal Division, England, and in Figs 6 and 7 by the Toyo Communication Company, Tokyo, to whom thanks are expressed.

10. Appendix 1: Single- and Double-Frequency Allocation

It is not generally realized that there are two distinct methods of allocating radio frequencies for mobile radio systems. These are known as the single- and double-frequency allocation methods.

The single-frequency allocation method is the simpler to understand. All fixed and mobile stations are on the same single frequency for transmit and receive functions in this system. All stations can inter-communicate directly when within direct range of each other. Single frequency allocation appears at first sight to be the more economical way of allocating frequencies but paradoxically it turns out in fact to be the more wasteful method.

The double-frequency allocation method requires two separate frequencies, one for the fixed to mobile transmission and the other the other for mobile to fixed. In this system mobiles can talk directly to each other only via the fixed station on a 'talk-through' basis. Co-channel fixed stations cannot hear each other and it is this feature which brings one of the main advantages of double-frequency allocation in terms of frequency economy.

Single-frequency fixed stations sharing the same channel are liable to interfere with each other unless the spacing between the two stations is quite large. Spacings of 50 miles (80 km), 100 miles (161 km), or even 180 miles (289 km) may be necessary if the two stations are to function without mutual interference. Since fixed stations transmit for about 50% of the total system time, co-channel interference of this kind between fixed stations can lead to a very serious impediment to traffic handling.

In the double-frequency case the two fixed stations cannot hear each other, no matter how close, because of the frequency transposition. Inter-system co-channel interference can occur only between fixed stations and the mobiles of the system sharing the channel. The two systems may, as a result, be more closely spaced by a factor typically of about three.

Since frequencies are allocated on an area basis, this system spacing improvement factor can be obtained in two dimensions, thus making possible 9 times as many stations in a given area using the double frequency method. Since two frequencies per system are used instead of one the advantage has however to be reduced again by a factor of two but the net gain is obviously substantial.

A further important advantage accrues in the double-frequency case in the protection it gives against other than co-channel interference. In covering a given area there is always a tendency for many fixed stations to

congregate on a small number of high sites, e.g. hill tops, water towers or skyscrapers. This leads in the single-frequency method to serious difficulties due to intermodulation blocking, desensitization, etc., because the receivers used cannot provide adequate 'front-end' protection. In this situation many channels quickly become unusable.

In the double-frequency method all transmitters are allocated in one frequency block and all receivers in a second block, well separated from the first. In these circumstances receiver front-end protection is achieved and channels are seldom lost from interference problems.

A third advantage of the double-frequency method, which is not practical with single-frequency working, is that it enables connexions to be made to the telephone network.

11. Appendix 2: Channel Spacings (kHz) in use in Various Leading Countries in the World

	25/50 MHz†	50/100 MHz	100/150 MHz	150/174 MHz‡	420/470 MHz
UNITED STATES	20	30	30	30	25
UNITED KINGDOM	—	12·5	12·5	12·5	25
GERMANY	20	20	20	20	20
FRANCE	25	25	20	20	25/50
BELGIUM	50	20/25	20/25	20/25	20/40
HOLLAND	—	25	20	20	20
DENMARK	25	25	25	25	25
NORWAY	25	25	25	25	25/50
SWEDEN	25	25	25	25	25
AUSTRALIA	25	25	30	30	25
NEW ZEALAND	12·5	25	25	25	25
FINLAND	25	25	25	25	25

† Excluding Citizen's Band. ‡ Excluding Maritime Band.

12. Appendix 3: A.M. and F.M. Aspects of 12·5 kHz Channel Spacing

The development of v.h.f. mobile radio in the U.K. has predominantly been based on amplitude modulation for reasons which are now largely historic. Nevertheless, about 10% of these v.h.f. installations, totalling several thousand systems, use f.m.

The results obtained with the two different modulation systems using 12·5 kHz equipment are not widely different. The communication range, for example, is substantially the same. Ignition noise levels are slightly different, the difference being more noticeable in the vehicles since these are exposed to higher ignition interference levels.

When both systems have zero frequency errors the ignition noise levels are about the same. When the systems are off-tune towards the limit of their frequency tolerances there is little doubt that the ignition noise levels are higher in the f.m. case. The audible characteristics of the ignition pulse are however different in the two systems and the assessment of the relative annoyance is therefore to some degree subjective.

The fact that a.m. suppresses ignition noise at least as well as f.m. does not even now seem to be appreciated in the U.S. A statement quite recently by Richard T. Buesing† infers that a.m. suffers badly from ignition interference. This has not been the experience in the U.K.

It is interesting to note that in 1968 when the new v.h.f. mid-band was opened in the U.K. the Fuel and Power Industry decided to standardize on a.m. This decision was made by an industry with a long experience of both systems and indicates that in its view there were marginal advantages in favour of a.m.

The a.m.–f.m. controversy which has enlivened the discussion of mobile development over the past 25 years is not likely to be a feature of u.h.f. development. There are a number of reasons for this. First, since there is negligible ignition interference at u.h.f., the scope for argument on this point disappears. Secondly, since the f.m. deviation ratios are low in the first place, the differences in the signal/noise ratios which can be achieved in the two systems are small.

In addition, for commercial reasons the U.K. manufacturers have not developed a.m. at u.h.f. Having no historic commitment to either system in this band and not wishing to develop and manufacture two systems where one will suffice, the manufacturers have independently and without exception chosen to concentrate exclusively on f.m. particularly since this is the system required in the export market.

† 'Modulation methods and channel separation in the land mobile service', *I.E.E.E. Trans. on Vehicular Technology*, **VT-19**, No. 2, pp. 187–206, May 1970.

Manuscript received by the Institution on 7th December 1970. (Paper No. 1389/Com. 46.)

U.D.C. 621.372.5:621.374.5

Method of Synthesis of Non-minimum-phase Transfer Functions for Time-delay Simulation

By

Professor B. D. RAKOVICH,
Dip.Eng., Ph.D.†

and

B. DJURICH,
Dip.Eng.‡

The paper presents a new method for designing non-minimum phase rational approximants of the ideal delay function e^{-s} suitable for the applications where the frequency spectrum of the input signal occupies large bandwidth. It is shown by theoretical considerations that for this purpose the most suitable type of delay characteristic is the one approximating to a constant delay over a frequency range extending from zero to a frequency $\omega < \omega_c$ and having a relatively large delay peak at the end of the passband (ω_c). The amplitude of the initial transient ringing (precursor) and the overshoot in the transient response of the filter mainly depend on the value of the peak in the delay characteristic. The polynomials with two variable parameters that exhibit this type of delay response are then introduced and shown that the precursor and overshoot can be adjusted to any prescribed value by varying the free parameters. Extensive tables are presented enabling direct determination of the fourth-order transfer function with three right-half-plane zeros for almost any practical prescribed values of the precursor and overshoot. The method can be extended to higher-order networks and as an example a table containing data on the sixth-order functions with five right-half-plane zeros is also included. A comparison of the transient responses reveals that the technique proposed yields an improvement over all other methods so far described.

1. Introduction

In analogue computer studies of systems having inherent transport delay rational fraction approximations of $e^{-\tau s}$ are usually employed to simulate time delay in the system since the ideal delay function $e^{-\tau s}$ cannot be synthesized by lumped parameter circuit used in the analogue computer.[1, 2, 3] Unfortunately, Chebyshev all-pass delay approximants are not useful in those applications where the frequency components of the signal to be delayed occupy very large bandwidth, for example, such as in the case of an input step signal. This is due to the fact that the transient response of all-pass circuits to a unit step input signal has a very large initial transient ringing or delay precursor.

King and Rideout[4] have described a method of time domain synthesis of the fourth-order non-minimum phase transfer functions with three zeros in the right half of the complex frequency plane that yield smaller initial transient ringing but with somewhat increased rise-time of the output pulses than the comparable Padé approximants. Recently, further considerable improvement in the transient response of delayed pulses has been obtained by Deliyannis.[5] His method is based on a technique of rational approximation of the delay operator e^{-s}, originally presented by Budak,[6] which consists in splitting e^{-s} into two parts

$$e^{-s} = \frac{e^{-ks}}{e^{-(k-1)s}} \quad 0 \leqslant k \leqslant 1 \qquad \ldots\ldots(1)$$

and then approximating independently the numerator and denominator with the third- and fourth-order Bessel polynomials respectively to obtain the maximally flat type of delay characteristic. Another set of solutions is obtained by approximating e^{-ks} and $e^{-(k-1)s}$ with the third- and fourth-order polynomials that provide a Chebyshev type of delay response. In Deliyannis's work[5] the parameter k is used to adjust -3 dB bandwidth of the resulting active filter. A different approach to the same problem has been proposed by Allemandou[7] who derived the transfer function approximating, in modulus and phase, the exponential function around the frequency origin. Unfortunately, these functions yield even larger values of the delay precursor than those obtained with the Padé approximants. Most recently, Pongrarit and Park[8] used the truncated continued fraction expansion of e^s to derive the approximating transfer functions. It is known, however, that continued fraction expansions are related to Padé approximations[9, 10] and it can be easily verified that, in fact, by truncating the continued fraction of e^s, Pongrarit and Park derived Padé approximants to e^s.

This paper presents a new method for determining the rational function approximations of equation (1) which enable any prescribed specifications in respect of the values of the precursor and overshoot to be met in practical design. The ratio of the time delay to the rise-time in the step response (T_d/T_r) for these transfer functions compares favourably with the results obtained by any other method so far described. First, a discussion on the dependence of the values of the precursor and overshoot in the step response on the shape of the phase characteristic of the filter is given. Then, based on these considerations the polynomials with two variable parameters in the numerator and denominator of the rational function approximation of $e^{-\tau s}$ are defined. The motivation for introducing two variable parameters in the approximating polynomials is that by optimizing these free parameters according to the prescribed specifications on the maximum tolerable precursor and overshoot various types of transient response of the resulting network can be obtained. Complete data for the design

† Faculty of Electrical Engineering, University of Belgrade, Yugoslavia.
‡ Faculty of Electronic Engineering, University of Nish, Yugoslavia.

of the fourth-order transfer function with three right-half-plane zeros are tabulated but it is shown that the procedure described can easily be extended to higher-order networks.

2. Approximation Technique

2.1. *Normalization*

The most important parameters in measuring the quality of a delay network are the delay/rise-time ratio (T_d/T_r), the overshoot ($p\%$) and the maximum value of the precursor ($a\%$). Since, if $F(\omega)$ is the Fourier transform of $f(t)$ then $1/kF(\omega/k)$ is the Fourier transform of $f(kt)$, these quantities are independent of any scaling in the frequency domain. Hence, instead of (1) we can write

$$e^{-(k+1)s} = \frac{e^{-s}}{e^{ks}} \simeq \frac{N_{n-1}(-ks)}{N_n(s)} \qquad \text{.....(1')}$$

To facilitate the comparison the polynomials $N_{n-1}(-ks)$ and $N_n(s)$ can also be normalized to unit cut-off angular frequency by substituting

$$ks\left(\frac{a_{n-1}}{a_0}\right)^{1/(n-1)} = kp$$

and

$$s\left(\frac{b_n}{b_0}\right)^{1/n} = p$$

respectively, so that the rational approximant to be determined takes the form

$$F_{n-1,n} = \frac{N_{n-1}(-kp)}{N_n(p)} = \frac{1-c_1(kp)+c_2(kp)^2-c_3(kp)^3+\cdots(-1)^{n-1}(kp)^{n-1}}{1+b_1p+b_2p^2+b_3p^3+\cdots+p^n} \qquad \text{.....(2)}$$

As the parameter k varies from zero to one, the zeros of the transfer function (2) move along radial lines from infinity towards the origin, while the pole locations remain unchanged. It should be noted that, in contrast with the technique used in Reference 5, in the procedure proposed in this paper the parameter k serves primarily to change the shape of the group delay response. On the other hand, the ω_{3dB} bandwidth of the filter can be adjusted to any prescribed value simply by multiplying both the zeros and poles of the resulting transfer function by a suitably chosen constant. As mentioned before, this will not affect the value of T_d/T_r.

