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. 1 JANUARY 1971

THE RADIO AND ELECTRONIC ENGINEER

The Journal of the Institution of Electronic and Radio Engineers

Expansion or Contraction

IN a computer program, every digit has a meaning. In comparable terms, authors and publishers have a reason for showing the date of publication which will thereby mark the pertinence of the contribution to human understanding of social and economic conditions or to scientific progress. Thus while all publications are, or should be, dated, periodicals go further by indicating the sequence of volumes, thereby establishing authority and demonstrating the continuing interest and concern of many people in the subject matter of the publication. History shows that if the publication is no longer socially relevant or has too broad interests, it either ceases publication or is overtaken by new publications specifically designed to meet current interests. In simple terms, there is either an expansion of special interest or a contraction of appeal.

In the field of radio and electronics there is, most certainly, no contraction of appeal, for there is scarcely a commercial, industrial, Government, scientific or engineering (in all its facets) publication which does not introduce some development or application of electronics into its content as a major development in particular interests and of value to the readers of the individual publication. In itself, this fact shows the expansion of interest in electronics: it is substantiated by the large number of commercial publications, some short lived, which have in the last decade or so reaped a financial reward in attempting to provide an understanding of the revolution which electronics has made on the social and economic life of most of the human race.

This issue of *The Radio and Electronic Engineer*, the commencement of the 41st volume of the Institution's *Journal*, represents a continuation of the need of the radio and electronic engineer to record his contribution to human progress. Further, it emphasizes the ability of the professional engineer to maintain a publication divorced, in part at least, from immediate commercial gain but nevertheless essential in terms of indicating a time when further contributions to understanding are necessary.

While the format and, to a certain extent, the content of the *Journal* has changed with this issue these changes represent further expansion. Indeed the Institution itself, through its publications, has always endeavoured to demonstrate the expansion which continues to be the foundation of its existence. Such evidence is apparent in the nostalgia with which one can look back at the early volumes containing papers on, for instance, colour television, waveguides and printed circuits, which today are taken as commonplace in trade and industry. Research papers now being published, though at present of apparently very narrow and specialized interest, may well be regarded in the years to come as foundation stones of whole new areas of technology.

The philosophy of contraction is to be seen in the approach of the contributors of review or survey papers which take a retrospective look at the way in which a particular subject has developed to its present stage. The more skilled the writer, the more effectively he can distinguish the significant steps and in this way contract, or condense, the work of perhaps a hundred papers by nearly as many individual scientists and engineers to yield a coherent statement that can make technical sense to the non-specialist.

The functions of the Institution in arranging meetings and conferences for the presentation and discussion of papers and in publishing a *Journal* for the greater dissemination of the ideas and achievements which these papers set forth, are germane to its purpose of promoting knowledge. In this work there is considerable dependence on the efforts and goodwill of many—contributors of papers and their organizations, those providing facilities for meetings and so on—and not least the backing of the Institution's members themselves. For it must not be overlooked that in effect the Institution subsidizes 'learned society' ventures throughout the world and through its *Journal*. The expansion of the scope of that publication to match the ever widening frontiers of electronics is of paramount importance and contraction to meet growing pressures is an unacceptable course of action which would hinder the progress of knowledge.

G.D.C.

NOTICES

Statement by the N.Z. Engineers Registration Board

The Engineers Registration Board of New Zealand has announced adoption of the following policy with regard to membership of the Engineering Institutions comprising the Council of Engineering Institutions:

(a) All persons who, on 9th July 1970 were corporate members of any one of the 14 constituent institutions of the Council of Engineering Institutions of the United Kingdom (C.E.I.) and who have been granted the status of 'Chartered Engineer' are deemed to possess the academic qualifications required for registration under the Engineers Registration Act, 1924. Each application for registration received from such a person will be examined and any such applicant may, at the discretion of the Board, be required to attend a Professional Interview for the purpose of establishing his practical experience and professional capability.

(b) All persons who henceforth are granted the status of 'Chartered Engineer' by the C.E.I. will be deemed to be academically qualified for registration but may, at the discretion of the Board, be required to attend a Professional Interview.

(c) All persons who qualify for admission to graduate membership (in some institutions known as associate membership) of a constituent institution of the C.E.I. by 31st December 1970 will be recognized as having met the educational requirements of the Board and will be required to attend a Professional Interview.

(d) Persons who obtained their educational qualifications before 31st December 1970 and who have not been able to satisfy the Board's training and professional experience requirements by 31st December 1973 may be required to sit and pass the whole or part of the C.E.I. Part 2 examination *unless* the qualification obtained before 31st December 1970 was a university degree in engineering or science recognized by the Board, or a pass in the C.E.I. examinations, Parts 1 and 2, or a diploma or certificate in engineering recognized by the Board.

The Training of Professional Engineers

The theme of the 5th International Congress of Engineers which will mark the 20th Anniversary of FEANI (Federation Europeene d'Associations Nationales d'Ingenieurs) will be 'The Training of Professional Engineers'. This Congress will take place in London between 27th September and 1st October 1971. Subjects of the main sessions are:

Training methods in the member nations of FEANI.

The provision of organized industrial training for

(i) Professional engineers

(ii) Supporting staff

The scope of training. Is technology enough?

Further training within the original discipline and retraining in a new or related engineering discipline.

The fee for attendance at the Congress will be £26 which will cover the cost of pre-prints of the papers to be given, a copy of the Congress Proceedings and luncheon on 28th, 29th and 30th September. A full programme of Social Events including a Ladies Programme is being arranged.

The Congress is being organized for FEANI by the British Institution of Civil Engineers. Please write to:

The Secretary, (FEANI Congress), Institution of Civil Engineers, Great George Street, London S.W.1.

The Importance of English to Engineers

The Council of Engineering Institutions is concerned at the tendency for technological education to concentrate on technology without comparable attention to lucid practical English. Professional engineers and technicians deal with people, and the majority have to give other people instructions. A man or woman with a technological education must at least be able to speak, write and sketch clearly. Without these abilities the technologist can only be inadequately equipped. The main instrument must remain the written and spoken word, and good communication still depends upon the proper use of words.