2.2. *Transient Response of Ideal Low-pass Filters*

It is well known that the ideal low-pass filter, Fig. 1, is defined by the following conditions

$$A(\omega) = 1 \quad \text{for } \omega < \omega_c$$
$$A(\omega) = 0 \quad \text{for } \omega > \omega_c \qquad \text{.....(3)}$$
$$\phi(\omega) = \tau_1\omega \quad \text{for } \omega < \omega_c$$

$\phi(\omega)$ is an arbitrary function for $\omega > \omega_c$, and is not physically realizable since the criterion of Paley and Wiener is not satisfied for this type of magnitude response.[11] Nevertheless, some important conclusions

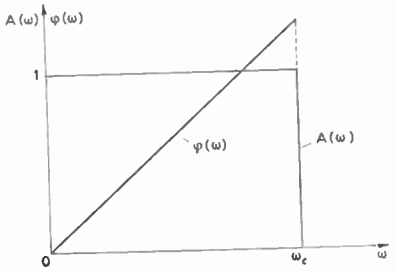


Fig. 1. Ideal low-pass filter.

for physically-realizable networks can be drawn from the study of the transient behaviour of this idealized filter function. The response of this filter to a unit step input as found by standard method[12, 13] is

$$V_2(t) = \frac{1}{2} + \frac{1}{\pi}\,\mathrm{Si}\,[\omega_c(t-\tau_1)] \qquad \text{.....(4)}$$

where

$$\mathrm{Si}\,(x) = -\int_x^\infty \frac{\sin t}{t}\,dt$$

is the sine integral.[14] The transient response (Fig. 2) is symmetrical with respect to the point $t = \tau_1$ so that the overshoot ($p\%$) is equal to the peak value of the precursor ($a\%$), $p = a = 9\%$.

Now, suppose the magnitude response of the filter remains unchanged while the phase characteristic in the passband is modified so that it takes the form shown in Fig. 3(a) or 3(b). In Fig. 3(a), the slope of the phase characteristic is increased near the end of the useful band corresponding to a peak in the group delay characteristic. In Fig. 3(b) the group delay is also constant over the largest part of the passband but it decreases near the cut-off frequency. The output voltage in response to a unit step input can be obtained in the following form (see Appendix 7.1):

$$V_2(t) = \frac{1}{2} + \frac{1}{\pi}\,\mathrm{Si}\,[\omega_c m(t-\tau_1)] + \frac{1}{\pi}\cos a_0\left\{\mathrm{Si}\left[\omega_c(t-\tau_1)-\frac{a_0}{m}\right] - \mathrm{Si}\,[\omega_c m(t-\tau_1)-a_0]\right\} - \frac{1}{\pi}\sin a_0\left\{\mathrm{Ci}\left[\omega_c(t-\tau_1)-\frac{a_0}{m}\right] - \mathrm{Ci}\,[\omega_c m(t-\tau_1)-a_0]\right\} \qquad \text{.....(5}$$

where $a_0 = \omega_c m(\tau_2-\tau_1)$ and $a_0 = \omega_c m(\tau_1-\tau_2)$ for the phase characteristics in Figs. 3(a) and 3(b) respectively, and

$$\mathrm{Ci}\,(x) = -\int_x^\infty \frac{\cos t}{t}\,dt$$

is the cosine integral.[14]

It can be seen from Fig. 4 in which the function (5) is shown for $\tau_2 > \tau_1$, and $\tau_2 < \tau_1$ ($m = 0{\cdot}7$, $a_0 = 2{\cdot}11$) that the time response is no longer symmetrical about the point for which $V_2(t) = 1/2$. What is more important,

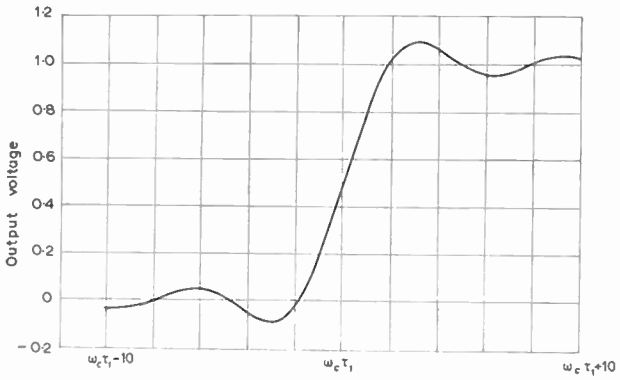


Fig. 2. The step response of an ideal low-pass filter.

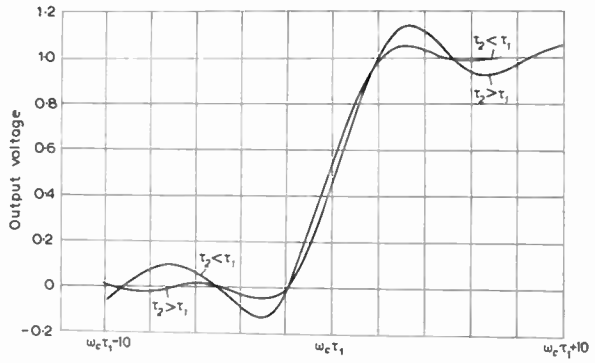


Fig. 4. The step responses of ideal low-pass filters with modified phase characteristics.

however, is the fact that the values of the precursor and overshoot depend on the shape of the delay response near the end of the useful band. Increasing the peak in the group delay response near $\omega = \omega_c$ (i.e. $\tau_2 > \tau_1$), decreases the precursor and increases the overshoot. The reverse is true in the case $\tau_2 < \tau_1$: the precursor is increased and the overshoot decreased when compared with the corresponding values obtained for the constant delay characteristic throughout the useful band (Fig. 2).

2.3 *Approximating Polynomials*

It follows from the above analysis of ideal filters that the approximating polynomials in the numerator and the denominator of (2) should preferably be of the type that provides an essentially constant delay over a frequency range occupying the largest part of the passband and with an adjustable delay peak near the ω_{3dB} frequency. Recently, these polynomials have been derived by one of the present authors[15], by applying the frequency transformation

$$p = \lambda_n z - \frac{1}{\lambda_n z} \qquad \ldots\ldots(6)$$

on the zeros of an auxiliary polynomial $Q_n(z)$ (see Appendix 7.2):

$$Q_n(z) = \sum_{k=0}^{n} A_k z^k = \xi_n B_{n-1}(z) + z^2 B_{n-2}(z) \qquad \ldots\ldots(7)$$

where λ_n is a constant depending on the order of the polynomial, $B_{n-1}(z)$ and $B_{n-2}(z)$ are Bessel polynomials of order $n-1$ and $n-2$ respectively and ξ_n is a variable parameter by which the peak in the delay characteristic can be changed. For the coefficients A_k of the auxiliary polynomial $Q_n(z)$ the following compact formula can easily be derived:

$$A_k = \frac{(2n-k)!}{2^{n-k}k!(n-k)!} - (2n-1-\xi_n)\frac{(2n-2-k)!}{2^{n-1-k}k!(n-1-k)!} \qquad \ldots\ldots(8)$$

For any particular n, the variable parameter ξ_n lies in the range $2 \leqslant \xi_n \leqslant 2n+1$. The lower limit $\xi_n = 2$ corresponds to a flat magnitude response in the auxiliary z-plane. Another particular value of ξ_n is $\xi_n = 2n-1$ for which the so-called quasi-Chebyshev polynomials in the p-plane are obtained. The former have also been discussed by Golay[16] and the latter by Cartianu and Constantin.[17, 18] It has been shown[15] that decreasing ξ_n increases the peak in the group delay characteristic. This is illustrated in Fig. 5 in which the normalized

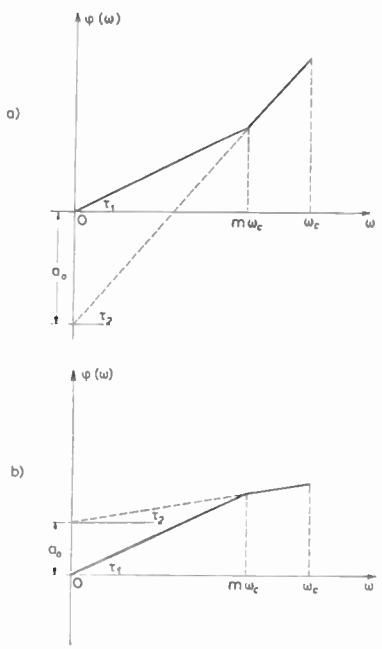


Fig. 3. Modified phase characteristics of an ideal low-pass filter.

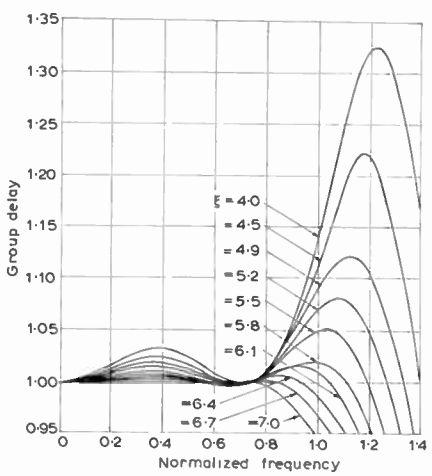


Fig. 5. The normalized group delay characteristics of $N_4(\xi_4, p)$.

group delay characteristics in the p-plane for the fourth order polynomials and some selected values of ξ_4 are presented.

2.4. *Design Procedures*

When the type of the polynomials in the numerator and the denominator of the transfer function (2) is selected, we adopt the following procedure to compute the values of the variable parameters ξ_{n-1}, ξ_n and k corresponding to any prescribed values of the precursor and the overshoot.

As discussed in Reference 15, the constant λ_n depends on the order of the polynomial and, for any particular n, λ_n is determined in such a way that the polynomial $Q_n(z)$ with $\xi_n = 2n-1$, when transformed in the original p-plane, yields a delay distortion of less than 0·2% in the useful band. For $n = 3$–7 the following values of λ have been found[15]: $\lambda_3 = 1{\cdot}4$, $\lambda_4 = 0{\cdot}8$, $\lambda_5 = 0{\cdot}6$, $\lambda_6 = 0{\cdot}4$, $\lambda_7 = 0{\cdot}3$.

With these values in hand, we now turn back to the synthesis problem and choose the initial values for k and ξ. The parameter k lies in the range $0 \leqslant k \leqslant 1$, with the lower limit $k = 0$ corresponding to the transfer function without finite zeros. Hence, for $k = 0$ there is no initial ringing in the transient response and the precursor increases with increasing k as the transmission zeros move from infinity towards the origin along the radial lines. Since in most practical cases the parameter k is in the range $0{\cdot}5 < k < 1$, it is advisable to start with $k = 1$ and then to decrease k when adjusting the value of the precursor. The upper bounds $\xi_{n-1} = 2n-1$, $\xi_n = 2n+1$ can also be chosen as the initial values for the parameters ξ_{n-1} and ξ_n in the numerator and the denominator polynomials of the transfer function. It has been found, however, that ξ_{n-1} is always smaller than $\xi_{n-1} = 2n-3$, i.e. the value of ξ_{n-1} corresponding to a Bessel polynomial in the auxiliary z-plane. So, in order to save the computational time we may choose $\xi_n = 2n+1$, $\xi_{n-1} = 2n-3$.

Once the initial values of the variable parameters for any particular n are known, the computation can start either by adjusting first the value of the precursor or the value of the overshoot according to the prescribed specifications. To this end a simple computer program has been written enabling the evaluation of the 10–90% rise-time in the step response T_r, the 50% delay-time T_d, the ratio of the delay time to the rise-time T_d/T_r, the $(n-1)$ peak values of the initial transient ringing or precursors $a\%$, the maximum overshoot $p\%$ and the normalized 3 dB amplitude bandwidth ω_{3dB}. There are exactly $n-1$ peaks in the initial transient ringing (precursors) the first of which is a minimum for n even and a maximum for n odd, while the last peak is always a minimum. The ringing at the top of the step response decays fairly quickly so that only the value of the first overshoot should be evaluated in the optimization process.

The main steps of the design procedure are as follows:

(i) Using equations (6)–(8), compute the parameters of the step response for $\xi_n = 2n+1$, $\xi_{n-1} = 2n-3$, $k = 1$.

(ii) If the last peak (minimum) in the initial ringing $a_{n-1}\%$ is higher than the maximum prescribed value $a_{max}\%$ decrease the parameter k and repeat computation until $a_{n-1}\%$ is approximately equal to $a_{max}\%$.

(iii) Decrease ξ_{n-1}, while keeping the other variable parameter unchanged, in order to adjust the first peak a_1 in the initial ringing so that $a_1 \simeq a_{max}$. Decreasing ξ_{n-1} increases the first peak a_1 while decreasing all other peaks in the initial ringing and also the overshoot.

(iv) Since $a_{n-1}\%$ decreases with decreasing ξ_{n-1}, increase k and readjust ξ_{n-1} so as to obtain $a_1 \simeq a_{n-1} \simeq a_{max}$.

(v) Decreasing ξ_n increases the maximum overshoot $p\%$. Hence, if the overshoot is smaller than the minimum prescribed value $p_{max}\%$, decrease ξ_n until $p\%$ is approximately equal to $p_{max}\%$.

(vi) Adjust the parameters in fine steps so as to obtain the results with required degree of accuracy.

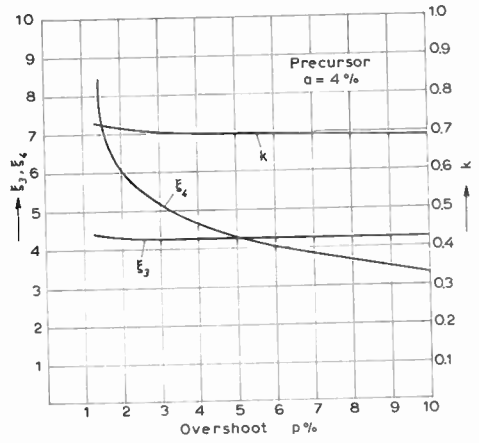


Fig. 6. Variation of ξ_3, ξ_4 and k with a percent overshoot for constant maximum value of the precursor ($a = 4\%$).

The final adjustment in step (vi) is greatly facilitated by the fact that the overshoot for a prescribed maximum value of precursor is almost entirely controlled by ξ_n. This can be seen from Fig. 6 in which ξ_3, ξ_4 and k as functions of the percent overshoot for $F_{3,4}(p)$ (the transfer function with 4 poles and 3 r.h.p. zeros) and the constant value of the precursor $a = 4\%$ are presented. On the other side, if the percent overshoot is kept constant, the maximum value of the precursor depends mainly on the value of the parameter k as shown in Fig. 7 in which ξ_3, ξ_4 and k as functions of the maximum value of the precursor are plotted for $F_{3,4}(p)$ and $p = 4\%$. It can be concluded that the maximum value of the precursor is determined almost entirely by the zeros and the overshoot by the poles of the transfer function.

From the foregoing description we observe that only the first and the last peak in the initial ringing are required in the optimization process. This fact may be used advantageously in reducing the computational time

when designing higher-order transfer functions since the evaluation of all $(n-1)$ peaks in each step of the optimization process would lead to an unnecessary increase of the total computational work.

An even more rapid design can be achieved at the expense of a very small reduction in the maximum ratio of the delay-time to the rise-time T_d/T_r of the resulting network by adjusting only the last peak of the precursor to the prescribed value. In this case a program for automatic computation can easily be written since steps (iii) and (iv) are completely eliminated and only k and ξ_n are variable parameters. Of course, all other peaks in the initial transient ringing must be smaller than the last one and this is automatically fulfilled if ξ_{n-1} is given its highest value $\xi_{n-1} = 2n-3$. This point will be illustrated by the following example. Suppose a fourth-order transfer function with three r.h.p. zeros $F_{3,4}(p)$ is required with equal maximum values of the overshoot and the precursors $p = a = 5\%$. Using the described procedure, after 20 iterations we obtain $\xi_3 = 4{\cdot}17$, $\xi_4 = 4{\cdot}75$ and $k = 0{\cdot}74$ with the following parameters of the step response: the overshoot $p = 5{\cdot}02\%$, the $(n-1)$ precursors $a_1 = -5{\cdot}01\%$, $a_2 = 4{\cdot}67\%$, $a_3 = -5{\cdot}02\%$, the delay/rise-time ratio $T_d/T_r = 2{\cdot}256$. On the other side, with $\xi_3 = 2n-3 = 5$, we found after 11 iterations $\xi_4 = 4{\cdot}365$, $k = 0{\cdot}655$ and the following values of the step response parameters: the overshoot $p = 5{\cdot}02\%$, the $(n-1)$ precursors $a_1 = -3{\cdot}02\%$, $a_2 = 4{\cdot}07\%$, $a_3 = -5{\cdot}04\%$. The delay/rise-time ratio in this case is $T_d/T_r = 2{\cdot}191$ which is to be compared with $T_d/T_r = 2{\cdot}256$ obtained when all variable parameters ξ_3, ξ_4 and k are optimized. The execution time of each iteration was approximately 2 minutes on the IBM 1130 computer but no attempt was made to reduce the computation time to a minimum.

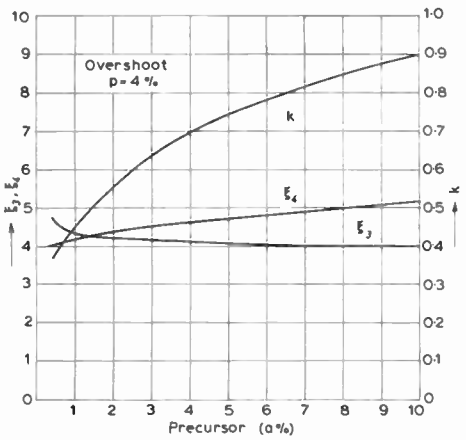


Fig. 7. Variation of ξ_3, ξ_4 and k with the percent precursor for constant value of the overshoot ($p = 4\%$).

Finally, it is of interest to compare the results obtained by the procedure proposed with prediction based on theoretical considerations in Sect. 2.2. For this purpose four transfer functions $F_{3,4}(p)$ with different values of k and ξ were chosen each having exactly the same value of the normalized magnitude bandwidth, $\omega_{3dB} = 1{\cdot}31$. The group delay characteristics and transient responses

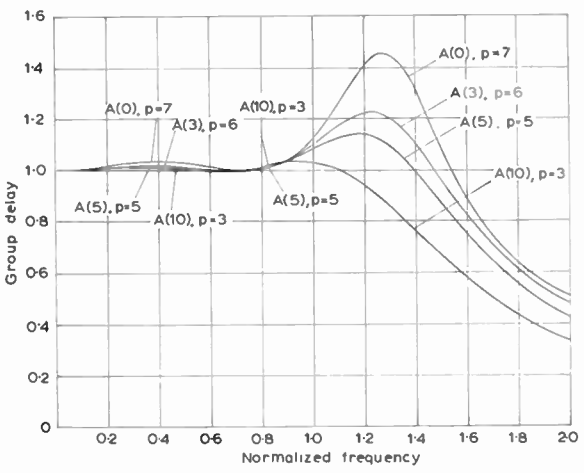


Fig. 8. Normalized delay characteristics of $F_{3,4}(p)$ for different values of the precursor and overshoot and equal ω_{3dB} bandwidth ($\omega_{3dB} = 1{\cdot}31$).

to a unit step input of these functions are shown in Figs. 8 and 9. It can be noticed that the larger the delay peak at the edge of the passband, the smaller maximum value of the precursor in the transient response is obtained at the expense of an increased overshoot. This is in good agreement with the results obtained in Sect. 2.2.

On the other side, the analysis of ideal filters predicts the constant delay throughout the passband in the case of equal maximum values of the precursor and overshoot, while the delay response of the function $F_{3,4}(p)$ with equal maximum values of the precursor and overshoot exhibits a delay peak of approximately 14%. In the latter case, however, the transient response is not symmetrical as in the case of the ideal filter (Fig. 2), since the initial transient ringing has three almost equal maxima and minima (5·01%, 4·7%, 5·02%) while the ringing at the top of the pulse decays quickly. We

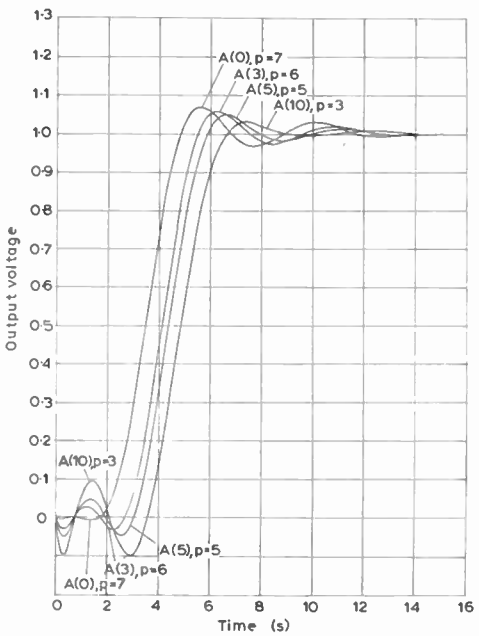


Fig. 9. The step responses of four different functions $F_{3,4}(p)$ with equal ω_{3dB} bandwidth and group delay characteristics shown in Fig. 8.

may conclude that in practical filters there should be a delay peak at the end of the passband even for equal specified values of overshoot and precursor. This gives further evidence in support of the results obtained by Jess and Schuessler[19, 20] and most recently by Crousel and Neirynck[21] and by Ariga and Sato[22] according to which the common statement that a good transient is always linked with a constant delay may not be well grounded. In any case, as far as the initial transient ringing is concerned this statement leads to what must be considered mistaken conclusions.