The Council therefore will attach much importance to clarity of expression in the two compulsory papers of its Examination, namely: 'The Presentation of Engineering Information' (a Part 1 paper, to be introduced in 1972), and 'The Engineer in Society'. It hopes that those responsible for the education of professional engineers and technicians will address themselves to this important matter, and encourage students as normal practice to read widely, to write lucidly and to take part in professional discussions.

Convention in Israel

The seventh National Convention of Electrical and Electronic Engineers in Israel will be held in Tel Aviv on 19th–22nd April, 1971. It is being organized by the I.E.E.E. Section in Israel with the association of the I.E.R.E. Israeli Section and will include original papers on theory and design and new technological developments drawn from the whole field of electrical and electronic engineering. Papers have been invited from many countries outside Israel and will be in English. Further information from:

Convention Technical Committee, P.O. Box 3386, Tel Aviv, or from the Honorary Secretary of the I.E.R.E. Israeli Section, P.O. Box 1214, Holon, Israel.

Correction

The following correction should be made in the paper 'The use of cathode-ray tubes in professional equipment' published in the December, 1970, issue of *The Radio and Electronic Engineer:* Page 290, Table 1, third column: the units of rise-time should read (ns).

U.D.C. 621.319

Electrolytic Capacitors

By

D. S. CAMPBELL,
B.Sc., D.I.C.. D.Sc., F.Inst.P.†

Presented at a Components and Circuits Group Symposium on Capacitors held in London on 28th April 1970.

A brief survey is given of aluminium and tantalum electrolytic capacitors. The fundamental behaviour of the dielectrics are discussed particularly with reference to the temperature coefficient of capacitance. It is shown, however, that the fundamental properties are masked by the presence of electrolyte, used to heal the dielectric film during use and to connect to the dielectric when it is deposited in the etched channels that can be formed in both aluminium and tantalum foils. The practical preparation of both etched aluminium and tantalum foil and also sintered tantalum capacitors are discussed. Scanning electron micrographs are used to illustrate the structures obtained.

1. Introduction

This paper is concerned with electrolytic capacitors; the dielectric is an oxide produced on a metal by anodization. The growth process of the oxide is analogous to the thermal growth of an oxide in corrosion. Although most metals will form an oxide during corrosion there are a few which form a highly tenacious continuous oxide which in fact rapidly protects the underlying metal from further attack. The thickness of oxide on these metals is normally limited but it can be enhanced by growing the oxide in an electrolytic bath with the metal the anode of the electrode system and a potential applied between the electrodes. This is the process known as anodization.[1, 2, 3] Metals that can be anodized to give continuous coherent oxides include zirconium, niobium, silicon, aluminium and tantalum. These last two are the most widely used in electrolytic capacitors and they will be discussed in detail.

The thickness of the oxide resulting from the process of anodization will depend on the voltage applied across the electrode system. The thinnest used in practice is about 8 nm and the thickest about 1 μm. In comparison with capacitors using plastic dielectric, this dielectric is, at its thickest, of the same order of thickness as the thinnest plastic sheet used but it can be two orders thinner in the extreme case.

The growth of an oxide on tantalum and aluminium during anodization is illustrated in Fig. 1. This shows a growth curve analogous to a thermal growth curve for both tantalum and aluminium and the thickness of the film is shown as a function of time for an applied voltage across the anodizing cell of 30 V. It can be seen that for a given voltage tantalum gives a thicker oxide than does aluminium. One can quantify this by noting that tantalum gives a thickness of 1·60 nm for every volt applied whereas aluminium gives a thickness of 1·36 nm for every volt applied. The type of growth illustrated in Fig. 1 is known as a constant voltage growth curve. Figure 2 shows a different type of characteristic obtained if a constant current is applied. In this case the constant current must imply a continual increase of voltage so that the thickness will increase as long as the current is applied and the voltage is available. There is however, a limit to the voltage that can be applied across the oxide above which impurities in the foil, the temperature of the bath and various other factors will cause the oxide film to break down. This limit in the case of aluminium is at 1100 V giving a maximum thickness of 1·5 μm. In the case of tantalum the limit is at 700 V giving a thickness of 1·1 μm.

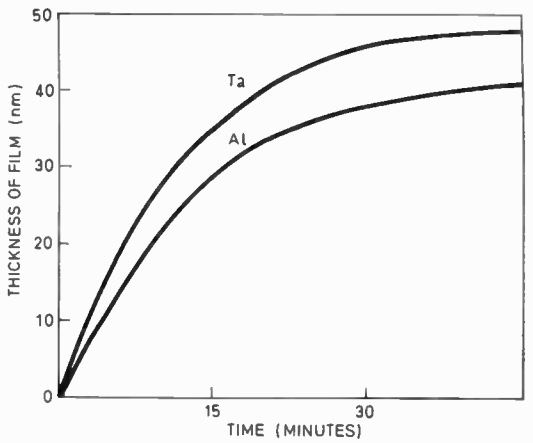


Fig. 1. Constant voltage growth of films (30 V).

The dielectric that is obtained by these anodization processes is amorphous,[4] i.e. shows no crystalline structure in electron diffraction. The behaviour of these dielectrics fits into the general picture of behaviour of dielectrics.[5, 6, 7] One aspect of this picture is that a general relationship can be established between temperature coefficient of capacitance, γ_c, and permittivity,

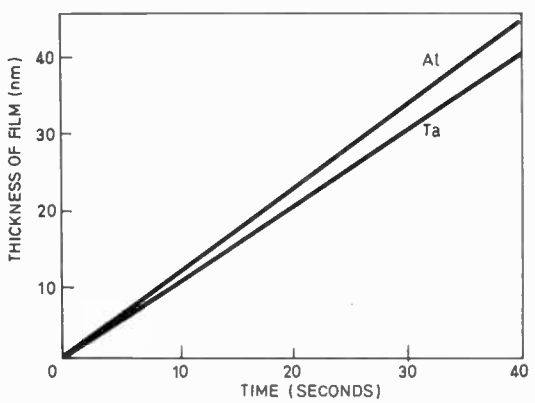


Fig. 2. Constant current growth of films (2 mA/cm²).

† The Plessey Co. Ltd., Components Group, T.C.C. Capacitor Division, Whiteside Works, Bathgate, West Lothian.

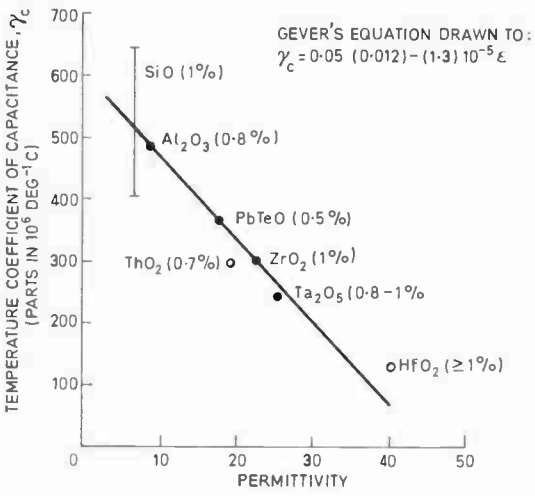


Fig. 3. Temperature coefficient of capacitance as a function of permittivity (showing experimental data for films with indicated loss $> 0{\cdot}1\%$, i.e. extrinsic γ_c).[9]

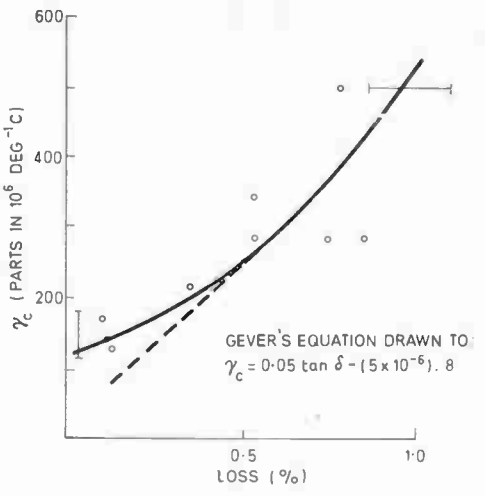


Fig. 5. Temperature coefficient of capacitance as a function of loss for Al_2O_3 films (the point at lowest loss is for bulk Al_2O_3).[9]

ε, for a given dielectric loss.[8,9] Figure 3 illustrates temperature coefficient of capacitance of a variety of oxide films shown as a function of permittivity for a loss of 1·2%, and it can be seen that alumina and tantala films fit the curve. Figure 4 illustrates the limit that is obtained for γ_c vs ε for very low loss ($\tan\delta < 0{\cdot}1\%$). The curve shows the limit below which the temperature coefficient cannot fall even if the loss is less than 0·1%. How rapidly alumina approaches this state is shown in Fig. 5, where the temperature coefficient of capacitance is shown as a function of loss. Not all the points shown on this graph are of anodized films; the lowest value obtained is using bulk material. However, the minimum value for the temperature coefficient of capacitance of alumina is just above +100 parts in 10^6 deg^{-1}C even at 0·1% loss. In the case of tantala the figure that is obtained is a minimum value of approximately −100 parts in 10^6 deg^{-1}C for the lowest loss. In practice films have not easily been obtained with such a value of loss and hence such a low temperature coefficient of capacitance. In practical terms a loss of 2% in alumina will be associated with a temperature coefficient of +1000 parts in 10^6 deg^{-1} C.

The general picture that emerges is that both alumina and tantala films prepared by anodization can be explained in modern dielectric terms. Under low fields the dielectric exhibits conduction by the charge carriers hopping between structure dependent traps in the dielectric.[10] The temperature coefficient behaviour which has been shown illustrates the viability of this approach. Unfortunately although one obtains by anodization a reasonably good dielectric, especially in the case of tantala, with a basically low loss, in practical terms this low loss is not of very much use.

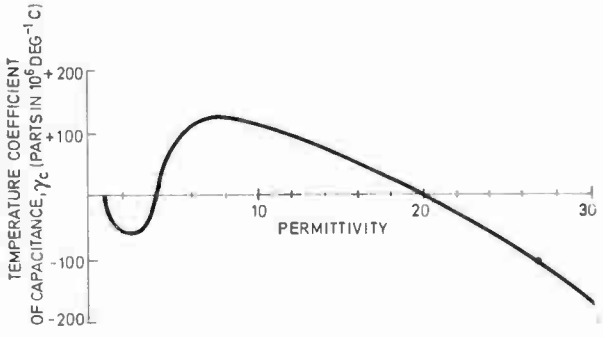


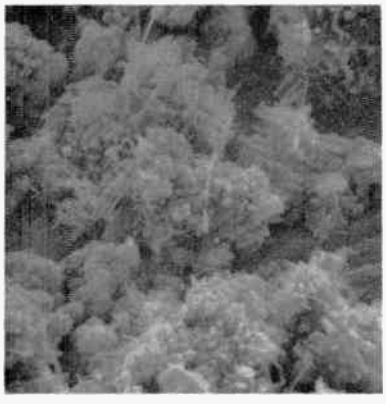
Fig. 4. Temperature coefficient of capacitance as a function of permittivity.[9]

2. Aluminium Capacitors

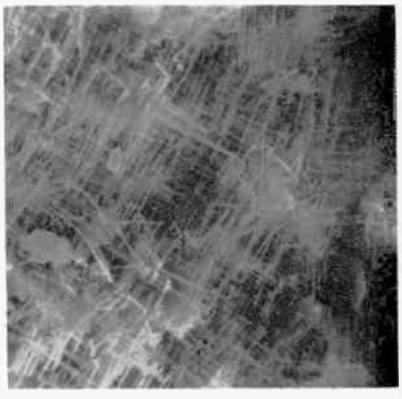
Let us now consider aluminium electrolytic capacitors in a little more detail. The first thing that we find that goes wrong with the basic concepts that we have evolved up to date is that it is found that in order to obtain a satisfactory thin dielectric (< 300 nm) it is necessary to use the film in an environment where any weaknesses in the film are either isolated or healed in use. A healing environment is arranged by immersing the total metal/oxide system in an electrolyte and always applying voltage so that the anodized metal is the anode, i.e. any weak points in the film will become anodized during use. However, the presence of the electrolyte adds series resistance to the capacitor and thereby substantially increases the loss.