3. Numerical Tables and Comparison of Results

In order to facilitate the practical design of these filters, a fairly complete set of tables for the fourth-order function with three r.h.p. zeros $F_{3,4}(p)$ is given in the Appendix (Tables 2–9). Each Table provides the zero and pole locations for a constant maximum value of the precursor ($a\%$) and different values of overshoot ($p\%$). The corresponding values of the normalized rise-time (T_r), delay-time (T_d) and delay/rise-time ratio (T_d/T_r) are also included in these tables together with the normalized ω_{3dB} bandwidth of the filter. Hence, from these Tables the transfer function corresponding to almost any particular specifications in respect of the precursor and overshoot can be directly obtained.

Using the procedures described in Sect. 2.4 similar numerical tables can be compiled for direct evaluations of the parameters of higher-order transfer functions. As an example, Table 10 in the Appendix presents the data for the sixth-order transfer functions with five r.h.p. zeros $F_{5,6}(p)$ for equal values of the precursor and overshoot ($a\% = p\%$).

Apart from the fact that the present method allows much greater flexibility in specifying the tolerable values of the precursor and overshoot, a comparison of results also shows that it provides better transient performance of the resulting network than any other method so far described. To demonstrate this, the most important parameters associated with the unit step response for the functions $CK_{3,4}$, $CR_{3,4}$ (Reference 5†), and for the function $R_{3,4}$ and $K_{3,4}$ (Reference 4) are presented in Table 1 together with corresponding figures for some functions selected from Tables 3 and 4.

Table 1. Comparison of the fourth-order networks

Function description	Precursor $a\%$	Overshoot $p\%$	T_d/T_r
$C - K_{3,4}$	1·1	2·9	1·47
$C - R_{3,4}$	1·6	3·0	1·59
$K_{3,4}$	4·1	3·3	1·42
$R_{3,4}$	4·3	1·8	1·61
$F_{3,4}$	0·5	3·0	1·56
$F_{3,4}$	2·0	2·0	1·72
$F_{3,4}$	2·0	3·0	1·82
$F_{3,4}$	3·0	3·0	1·93

† Some mistakes have been noticed in Reference 5, Table 2, the most serious of which are the value of the precursor for $CR_{3,4}$ given as 1·06% instead of 1·6% and the value of the overshoot for $CP_{3,4}$ given as 2·19% instead of 4·7%.

Various techniques for practical realization of the transfer functions described are available using two or more operational amplifiers. They are well covered in recent Deliyannis papers together with the examples[5, 23, 24] and will not be repeated here. Of course, these transfer functions can be realized by a variety of other synthesis techniques that can be found in the literature.[25] Practical aspects like number of amplifiers and other components, size and sensitivity to variations in RC component values may determine which technique is most suitable. However, these topics are beyond the scope of this paper.

4. Conclusions

The problem of simulating the various systems with inherent transport delay on analogue computer has emphasized the need for improved time delay networks. For band-limited applications Chebyshev all-pass delay approximations of the ideal delay function $e^{-\tau s}$ has been proved to be the most efficient and the complete data for determining transfer functions of these networks have recently been published.[3] However, because of large initial transient ringing the all-pass Chebyshev approximants are useless in those applications where the input signal spectrum occupies very large bandwidth. Although some useful techniques for improving the transient response of delay networks have been published, this problem has not yet found a satisfactory solution. None of the existing methods, for example, provides possibilities for the circuit designer to choose freely the specified values for the precursor and overshoot according to the problem at hand.

A new approximation technique has been developed in this paper to deal exactly with the situation where the maximum tolerable values of the precursor and overshoot are prescribed beforehand. The method is based on the application of a recently introduced class of polynomials with two variable parameters that yield a quasi-Chebyshev type of delay response with a variable delay peak near the end of the passband. Opposite to the common belief, this type of delay response has been found to provide better transient response of the filter than in the case of constant delay approximations throughout the passband especially from the point of view of the synthesis of delay networks with smaller values of the precursor.

Fairly complete data enabling direct determination of the fourth-order non-minimum phase transfer function with three right-half-plane zeros for almost any practical values of the precursor and overshoot have been tabulated. The important feature, however, is that this method is quite general and can be easily applied to higher order networks since the polynomials used in approximation are defined in closed analytical form in the transformed frequency plane. As an example, Table 10 has been compiled giving all necessary data for the construction of the sixth-order transfer function with five right-half-plane zeros for equal values of the precursor and overshoot.

Finally, in order to examine the efficiency of the present technique a comparison of some most important parameters associated with the transient response to a unit step excitation for delay functions designed by various known methods has been made. It confirms that the transient responses obtained by the method described show an improvement over the rest.

5. Acknowledgments

The authors wish to acknowledge the Research Fund of S. R. Serbia for financial support of work of which that described forms a part. All numerical computations were carried out on the computer of the Faculty of Electronic Engineering of the University of Nish.

6. References

1. Morill, C. D., 'A sub-audio time delay circuit', *Trans. Inst. Radio Engrs on Electronic Computers*, **EC-3**, pp. 45–49, June 1954.
2. Hepner, C. F., 'Improved methods of simulating time delays', *Trans. I.E.E.E. on Electronic Computers*, **EC-14**, pp. 239–43, April 1965.
3. Hausner, A. and Furlani, C. M., 'Chebyshev all-pass approximants for time-delay simulation', *I.E.E.E. Trans. on Electronic Computers*, **EC-15**, pp. 314–21, June 1966.
4. King, W. J. and Rideout, V. S., 'Improved transport delay circuits for analog computer use', Proc. of the Third International Analogue Computation Meetings, Opatija, Yugoslavia, September 1961, pp. 560–7.
5. Deliyannis, T., 'Six new delay functions and their realization using active RC networks', *The Radio and Electronic Engineer*, **39**, pp. 139–44, March 1970.
6. Budak, A., 'A maximally flat phase and controllable magnitude approximation', *I.E.E.E. Trans. on Circuit Theory*, **CT-12**, p. 279, June 1965.
7. Allemandou, P., 'Low-pass filters-approximating in modulus and phase—the exponential function', *Trans. I.E.E.E. on Circuit Theory*, **CT-13**, pp. 218–301, September 1966.
8. Pongrarit, V. and Park, S. B., 'Rational delay functions based on continued-fraction expansion of e^{z}', *Electronics Letters*, **6**, No. 20, pp. 656–8, 1st October 1970.
9. Perron, O., 'Die Lehre von den Kettenbrücken', Band II (B. G. Teubner Verlagsgesellshaft, Stuttgart, 1957).
10. Wall, H. S., 'Analytic Theory of Continued Fractions' (Van Nostrand, New York, 1948).
11. Valley, G. E. and Wallman, H., 'Vacuum Tube Amplifiers', Appendix A, 'Realizability of filters' (McGraw-Hill, New York, 1948).
12. Guillemin, E. A., 'Theory of Linear Physical Systems' (Wiley, New York, 1963).
13. Doborovolski, G. V., 'Transmission of Pulses Through Communication Channels', (Gosudarstvenoe Izdatelstvo Literaturi, Moskva 1960) (In Russian).
14. Abramovitz, M. and Stegun, I. A., 'Handbook of Mathematical Functions (Dover Publications, New York, 1965).
15. Rakovich, B. D., 'Transfer functions approximating to a constant group delay—Parts 1 and 2', *Electronic Engng*, **40**, pp. 242–6, May 1968; **40**, pp. 326–8, June 1968.
16. Golay, M. J. E., 'Polynomials of transfer functions with poles only satisfying conditions at the origin', *Trans. Inst. Radio Engrs on Circuit Theory*, **CT-7**, pp. 224–9, September 1960.
17. Cartianu, G. and Constantin, I., 'Transfer functions obtained by a transform derived from the Darlington transform', *Electronics Letters*, **4**, pp. 327–8, 9th August 1968.
18. Cartianu, G. and Constantin, I., 'Transfer functions with quasi-Chebyshev type characteristics obtained by a Darlington-derived transformation', *Electronics Letters*, **4**, pp. 328–31, 9th August 1968.
19. Jess, J. and Schuessler, H. W., 'A class of pulse forming networks', *I.E.E.E. Trans. on Circuit Theory*, **CT-12**, pp. 296–9, June 1965.
20. Jess, J. and Schuessler, H. W., 'On the design of pulse-forming networks', *I.E.E.E. Trans. on Circuit Theory*, **CT-12**, pp. 394–400, September 1965.
21. Crousel, L. and Neirynk, J. J., 'Polynomial Chebyshev approximations of the ideal filters', *I.E.E.E. Trans. on Circuit Theory*, **CT-15**, pp. 307–15, December 1968.
22. Ariga, M. and Sato, M., 'An extremum approach to constant-delay transfer functions providing large amplitude bandwidth', *I.E.E.E. Trans. on Circuit Theory*, **CT-17**, pp. 121–5, February 1970.
23. Deliyannis, T., 'RC active all-pass sections', *Electronics Letters*, **5**, pp. 59–60, 6th February 1969.
24. Bedri, Y. and Deliyannis, T., 'Realization of a quadratic with a positive real zero', *The Radio and Electronic Engineer*, **39**, pp. 271–2, May 1970.
25. Mitra, S. K., 'Analysis and Synthesis of Linear Active Networks' (Wiley, New York, 1969).

7. Appendices

7.1. *Derivation of Equation (5)*

Let $F(j\omega) = R(\omega)+jX(\omega) = A(\omega)\,e^{-j\psi(\omega)}$ be the transfer function of the network and

$$F(\omega) = \lim_{\alpha\to 0} \frac{j}{\sqrt{2\pi}} \frac{\omega}{\omega^2+\alpha^2} = \frac{j}{\sqrt{2\pi}} \frac{1}{\omega} \qquad \text{.....(9)}$$

the frequency spectrum of the unit step excitation. Then, using the inversion Fourier integral, we obtain the output voltage as a function of time:

$$V_2(t) = \tfrac{1}{2}R(0) + \frac{1}{\pi}\int_0^\infty \frac{R(\omega)}{\omega} \sin \omega t\, d\omega + \\ + \frac{1}{\pi}\int_0^\infty \frac{X(\omega)}{\omega} \cos \omega t\, d\omega \qquad \text{.....(10)}$$

Since $V_2(-t) = 0$ we have also

$$V_2(-t) = \tfrac{1}{2}R(0) - \frac{1}{\pi}\int_0^\infty \frac{R(\omega)}{\omega} \sin \omega t\, d\omega + \\ + \frac{1}{\pi}\int_0^\infty \frac{X(\omega)}{\omega} \cos \omega t\, d\omega = 0 \qquad \text{.....(11)}$$

so that $V_2(t)$ can be written in the following form:

$$V_2(t) = V_2(t)-V_2(-t) = \frac{2}{\pi}\int_0^\infty \frac{R(\omega)}{\omega} \sin \omega t\, d\omega \qquad \text{.....(12)}$$

Now, for the ideal magnitude response and the phase characteristic shown in Fig. 3(a) or 3(b) we have

$$V_2(t) = \frac{2}{\pi}\int_0^{m\omega_c} \frac{\sin \omega t \cos \tau_1 \omega}{\omega} d\omega + \\ + \frac{2}{\pi}\int_{m\omega_c}^{\omega_c} \frac{\sin \omega t \cos (a_0+\tau_2 \omega)}{\omega} d\omega \qquad \text{.....(13)}$$