Secondly, in order to obtain a large enough capacitance in a small space it is necessary to etch the aluminium foil,[11,12] so as to increase the surface area of the foil to be anodized. Figure 6(a) shows a scanning electron micrograph of the etch pits[12,13] at the initial stages of etching. This micrograph was obtained by etching to a certain depth, removing the aluminium foil from the etching bath and subsequently anodizing to such a thickness that all the etched pits have been filled in with oxide. The metal is then removed by immersing the foil in bromine in methanol so that an oxide is left in which the etch pits are standing proud of the oxide surface. The micrograph shown is thus looking at this oxide surface and the etch pits from the inside of the original metal. Figure 6(b) shows a further stage in the etching process. The original etch pits which were cubic in shape have now branched out and tunnels are beginning to form and grow out of the pits. At the limit a complete maze of etched

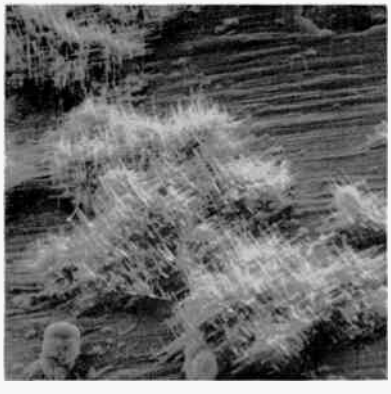
(a) Early stages of growth. Length of edge 50 μm.



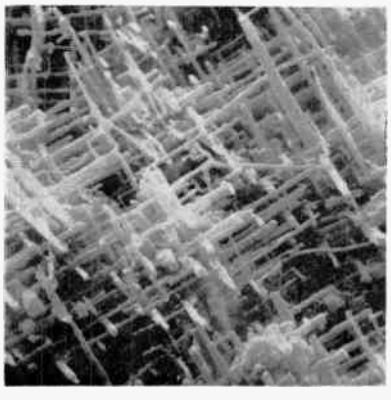
(b) Stage of growth showing commencement of tunnels. Length of edge 50 μm.



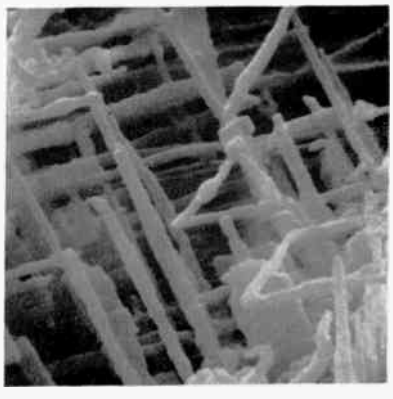
(c) Highly developed tunnel formation. Length of edge 50 μm.



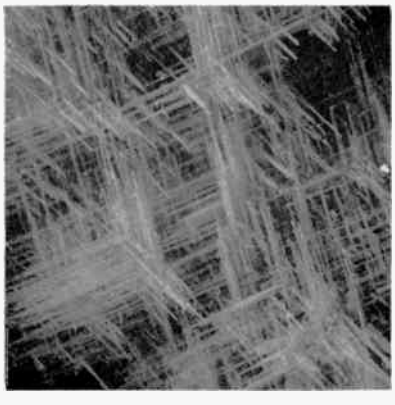
(d) Low magnification. Length of edge 100 μm.



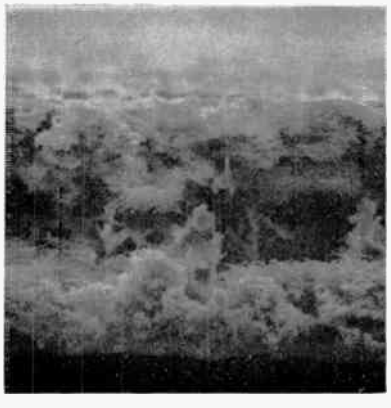
(e) Medium magnification. Length of edge 25 μm.



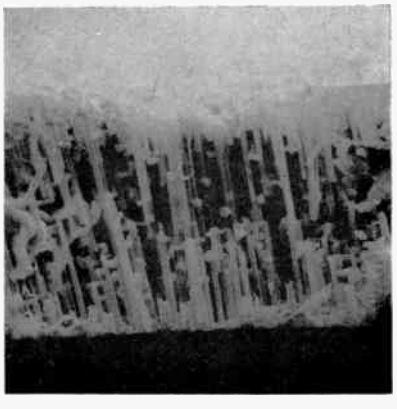
(f) High magnification. Length of edge 10 μm.



(g) Showing orientation effect. Length of edge 50 μm.



(h) Edge view of foil used for low-voltage applications. Length of edge 100 μm.



(i) Edge view of foil used for high-voltage applications. Length of edge 100 μm.

Fig. 6. Scanning electron micrographs of etched aluminium foil.

tunnels is obtained and this is illustrated in Fig. 6(c). Figures 6(d), (e) and (f) show tunnels under increasing degrees of magnification; Fig. 6(d) has an edge to the micrograph which is 100 μm in length, Fig. 6(e) 25 μm in length and Fig. 6(f) 10 μm. The tunnels run along crystallographic directions, and Fig. 6(g) shows the (100) orientation (N. F. Jackson, unpublished) in a single crystal part of the substrate. It can be seen that the tunnels are in fact very narrow and useless from the capacitor point of view if in the anodization the tunnels were completely filled up as they have been to obtain Fig. 6. In these circumstances any surface gain obtained by etching would be lost. There is thus a limiting thickness of oxide that can be grown in order to preserve the surface gain obtained by etching and this limitation implies a limiting voltage in use. In practice foils can be etched in one or two basic ways, either for use in high-voltage capacitors (> 100 V) or in low-voltage capacitors

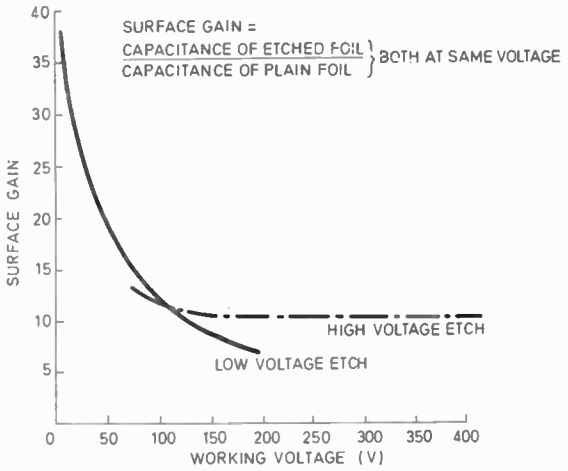
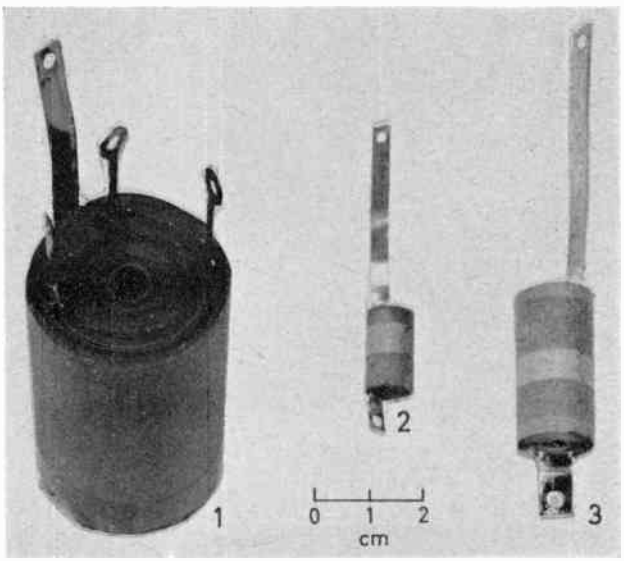


Fig. 7. Etched aluminium foil capacitors. Foil surface gain as a function of working voltage.

(< 100 V). In low-voltage capacitors very thin tunnelling structure can be used and hence very high gains can be utilized. However, for high-voltage applications it is necessary to have wide tunnels so that a thick oxide can be grown without filling in the tunnels completely. The two types of structures that are used are illustrated in Figs. 6(h) and (i). Figure 6(h) shows an electron micrograph of the edge of a high-gain foil that would be used for low-voltage application. It can be seen that the tunnelling structure extends a considerable way into the foil and in fact only just stops short of penetrating the foil completely. In contrast Fig. 6(i) shows, at the same magnification, the edge of a foil that has been etched for use at high voltage. In this case, it can be seen that wide straight tunnels have been formed. Figure 7 illustrates the surface gains obtained from these two types of etches and it can be seen that the cross-over is around 100 V. Above 100 V the initially high gain tunnelling structure which can be used in a low-voltage capacitor is lost as the tunnels fill in completely, whereas with the high-voltage etch the initial lower gain is nevertheless maintained up to large thicknesses of oxide.

Because of the presence of the tunnels it is always necessary to introduce an electrolyte into the capacitor to enable contact to be made to the opposite side of the dielectric from the metal. The effect of the series resistance introduced by this electrolyte is illustrated in Fig. 8 where loss is plotted as a function of working voltage for both an etched and, for comparison, a plain foil. It can be seen that the loss is higher in the etched foil case, as the series resistance in the etched and anodized tunnels will be higher than in the case of the plain foil.



(1) Triple type 250 μF+500 μF+50 μF, 300 V.
(2) Sub-miniature type 8 μF, 275 V.
(3) Tubular type 500 μF, 25 V.

Fig. 9. Aluminium electrolytic capacitors. Examples of wound unimpregnated cartridges

In practice aluminium electrolytic capacitors take the form of wound cylinders similar in external appearance to those obtained in paper or plastic capacitor construction. The windings taken in sequence are firstly the etched and anodized foil, secondly a porous paper wick which can subsequently be soaked in electrolyte and finally a cathode foil. Three examples of the wound cylinders are shown in Fig. 9.

The cathode foil will be very highly etched and very lightly anodized, if at all, so as to give a very high capacitance in comparison with the anode foil. As the anode foil and cathode foil will be effectively in series the effect of the cathode foil can be ignored if its capacitance is sufficiently above that of the anode foil.

After soaking the completed capacitor in electrolyte it is necessary to 'age' the capacitor, i.e. re-anodize at the necessary voltage. This will oxidize any surfaces which

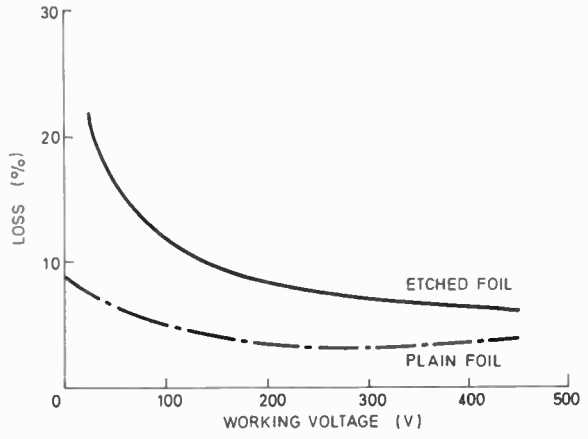


Fig. 8. Etched and plain aluminium foil electrolytic capacitors. Variation of power factor with working voltage.

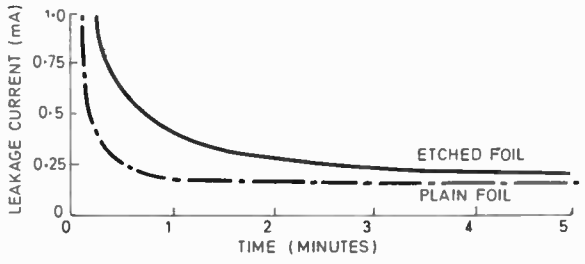


Fig. 10. Etched and plain aluminium foil electrolytic capacitors. Variation of leakage current with time for 8 μF, 450 V unit.