The first integral on the right-hand side corresponds to the ideal filter (Fig. 1) having the cut-off frequency $m\omega_c$, and hence

$$\frac{2}{\pi}\int_0^{m\omega_c} \frac{\sin \omega t \cos \tau_1 \omega}{\omega} d\omega = \tfrac{1}{2}+\mathrm{Si}\,[m\omega_c(t-\tau_1)] \qquad \text{.....(14)}$$

Table 2. $F_{3,4}$ $a = 0{\cdot}5\%$

$p\%$	T_d	T_r	T_d/T_r	ω_{3dB}	$\sigma_{1,0}$	$\omega_{1,0}$	$\sigma_{2,0}$	$\omega_{2,0}$	$\sigma_{1,p}$	$\omega_{1,p}$	$\sigma_{2,p}$	$\omega_{2,p}$
1·0	4·022	3·118	1·290	0·72	2·1574	0	1·7568	2·1564	−0·9707	0·0203	−0·5924	0·8426
2·0	3·699	2·481	1·491	0·93	2·2082	0	1·7982	2·2072	−0·6765	0·3940	−0·4815	1·1831
3·0	3·636	2·336	1·556	1·02	2·2140	0	1·8029	2·2130	−0·6539	0·4069	−0·4271	1·2262
4·0	3·589	2·229	1·610	1·12	2·2215	0	1·8090	2·2205	−0·6397	0·4165	−0·3849	1·2522
5·0	3·559	2·155	1·652	1·20	2·2215	0	1·8090	2·2205	−0·6307	0·4234	−0·3553	1·2674
6·0	3·532	2·090	1·690	1·27	2·2215	0	1·8090	2·2205	−0·6234	0·4298	−0·3290	1·2790
7·0	3·505	2·036	1·721	1·31	2·2333	0	1·8186	2·2323	−0·6172	0·4356	−0·3058	1·2879
10·0	3·453	1·911	1·806	1·41	2·2333	0	1·8186	2·2323	−0·6035	0·4502	−0·2525	1·3039

Table 3. $F_{3,4}$ $a = 1\%$

$p\%$	T_d	T_r	T_d/T_r	ω_{3dB}	$\sigma_{1,0}$	$\omega_{1,0}$	$\sigma_{2,0}$	$\omega_{2,0}$	$\sigma_{1,p}$	$\omega_{1,p}$	$\sigma_{2,p}$	$\omega_{2,p}$
1·1	4·227	3·055	1·384	0·74	1·6468	0	1·3302	1·8221	−0·9707	0·0203	−0·5924	0·8426
2·0	3·916	2·465	1·589	0·94	1·6961	0	1·3700	1·8766	−0·6863	0·3890	−0·5008	1·1645
3·0	3·850	2·301	1·673	1·04	1·6961	0	1·3700	1·8766	−0·6592	0·4036	−0·4412	1·2162
4·0	3·804	2·185	1·741	1·14	1·6969	0	1·3710	1·8751	−0·6436	0·4137	−0·3972	1·2452
5·0	3·769	2·112	1·784	1·22	1·7103	0	1·3828	1·8824	−0·6344	0·4205	−0·3676	1·2614
6·0	3·743	2·048	1·828	1·28	1·7103	0	1·3828	1·8824	−0·6269	0·4267	−0·3417	1·2736
7·0	3·722	1·993	1·867	1·32	1·7121	0	1·3850	1·8794	−0·6208	0·4322	−0·3194	1·2829
10·0	3·667	1·864	1·967	1·42	1·7285	0	1·4006	1·8769	−0·6065	0·4468	−0·2644	1·3001

Table 4. $F_{3,4}$ $a = 2\%$

$p\%$	T_d	T_r	T_d/T_r	ω_{3dB}	$\sigma_{1,0}$	$\omega_{1,0}$	$\sigma_{2,0}$	$\omega_{2,0}$	$\sigma_{1,p}$	$\omega_{1,p}$	$\sigma_{2,p}$	$\omega_{2,p}$
2·0	4·172	2·422	1·722	0·96	1·3901	0	1·1217	1·5462	−0·6988	0·3829	−0·5221	1·1411
3·0	4·088	2·240	1·825	1·07	1·3983	0	1·1277	1·5594	−0·6658	0·3998	−0·4577	1·2035
4·0	4·040	2·135	1·893	1·16	1·4056	0	1·1336	1·5675	−0·6505	0·4091	−0·4175	1·2326
5·0	4·006	2·050	1·954	1·24	1·4100	0	1·1377	1·5683	−0·6397	0·4165	−0·3849	1·2522
6·0	3·980	1·989	2·001	1·29	1·4117	0	1·1390	1·5710	−0·6323	0·4222	−0·3604	1·2649
7·0	3·952	1·932	2·046	1·34	1·4204	0	1·1465	1·5770	−0·6254	0·4280	−0·3364	1·2759
8·0	3·931	1·886	2·084	1·38	1·4298	0	1·1555	1·5778	−0·6202	0·4327	−0·3172	1·2837
10·0	3·896	1·806	2·157	1·43	1·4338	0	1·1587	1·5823	−0·6111	0·4418	−0·2824	1·2957

Table 5. $F_{3,4}$ $a = 3\%$

$p\%$	T_d	T_r	T_d/T_r	ω_{3dB}	$\sigma_{1,0}$	$\omega_{1,0}$	$\sigma_{2,0}$	$\omega_{2,0}$	$\sigma_{1,p}$	$\omega_{1,p}$	$\sigma_{2,p}$	$\omega_{2,p}$
2·0	4·345	2·408	1·805	0·96	1·2347	0	0·9934	1·3912	−0·7127	0·3764	−0·5418	1·1162
3·0	4·251	2·205	1·928	1·09	1·2418	0	0·9979	1·4063	−0·6724	0·3961	−0·4728	1·1909
4·0	4·199	2·092	2·007	1·19	1·2495	0	1·0041	1·4150	−0·6553	0·4060	−0·4309	1·2236
5·0	4·163	2·010	2·071	1·26	1·2528	0	1·0064	1·4205	−0·6443	0·4132	−0·3994	1·2439
6·0	4·134	1·944	2·127	1·31	1·2571	0	1·0100	1·4243	−0·6361	0·4192	−0·3733	1·2585
7·0	4·109	1·889	2·176	1·36	1·2629	0	1·0153	1·4277	−0·6294	0·4246	−0·3506	1·2696
8·0	4·089	1·838	2·224	1·40	1·2649	0	1·0171	1·4288	−0·6237	0·4295	−0·3300	1·2786
10·0	4·053	1·759	2·304	1·45	1·2711	0	1·0224	1·4340	−0·6145	0·4383	−0·2956	1·2915

Table 6. $F_{3,4}$ $a = 4\%$

$p\%$	T_d	T_r	T_d/T_r	ω_{3dB}	$\sigma_{1,0}$	$\omega_{1,0}$	$\sigma_{2,0}$	$\omega_{2,0}$	$\sigma_{1,p}$	$\omega_{1,p}$	$\sigma_{2,p}$	$\omega_{2,p}$
2·0	4·470	2·368	1·888	0·99	1·1329	0	0·9088	1·2910	−0·7182	0·3738	−0·5486	1·1065
3·0	4·378	2·180	2·008	1·11	1·1429	0	0·9159	1·3072	−0·6790	0·3927	−0·4866	1·1784
4·0	4·320	2·056	2·102	1·21	1·1493	0	0·9205	1·3175	−0·6592	0·4036	−0·4412	1·2162
5·0	4·286	1·975	2·170	1·28	1·1507	0	0·9209	1·3227	−0·6481	0·4106	−0·4107	1·2370
6·0	4·253	1·905	2·233	1·34	1·1561	0	0·9252	1·3288	−0·6391	0·4170	−0·3829	1·2533
7·0	4·230	1·848	2·289	1·38	1·1574	0	0·9260	1·3313	−0·6323	0·4222	−0·3624	1·2649
8·0	4·209	1·803	2·334	1·41	1·1610	0	0·9290	1·3348	−0·6269	0·4267	−0·3417	1·2736
10·0	4·172	1·721	2·424	1·47	1·1686	0	0·9359	1·3396	−0·6174	0·4355	−0·3065	1·2877

Table 7. $F_{3,4}$ $a = 5\%$

$p\%$	T_d	T_r	T_d/T_r	ω_{3dB}	$\sigma_{1,0}$	$\omega_{1,0}$	$\sigma_{2,0}$	$\omega_{2,0}$	$\sigma_{1,p}$	$\omega_{1,p}$	$\sigma_{2,p}$	$\omega_{2,p}$
2·0	4·589	2·364	1·941	1·00	1·0608	0	0·8501	1·2133	−0·7319	0·3674	−0·5632	1·0835
3·0	4·476	2·148	2·083	1·14	1·0752	0	0·8604	1·2358	−0·6828	0·3908	−0·4942	1·1711
4·0	4·420	2·026	2·182	1·24	1·0765	0	0·8596	1·2464	−0·6628	0·4015	−0·4504	1·2092
5·0	4·384	1·943	2·256	1·31	1·0774	0	0·8590	1·2535	−0·6511	0·4086	−0·4194	1·2314
6·0	4·354	1·874	2·323	1·36	1·0797	0	0·8603	1·2582	−0·6422	0·4147	−0·3929	1·2477
7·0	4·329	1·816	2·383	1·41	1·0823	0	0·8622	1·2622	−0·6350	0·4200	−0·3697	1·2603
9·0	4·286	1·727	2·482	1·47	1·0903	0	0·8690	1·2697	−0·6243	0·4290	−0·3322	1·2777
10·0	4·268	1·687	2·529	1·49	1·0937	0	0·8720	1·2725	−0·6196	0·4333	−0·3151	1·2845

Table 8. $F_{3,4}$ $a = 7\%$

$p\%$	T_d	T_r	T_d/T_r	ω_{3dB}	$\sigma_{1,0}$	$\omega_{1,0}$	$\sigma_{2,0}$	$\omega_{2,0}$	$\sigma_{1,p}$	$\omega_{1,p}$	$\sigma_{2,p}$	$\omega_{2,p}$
2·0	4·775	2·336	2·044	1·04	0·9589	0	0·7668	1·1049	−0·7519	0·3576	−0·5798	1·0519
3·0	4·649	2·103	2·211	1·20	0·9659	0	0·7682	1·1317	−0·6917	0·3864	−0·5105	1·1543
4·0	4·590	1·982	2·316	1·30	0·9696	0	0·7691	1·1440	−0·6705	0·3972	−0·4684	1·1946
5·0	4·549	1·895	2·401	1·37	0·9721	0	0·7697	1·1522	−0·6574	0·4047	−0·4365	1·2196
6·0	4·515	1·825	2·474	1·42	0·9757	0	0·7718	1·1591	−0·6478	0·4108	−0·4097	1·2376
7·0	4·488	1·764	2·544	1·47	0·9776	0	0·7727	1·1641	−0·6400	0·4163	−0·3859	1·2516
8·0	4·467	1·717	2·602	1·50	0·9799	0	0·7741	1·1687	−0·6341	0·4208	−0·3666	1·2619
10·0	4·429	1·635	2·708	1·55	0·9850	0	0·7777	1·1755	−0·6243	0·4290	−0·3322	1·2777