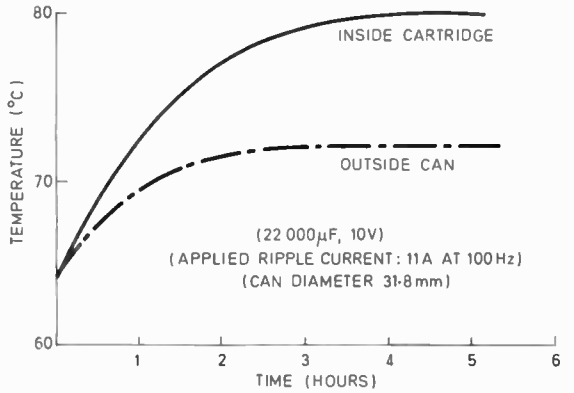


Fig. 11. Etched aluminium foil capacitors. Effect of ripple current on capacitor temperature.

have not been anodized previously. Furthermore, any weak points in the anode oxide will also be strengthened. Figure 10 shows examples of how the leakage current reduces during ageing as a function of time. It can be seen that a steady state is reached after approximately five minutes ageing.

One of the disadvantages of the series resistance due to the electrolyte is that there is a limitation to the maximum alternating current that the capacitor can handle without getting hot. The effect of this current is illustrated in Fig. 11 for an applied ripple (i.e. superimposed a.c.) current of 11 A at 100 Hz. It can be seen that when a steady state condition has been reached the temperature of the outside of the can is 70 deg C above its surroundings whereas the temperature in the inside of the can is a further 10 deg C above this. Figure 12 gives a more general picture of the effect of ripple current and the ripple current is here shown in terms of capacitor volume. In this case one is dealing with a small capacitor, the so-called sub-miniature type capacitor. The maximum ripple current is defined as that which gives a steady state temperature difference between outside and centre of the can of 10 deg C. It can be seen that the bigger the capacitor volume and hence can area the higher is the value of maximum ripple current.

Figure 13 shows the temperature coefficient behaviour of an etched aluminium foil capacitor. Various curves are shown for various sizes of capacitance. It should be noted that the temperature coefficient obtained from such

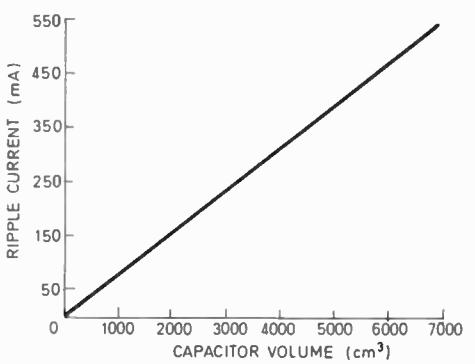


Fig. 12. Etched aluminium foil capacitors. Maximum ripple current as a function of capacitor volume. (Sub-miniature electrolytic capacitors.)

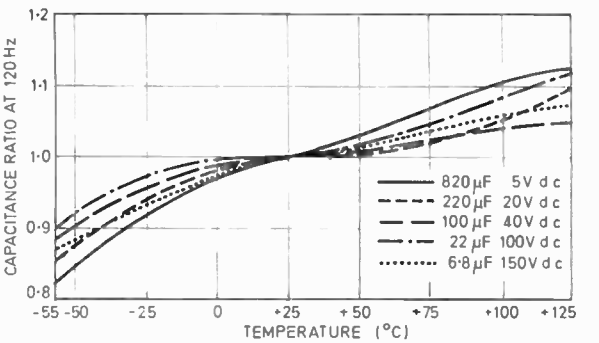


Fig. 13. Etched aluminium foil capacitors. Variation of capacitance with temperature.

a plot, given that a series bridge has been used to measure the capacitance, will be due to the basic dielectric itself, and not due to the series resistance introduced by the electrolyte. This property is in fact the only one remaining which results from the high grade of dielectric that one is able to make from an anodized film. The temperature coefficient shown agrees with the value of 2% loss and + 1000 parts in 10^6 deg^{-1} C quoted earlier.

Figure 14 shows another characteristic of an aluminium foil capacitor, obtained when impedance is plotted against frequency. At low frequencies the impedance is high and due to the capacitance of the capacitor. At very high frequencies, on the other hand, impedance is due to the inductance of the capacitance. In the middle frequency range the capacitor exhibits a minimum in impedance, the value of which is due to the electrolyte and to any resistance associated with tags and other connexions.

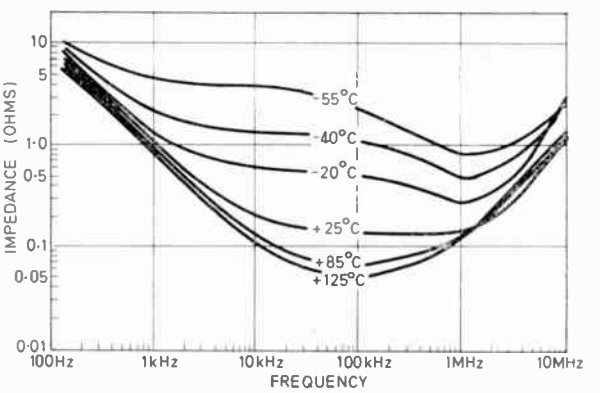


Fig. 14. Etched aluminium foil capacitors. Variation of impedance with frequency and temperature (150 μF, 20 V).

It can now be seen that the equivalent circuit of an aluminium electrolytic capacitor contains several elements and these are identified in Fig. 15. Figure 15(a) shows the complete circuit. Experimental values found on a 300 μF, 7 V capacitor are shown in Table 1.