Table 9. $F_{3,4}$ $a = 10\%$

$p\%$	T_d	T_r	T_d/T_r	ω_{3dB}	$\sigma_{1,0}$	$\omega_{1,0}$	$\sigma_{2,0}$	$\omega_{2,0}$	$\sigma_{1,p}$	$\omega_{1,p}$	$\sigma_{2,p}$	$\omega_{2,p}$
2·0	5·026	2·347	2·142	1·10	0·8578	0	0·6857	0·9896	−0·8015	0·3291	−0·6023	0·9845
3·0	4·848	2·063	2·350	1·31	0·8728	0	0·6929	1·0274	−0·7064	0·3794	−0·5333	1·1275
4·0	4·774	1·926	2·479	1·43	0·8768	0	0·6930	1·0440	−0·6794	0·3925	−0·4875	1·1776
5·0	4·733	1·833	2·582	1·50	0·8744	0	0·6878	1·0530	−0·6647	0·4004	−0·4550	1·2057
6·0	4·692	1·760	2·666	1·55	0·8818	0	0·6936	1·0619	−0·6539	0·4069	−0·4271	1·2262
7·0	4·670	1·701	2·745	1·59	0·8766	0	0·6859	1·0672	−0·6461	0·4120	−0·4048	1·2406
10·0	4·602	1·571	2·930	1·65	0·8879	0	0·6948	1·0809	−0·6296	0·4244	−0·3511	1·2694
11·0	4·588	1·537	2·984	1·67	0·8890	0	0·6949	1·0846	−0·6254	0·4280	−0·3364	1·2759

Table 10. $F_{5,6}$ $a\% = p\%$

$p\%$	T_d	T_r	T_d/T_r	ω_{3dB}	$\sigma_{1,0} \pm j\omega_{1,0}$	$\sigma_{2,0} \pm j\omega_{2,0}$	$\sigma_{3,0} \pm j\omega_{3,0}$	$\sigma_{1,p} \pm j\omega_{1,p}$	$\sigma_{2,p} \pm j\omega_{2,p}$	$\sigma_{3,p} \pm j\omega_{3,p}$
2·0	6·710	2·298	2·920	1·00	0·8760 ± j0·8495	0·6253 ± j1·6875	0·9344 ± j0	−0·5686 ± j0·3036	−0·5343 ± j0·8895	−0·3890 ± j1·4438
3·0	6·775	2·060	3·289	1·16	0·8375 ± j0·8063	0·6009 ± j1·5999	0·8927 ± j0	−0·5419 ± j0·3124	−0·5062 ± j0·9140	−0·3383 ± j1·4921
4·0	6·850	1·905	3·597	1·37	0·8009 ± j0·7757	0·5723 ± j1·5405	0·8542 ± j0	−0·5305 ± j0·3167	−0·4940 ± j0·9255	−0·3107 ± j1·5112
5·0	6·937	1·790	3·875	1·55	0·7603 ± j0·7497	0·5352 ± j1·4927	0·8121 ± j0	−0·5247 ± j0·3190	−0·4878 ± j0·9317	−0·2954 ± j1·5202
6·0	7·006	1·695	4·134	1·65	0·7322 ± j0·7300	0·5100 ± j1·4559	0·7829 ± j0	−0·5200 ± j0·3209	−0·4828 ± j0·9368	−0·2827 ± j1·5269
7·0	7·072	1·616	4·376	1·74	0·7066 ± j0·7132	0·4855 ± j1·4246	0·7564 ± j0	−0·5169 ± j0·3222	−0·4795 ± j0·9403	−0·2739 ± j1·5312
8·0	7·117	1·543	4·611	1·81	0·6914 ± j0·7007	0·4728 ± j1·4004	0·7405 ± j0	−0·5135 ± j0·3237	−0·4759 ± j0·9442	−0·2642 ± j1·5355
10·0	7·196	1·426	5·046	1·94	0·6644 ± j0·6808	0·4512 ± j1·3620	0·7143 ± j0	−0·5085 ± j0·3259	−0·4705 ± j0·9501	−0·2496 ± j1·5415

The second integral

$$\frac{2}{\pi}\int_{m\omega_c}^{\omega_c}\frac{\sin\omega t\cos(a_0+\tau_2\omega)}{\omega}\,d\omega$$

$$=\frac{2}{\pi}\int_{m\omega_c}^{\omega_c}\frac{\sin\omega t}{\omega}(\cos a_0\cos\tau_2\omega-\sin a_0\sin\tau_2\omega)\,d\omega$$

$$=\frac{\cos a_0}{\pi}\left[\int_{m\omega_c}^{\omega_c}\frac{\sin\omega(t+\tau_2)}{\omega}d\omega+\int_{m\omega_c}^{\omega_c}\frac{\sin\omega(t-\tau_2)}{\omega}d\omega\right]-$$

$$-\frac{\sin a_0}{\pi}\left[\int_{m\omega_c}^{\omega_c}\frac{\cos\omega(t+\tau_2)}{\omega}d\omega-\int_{m\omega_c}^{\omega_c}\frac{\cos\omega(t-\tau_2)}{\omega}d\omega\right]$$

.....(15)

For higher values of the argument the integrals

$$\int_{m\omega_c}^{\omega_c}\frac{\sin\omega(t+\tau_2)}{\omega}d\omega \quad\text{and}\quad \int_{m\omega_c}^{\omega_c}\frac{\cos\omega(t+\tau_2)}{\omega}d\omega$$

tend to zero so that

$$\frac{2}{\pi}\int_{m\omega_c}^{\omega_c}\frac{\sin\omega t}{\omega}\cos(a_0+\tau_2\omega)\,d\omega$$

$$=\frac{\cos a_0}{\pi}[\mathrm{Si}\,\omega_c(t-\tau_2)-\mathrm{Si}\,m\omega_c(t-\tau_1)]+$$

$$+\frac{\sin a_0}{\pi}[\mathrm{Ci}\,\omega_c(t-\tau_2)-\mathrm{Ci}\,m\omega_c(t-\tau_1)] \quad .....(16)$$

Finally, substituting $\omega_c\tau_2=\omega_c\tau_1+a_0/m$ and adding (14) and (16) the equation (5) is obtained.

7.2. *The Auxiliary Polynomial $Q_n(Z)$*

The auxiliary polynomial $Q_n(Z)$ in the Z-plane is derived from the nth order Bessel polynomial

$$B_n(z)=\sum_{k=0}^{n}\frac{(2n-k)!\,z^k}{2^{n-k}k!(n-k)!}=\sum_{k=0}^{n}b_kz^k \quad .....(17)$$

by substituting a variable parameter ξ for $(2n-1)$ in the degree varying recurrence formula

$$B_n(z)=(2n-1)B_{n-1}(z)+z^2B_{n-2}(z) \quad .....(18)$$

Hence,

$$Q_n(z)=\xi B_{n-1}(z)+z^2B_{n-2}(z) \quad .....(19)$$

which can be put in the following form

$$Q_n(z)=\sum_{k=0}^{n}A_kz^k$$

$$=(2n-1)B_{n-1}(z)+z^2B_{n-2}(z)-(2n-1-\xi)B_{n-1}(z)$$

$$=B_n(z)-(2n-1-\xi)B_{n-2}(z)$$

By matching coefficients in (17) and (20) we get directly,

$$A_k=\frac{(2n-k)!}{2^{n-k}k!(n-k)!}-(2n-1-\xi)\frac{(2n-2-k)!}{2^{n-1-k}k!(n-1-k)!}$$

which is the equation (8).

7.3. *Numerical Tables of Rational Approximants*

Tables 2 to 9 refer to the fourth-order function $F_{3,4}(p)$ and each of them gives the normalized zero $(\sigma_{i,0}\pm j\omega_{i,0})$ and pole locations $(\sigma_{i,p}\pm j\omega_{i,p})$ for a constant maximum value of the precursor $(a\%)$ and for different values of the overshoot $(p\%)$. The normalized delay-time (T_d), rise-time (T_r) and ω_{3dB} bandwidth are also included. Table 10 refers to the sixth-order function $F_{5,6}(p)$ for equal values of the precursor and overshoot.

Manuscript first received by the Institution on 25th September 1970, in revised form on 3rd November 1970 and in final form on 26th March 1971. (Paper No. 1390/CC.102.)

U.D.C. 621.318.57

The Application of a Synchronous Switch for Educational Use

By

R. J. TODD,
M.Sc., C.Eng., M.I.E.E., M.I.E.R.E.†

A simple synchronous switch having a control circuit comprising integrated circuits and a triac switch is described. The synchronous switch is primarily designed to demonstrate the transient surge current taken by a transformer when its supply is applied but has many other applications.

1. Introduction

It is well known that the initial magnetizing current taken by a transformer depends not only on the instantaneous magnitude of the applied voltage but also on the magnetic state of the core of the transformer.[1] With the worst conditions the maximum magnitude of the current inrush approaches a value given by the peak value of the applied voltage divided by the resistance of the transformer winding.

In order to investigate these effects, it is necessary to set the transformer core to a known remanent magnetic state. If then the alternating supply voltage is applied with a predetermined instantaneous magnitude to the magnetizing winding of the transformer, the transient current may be observed using a cathode-ray oscilloscope.

2. Brief Specification of the Synchronous Switch

The synchronous switch is intended for use with 29V or 240V 50 Hz single-phase supplies. It may be preset to apply the supply voltage to the load during either a positive or a negative half-cycle of the supply voltage waveform. Further, the instant at which the voltage is applied may be preset to any time between 40μs and 10 ms of the selected half-period. After the initial half-cycle, the switch conducts both half-cycles of the waveform and the supply may be left connected, at will, until steady-state conditions are established.

When desired the synchronous switch disconnects the supply when the load current passes through zero after either a positive or a negative half-cycle, as preselected, of the supply voltage waveform. The application and disconnexion of the supply may be initiated manually. Alternatively, the preselected sequence may be set to repeat automatically.

When the load is a transformer, the steady-state magnetizing current swings the core flux around a given hysteresis loop. The removal of the magnetizing current at a known current zero then leaves the core with a given magnitude and polarity of remanent flux. In this way the magnetic state of the core can be preset before the synchronous switch is used to reconnect the supply to the transformer magnetizing winding.

3. Circuit Description

Figure 1 is a schematic diagram of the synchronous switch. The positive and negative supplies for the integrated circuits are 0V and −5V respectively. In the following description, logical '1' represents, for convenience, 0V but in practice any direct voltage which is more positive than −3V. Conversely, '0' represents −5V.

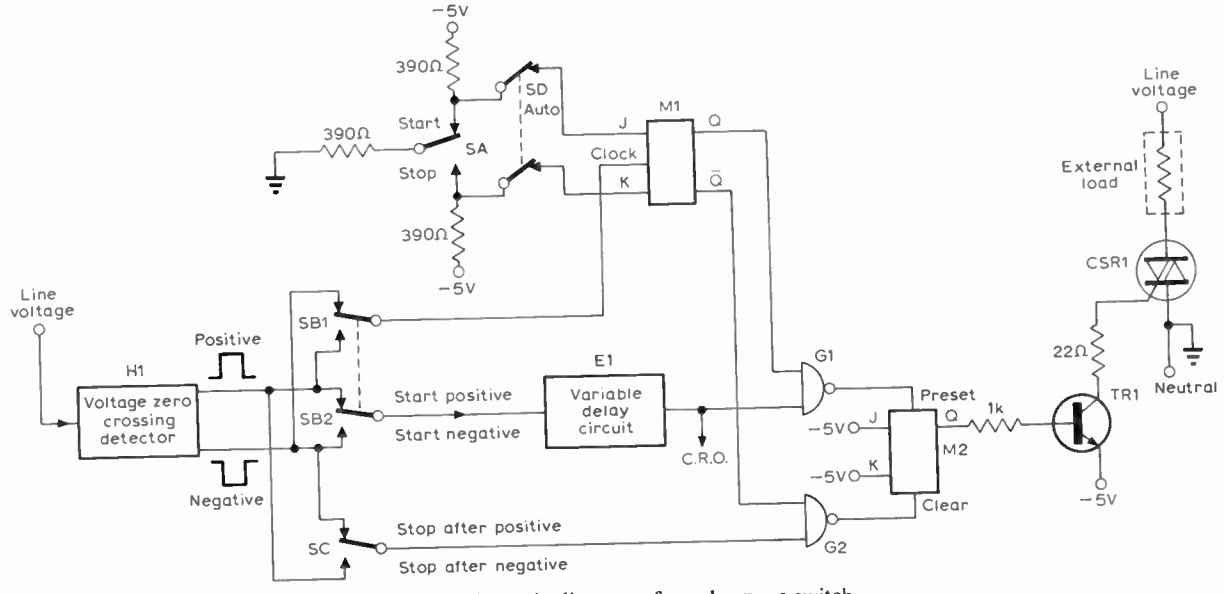


Fig. 1. Schematic diagram of synchronous switch.

† School of Electrical Engineering, South Australian Institute of Technology, Adelaide, South Australia, 5000.

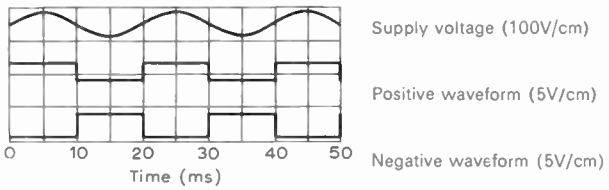


Fig. 2. Output waveform of the voltage zero crossing detector.

The supply voltage zero crossing detector H1 and the variable delay circuit E1 are described in more detail later. As shown by Fig. 2, the voltage zero crossing detector generates two rectangular waveforms, designated positive and negative, which are most positive during the positive and negative half-cycles of the supply waveform, respectively.

3.1. *Application of the Supply Voltage to the Load*

Switch SA is the manual control switch by which the operator initiates the connexion of the supply to the load. Assuming that switch SD is closed then when SA is switched to 'start' a logical '1' is applied to the J input of J–K flip-flop M1. Consequently, the Q output of M1 will change to '1' when its clock input next changes from '1' to '0'.

Switch SB is a preselection switch by which the operator may select to apply the mains voltage to the load during either a positive or a negative half-cycle of the supply voltage. The time in the chosen half-cycle at which the voltage is applied is determined by the variable delay circuit E1.

Assuming that a positive half-cycle is selected, the Q output of M1 will change to a '1' at the beginning of the positive half-cycle. The change is produced by the negative waveform applied to the clock input of M1 changing from '1' to '0'.

Flip-flop M1 then applies a '1' to NAND gate G1. However, at the instant that M1 changes state, the output of E1 falls to '0' (see Sect. 3.4) and inhibits G1 until the preselected delay period of E1 has expired. At the end of this period both inputs to G1 are '1' and G1 presets J–K flip-flop M2. This causes TR1 to conduct, which in turn triggers the triac CSR1; the triac then switches to the conducting state and completes the circuit of the external load under test.

In this way the instant at which CSR1 is triggered during the positive half-cycle of the supply voltage is determined by the time delay of E1. This time-delay is continuously variable between 40μs and 10 ms.

Conversely, switch SB may be used to select a negative half-cycle of the supply voltage. Flip-flop M1 then changes state at the beginning of the negative half-cycle and E1 inhibits G1 for a predetermined time as before.

To simplify the triggering circuit of the triac d.c. triggering is used. This ensures that the triac does not revert to the non-conducting state when the alternating load current falls below the specified holding current of the device.[2] The circuit is further simplified by using the triac in the I- and III-modes, that is with negative gate current and voltage for both the positive and negative half-cycles of the supply voltage.[3]

3.2. *Disconnexion of the Supply Voltage from the Load*

The inherent property of the triac to revert to the non-conducting state when the load current falls below a specified holding current[2] is used to disconnect the supply from the load.

Referring to Fig. 1, the supply may be disconnected at a load current zero following either a positive or a negative half-cycle of the supply voltage as preselected by switch SC. Assuming the former is selected, then gate G2 is enabled each time the negative waveform rises to the '1' state. That is at the end of each positive half- cycle.

However, G2 is inhibited by the $\bar{Q}$ output of M1 which is '0' until the operator throws switch SA to 'stop'. This applies a '1' to the K input of M1 so that M1 changes state when its clock input next changes from '1' to '0'. Gate G2, therefore, opens at the end of the next positive half-cycle of the supply voltage and changes the state of M2.

This switches off TR1 and deprives CSR1 of triggering current. Triac CSR1 therefore remains in the non-conducting state after the load current next passes through zero. Similarly, switch SC may be used to preselect a current zero occurring after a negative half-cycle of the supply voltage.

3.3. *Supply Voltage Zero Crossing Detector*

Figure 3 shows the supply voltage zero crossing detector circuit in more detail.

Transistor TR2 is used to detect when the line voltage of the supply passes through zero. Diode D1 biases the emitter of TR2 so that its base is approximately at earth potential. Transistor TR3 and gates G3 and G4 form a regenerative pulse-squaring circuit which is described elsewhere.[4] In practice the circuit regenerates when the line voltage passes through the threshold levels within the range of ± 0·3V. This is sufficiently sensitive for the immediate application of the circuit but can readily be improved by the use of a differential comparator.

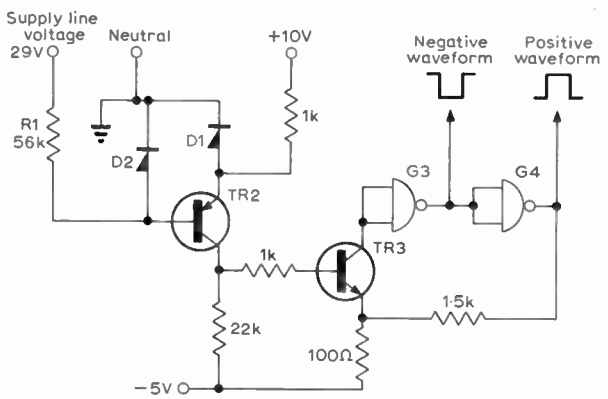


Fig. 3. Supply voltage zero crossing detector.

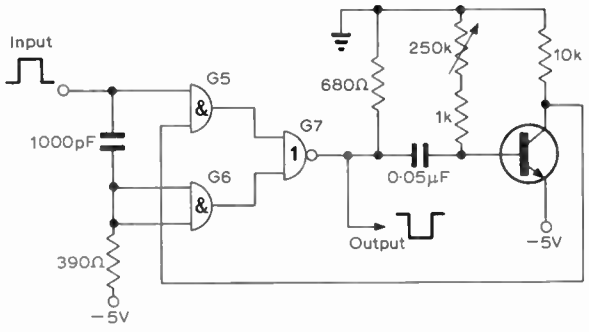


Fig. 4. Variable delay circuit.

3.4. *Variable Delay Circuit*

Figure 4 shows the variable delay circuit in more detail. This is described elsewhere.[5] Gates G5, G6 and G7 are contained in one exclusive-OR integrated circuit. The duration of the delay is continuously variable between 40 μs and 10 ms with a 50 Hz input square wave. The circuit is relatively insensitive to supply voltage variations. A change of ±25% in the −5V supply voltage produces only approximately ∓2% change in the delay period. In practice this change can be greatly reduced by the use of a stabilized supply.

Other means of incorporating the delay in the logical operation of the synchronous switch are possible. However, the operation described earlier was chosen because, at the sacrifice of the first 40 μs of the selected starting half-cycle of the supply waveform, it has the advantage that the delay circuit is triggered continuously. This enables the operator to use a double-beam cathode-ray oscilloscope accurately to align the end of the delay period with a chosen point in the supply voltage waveform, before switching switch SA to the 'start' position.

3.5. *Automatic Operation*

If switch SD is opened, the circuit is independent of the manual control switch SA and will cycle continuously through the preselected start-stop sequence. This utilizes the facility of the J–K flip-flop by which it changes state in response to every negative going change of its clock input when both its J and K inputs are '1' or open-circuited. The performance is best illustrated by Fig. 5 (see Sect. 3.6).

3.6. *Examples of Load Current Waveforms Obtained*

Figure 5 shows the load current waveforms obtained with a resistive load for various combinations of 'start' and 'stop' conditions provided by the synchronous switch. For convenience these waveforms were obtained with switch SD open to provide automatic operation as described in section 3.5. With manual operation similar waveforms are obtained but the number of complete cycles of the supply current is determined by the operator.

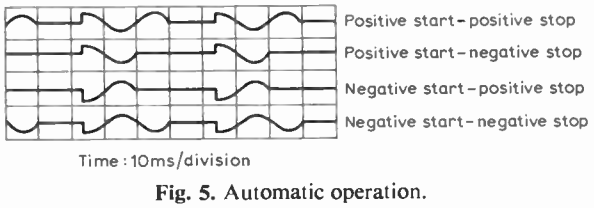


Fig. 5. Automatic operation.

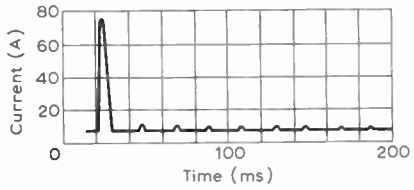


Fig. 6. Initial magnetizing current of transformer.

Figure 6 shows the transient inrush current taken by a small auto-transformer with a steady state magnetizing current of 85 mA. The triac used has a peak one-cycle surge forward current rating of 100 A. Other triacs with larger surge current ratings are available.

4. Practical Considerations

All the integrated circuits used are contained in three dual in-line packages, namely, one dual master-slave J–K flip-flop, one quadruple NAND gate and one exclusive-OR circuit. It will therefore be appreciated that the circuit is simple and inexpensive to construct.

The −5V supply for the integrated circuits is derived from the alternating mains supply by means of a transformer, a full-wave rectifier and a simple series stabilizer. The +10V supply is derived in a similar way but is unregulated. The primary of the transformer is tapped so that the synchronous switch can be used with a number of different supply voltages. However, it should be noted that resistor R1 of the voltage zero crossing detector circuit is permanently connected to the 29V tapping of the primary winding rather than directly to the line of the alternating supply used. This reduces the phase error of the detector circuit when other alternating supply voltages are used.

5. Conclusions

A simple synchronous switch has been described that has been designed to fulfil a specific requirement. Many extensions or variations of the principle of operation described are possible. For example, a counter might be used to control the state of J–K flip-flop M1 so that the number of cycles of the mains supply applied to the load can be accurately controlled. Alternatively, the principle could be extended to the control of three-phase supplies.