Figure 15(b) is a simplified circuit showing the important features for low-frequency applications. The simplified circuit is obtained because at all times the anode and cathode foil resistances can be ignored. Furthermore as the capacitance of the cathode is usually much greater than that of the anode, the cathode capacitance can be ignored. The anode dielectric leakage resistance is very

1. 68,000 μF, 16 V. Computer power supply.
2. 1750 μF, 325 V. Photoflash.
3. 250 μF+500 μF+50 μF, 300 V. Triple—for television circuits.
4. 100 μF+200 μF+50 μF+25 μF, 300 V. Quadruple—for television circuits.
5. 20 000 μF, 2·5 V. Low voltage.
6. 500 μF, 25 V. Tubular.
7. 560 μF, 10 V. Sub-miniature.
8. 220 μF, 5 V. 125°C working. Extended temperature range.
9. 50 μF, 12 V. Plastic case.
10. 4 μF, 350 V. Plastic case, plug-in.
11. 1 μF, 250 V. Plastic case, plug-in.
12. 2 μF, 63 V. Plastic case, plug-in.

Fig. 16. Aluminium electrolytic capacitors. Complete units.

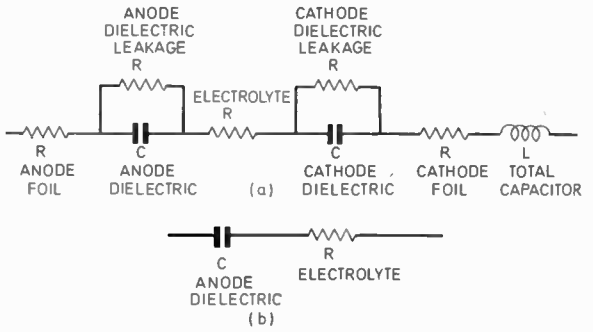


Fig. 15. Equivalent circuits of an aluminium electrolytic capacitor:
(a) Complete circuit.
(b) Simplified circuit at low frequencies.

Table 1

Experimental values of equivalent circuit elements of aluminium electrolytic capacitor (300 μF; 7 V)

R anode foil	$6\times10^{-3}\ \Omega$
R cathode foil	$13\times10^{-3}\ \Omega$
R corresponding to leakage of anode dielectric	$5\times10^{2}\ \Omega$ †
R corresponding to leakage of cathode dielectric	$8{\cdot}5\times10^{2}\ \Omega$ ‡
R electrolyte	$5\times10^{-2}\ \Omega$ §
C anode	370 μF
C anode	2000 μF
L total capacitor	0·008 μH

† The temperature coefficient of capacitance of the anode was +1200 parts in 10^6 deg^{-1} C. For a permittivity of 8, this corresponds to a loss of approximately 2% and hence a parallel resistance of $5\times10^{2}\ \Omega$.

‡ Calculated, assuming a resistivity of $10^{14}\ \Omega$cm. The cathode foil will have a dielectric thickness of 4 nm and a surface gain of 17 times whereas the anode foil had a dielectric thickness of 25 nm and a surface gain of 20.

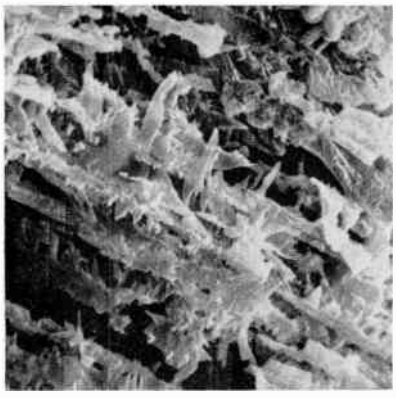
§ The impedance/frequency minimum was 0·05 Ω at 100 kHz. This figure was confirmed by plotting series resistance against $1/f^2$, where f is the measured frequency, and extrapolating to zero value for $1/f^2$.[14]

high in normal use, but after a capacitor has not been used for some time (e.g. > 6 months) it will be necessary to apply the working voltage for a short period (e.g. a few minutes) to increase the leakage resistance to its normal high value. At low frequencies the inductance can be ignored so that the equivalent circuit is reduced to a series capacitance and resistance.

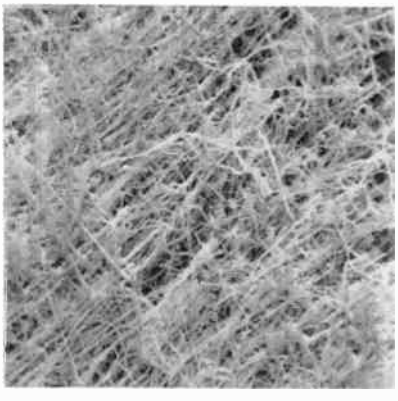
Figure 16 shows some typical capacitors ranging from computer power supply types that go up to 160 000 μF, 10 V in value, down to small plug-in capacitors of 1 μF, 63 V.

3. Tantalum Capacitors

With tantalum we are dealing with a system in which we are usually able to prepare a better dielectric by anodization than with aluminium. The result of this is that the capacitor has a better shelf-life than does the equivalent aluminium one. However, the disadvantage is that it is a material which has the cost of silver and the weight of lead and therefore it is not such an economic proposition to use tantalum as it is aluminium. Tantalum can be etched in a similar way to aluminium foil[15] and Fig. 17(a) shows a typical scanning electronmicrograph of an etched foil. The gain obtained is usually low but recent, unpublished work by K. Harrison has shown that



(a) Low gain etch. Length of edge 25 μm.



(b) High gain etch. Length of edge 25 μm.

Fig. 17. Scanning electron micrographs of etched tantalum foil.

it is possible to etch tantalum to give high surface gains and this is illustrated in Fig. 17(b). The relationship between gain and voltage is similar to the one obtained in aluminium and is shown in Fig. 18, both with regard to the high-gain etch developed at Plessey, Bathgate and with the standard commercial etch used for making tantalum foil capacitors.