6. Acknowledgments

The author wishes to acknowledge the assistance of Mr. W. G. Forte, Head of the School of Electrical Engineering, South Australian Institute of Technology, who initiated the project and made the facilities available for the investigation.

7. References

1. Puchstein, A. F., and Lloyd, T. C., 'Alternating-Current Machines', p. 147 (Wiley, 1936).
2. 'SCR Manual', 4th Edition, p. 14 (General Electric, New York, 1967).
3. *ibid.*, p. 13.
4. 'Semiconductor and Components Data Book—1', p. 41 (Texas Instruments Ltd., 1968).
5. *ibid.*, p. 49.

Manuscript first received by the Institution on 14th July 1970 and in final form on the 17th March 1971. (Paper No. 1391/CC103.)

The Problem of Safe Navigation in Confined Waters

Report of a Conference called by the E.E.A.

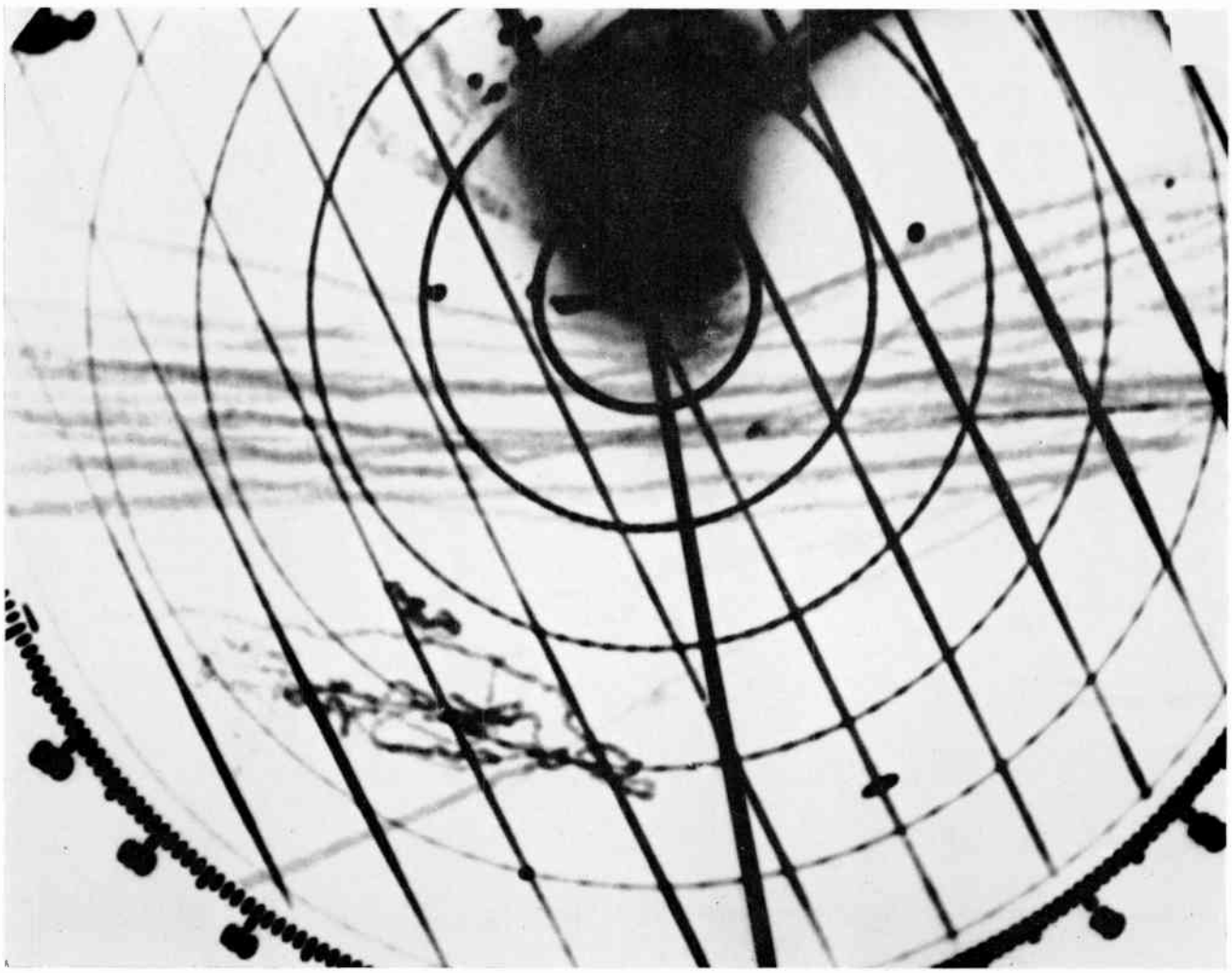
A short conference was held in London at the Royal United Service Institution on 11th May 1971, on the subject of 'The Problem of Safe Navigation in Confined Waters'. It was organized by the Electronic Engineering Association, whose Director, Captain R. A. Villiers, C.B.E., R.N., C.Eng., was the chairman, supported by a panel of four members:

Captain D. S. Tibbits, D.S.C., R.N.(Retd.) of Trinity House
Captain E. O. Jones of Shell International Marine
Mr. B. N. Steele of the National Physical Laboratory Ship Division
Captain A. C. Manson of the Department of Trade and Industry

Those invited to attend were from the fields of maritime administration and research and development, both Government and commercial, ship operators, maritime safety authorities, maritime insurance and salvage organizations, training organizations and authorities; observers attended from this Institution and other learned societies. Members of the E.E.A. who have interests in the manufacture of equipment for maritime use also sent representatives.

After a brief introduction by Capt. Villiers, the panel members each put forward their views on the problem and possible solutions.

Capt. Jones quoted recent marine incidents, particularly collisions in the English Channel, as the reason for the conference. He emphasized that in searching for a solution a sense of proportion must be preserved; new equipment evolved from Services requirements deserved consideration, but



Long-exposure photograph of film of radar screen obtained during the N.P.L.'s exploratory marine traffic survey in the Dover Strait. Dungeness Point is at the centre of the ring. A frame of cine film was exposed every 20 seconds. This still photograph shows the combination of 180 such exposures so producing streaks rather than spots to represent each vessel. The streaks across the photograph show vessels in through passage and those on the bottom left-hand side show fishing vessels. *N.P.L. photograph Crown copyright.*

'gimmicks' should be treated with caution. While equipment deficiencies played a part in some accidents, human error or failure also contributed—even undue reliance on a single source of information. He felt that the solution should be sought through improvements in equipment and in user training. There were also external solutions such as routing, particularly in the difficult situation of one traffic lane meeting or crossing other lanes. The presentation of information in a form acceptable to the user was important, some of the newer equipments failing in this respect (e.g. digital data presentation). Other fields in which improvement was required was the measurement of speed, particularly very low speeds over the ground in berthing, and in the Collision Regulations, now being undertaken. He did not believe there was any real conflict between safety and cost, and while there was still much to be done, the electronics industry had justly earned a good reputation in this field.

Mr. Steele described how the N.P.L. had become involved in the problems of data collection with regard to the traffic patterns in the Dover Straits and statistics on collisions. They would continue with studies based on the simulation of traffic patterns, and how these might develop in the future. Initially the collection of data by surveillance radar was wrongly thought to be simple, but further data were found essential, e.g. why were some ships using particular lanes. Identification (by aircraft or surface craft) and correlation with departure and arrival ports was necessary, but so far this had not been possible in darkness or fog. Time-lapse photography of a radar display yielded valuable data on the proximity of ships before taking avoiding action which was often as close as 1 or even $\frac{1}{2}$ mile. He suggested that a system by which each ship reported its arrival at an identifying position would be valuable, but there was a problem of finance.

Capt. Tibbits recommended that the problem should be clearly isolated and defined. The majority of ships were well equipped, manned and handled, but faced dangers from navigational hazards and other ships. What put them into those dangers? Not all ships could be so described, and 'bad' ships were increasing in number, but on the other hand most 'flag' countries had reasonable regulations which were subjected to continuous review. If all ships obeyed the collision and routing regulations, the problem would be small and human error needed two humans to result in a collision. Yet with 200,000 ships passing through the Dover Straits annually, some 20 serious collisions occurred and an unknown number of near misses—it would be interesting to know how many. Comparison with air traffic control requirements had been made, but it was his firm opinion that the many differences far outweighed the few similarities, so that common solutions did not apply, particularly 'control'. One problem was how to enforce obedience to the present reasonable laws; increasing the legal requirements merely hampered those who obeyed, but had no effect on those who ignored the law. A balance had to be struck between the regulators and the regulated, and the regulations should give maximum freedom to operate well-equipped ships by well-trained crews.

Capt. Manson pointed out that as ships had become larger and more numerous, the old regulations about keeping to the starboard side of a channel were becoming extended to apply outside these channels. The U.K. had tried to obtain international backing for this by introducing at I.M.C.O. the mandatory rule that a ship in a routing lane *must* go in the direction laid down for that lane. This arose because although only 5% of ships in these lanes were travelling in the wrong direction, these ships were concerned in 70% of the actual collisions. Proposals for position reporting procedures conflicted with the watchkeeping hours of radio operators and while this might be solved by the carriage of ship-to-shore v.h.f. radio this was not yet universal. He believed it was wrong to assume that 'good' ships never had accidents, and could quote examples in support. Similarly, the competence of certificated personnel was no guarantee of safety. The problem which did require study was the interface between the human observer and his equipment. Undue reliance on one source of information without cross-checking was believed to be a fruitful cause of accidents. In many proposals international agreement was essential and this had proved difficult and slow of achievement.

General discussion then followed, in which the more important points were:

Improvement in communication and identification, including language problems.

Study of the man-machine interface, possibly with assistance from the Medical Research Council.

The root causes of wrong appreciation of correct data—what are they?

The difficulty of U.K./French 'control' of the 80% of foreign shipping using the Dover Straits.

The financing of improvements, and the cost of safety.

Bridge design and lighting and its influence on fatigue, distraction and human error.

The use of racons for ship identification.

Manufacturers' representatives pointed out that considerable study of the man-machine interface had contributed to the design of new equipment, particularly in the anti-collision field; support for these views came from the Services research establishments. Many suggestions were not new, but were impractical for reasons stated by the panel and other speakers.

Summing up the discussion, Capt. Villiers said there had been a wide coverage of the problems and possible solutions—surveillance, radio reporting, etc. The panel had to point out the difficulties in implementing these proposals, but the resultant suggestion from the floor that they were 'complacent pessimists' was untrue, since it was a fact of life that time was needed for the implementation of the proposals discussed. No universal panacea had been proposed, but training, understanding of the human factors, careful analysis of the present developing situation, all had to contribute. He felt that the progressing development of international specifications for equipment would widen markets and reduce costs, and British industry had been well to the fore in this field from the start. In conclusion he thanked the members of the panel for their opening remarks and comments, and the audience for the interesting discussion.

This observer would agree that no new proposals arose in the course of the conference, but the various aspects of the problem and some possible solutions were frankly discussed. The implementation of many proposals will encounter difficulty, sometimes technical, but more often due to the international nature of maritime operations and administration. Progress can only be slow, and the conference made some contribution to the more widespread understanding of the difficulties to be overcome in solving this growing problem.

A. J. HARRISON