The etched foil so obtained can be wound up in the same way as the aluminium body we previously discussed,

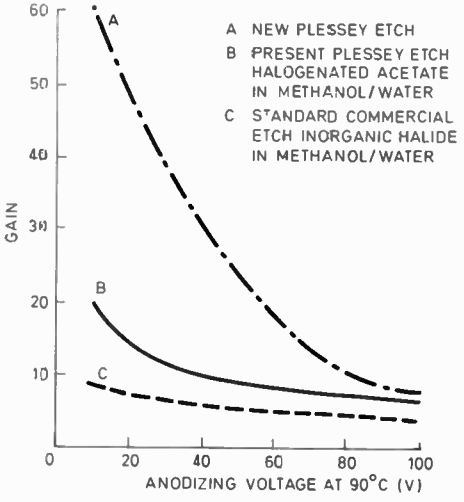
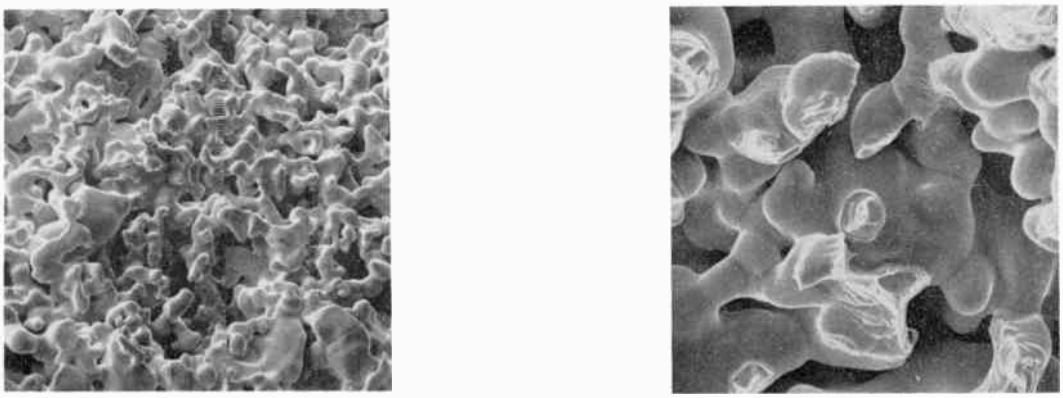


Fig. 18. Etched tantalum foil capacitors. Gain of various etched tantalum foils as a function of anodizing voltage.



(a) Sintered anode. Length of edge 100 μm.

(b) Sintered and anodized body. Length of edge 25 μm.

Fig. 19. Scanning electron micrographs of sintered tantalum.

using a paper wick, and an electrolyte is then absorbed in the wick. Such capacitors tend to be used at voltages above 100 V.

For lower voltages it is found best to form the high gain surface by sintering tantalum powder. Figure 19(a) shows a scanning electron micrograph[16] of such a sinter obtained by heating approximately 10 μm diameter tantalum particles to around 2000°C. Figure 19(b) shows a similar sinter which has been anodized and broken open so as to show the oxide which has grown on the individual grains. Figure 20 shows four sizes of anodes that are used for different values of capacitance and voltage; the connexion wire to the body is also of tantalum and is sintered into the block when it is formed. The problems associated with the preparation of tantalum capacitors using sintered powder are partly concerned with the variation of capacitance that one gets with sintering conditions.[17] Figure 21 shows the variation of capacitance obtained as a function of sintering temperature for both 4 μm and 10 μm diameter powder. It can be seen that the 10 μm powder, compared with 4 μm powder, gives a higher surface area for a firing temperature of 2100°C due to the lower sintering rate for a higher diameter powder,[18, 19] and a lower surface area if fired at 1800°C. An advantage

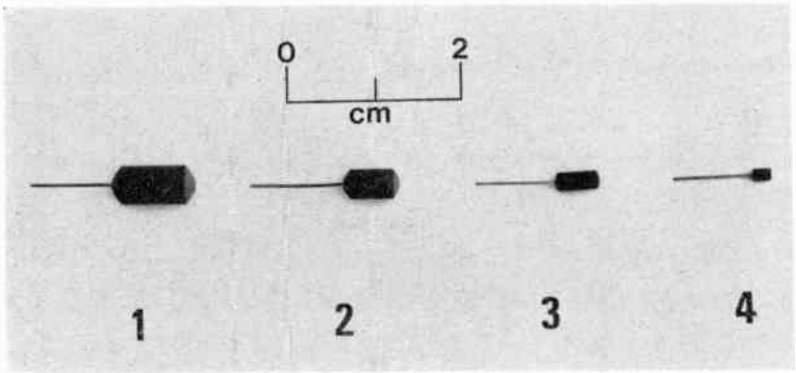


Fig. 20. Solid tantalum capacitors. Sintered untreated anodes.
1. $CV = 50$ μF V.
2. $CV = 200$–300 μF V.
3. $CV = 700$–1200 μF V.
4. $CV = 1500$–2000 μF V.

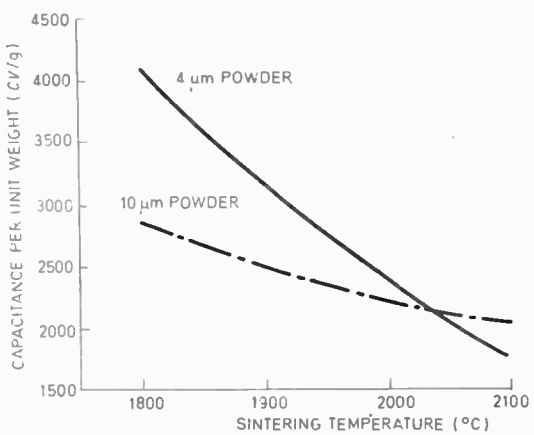


Fig. 21. Solid tantalum capacitors. Variation of capacitance with sintering temperature.

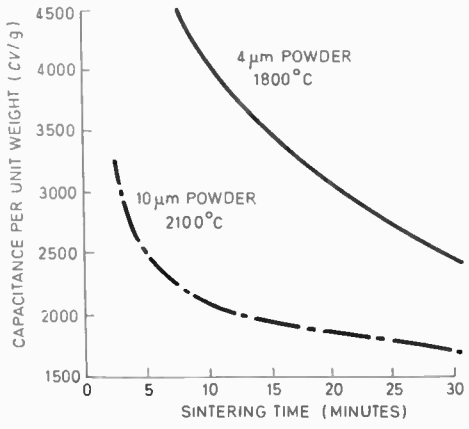


Fig. 22. Solid tantalum capacitors. Variation of capacitance with sintering time.

of the use of high sintering temperatures is that any impurities in the surface of the individual particles will be removed more effectively at high temperature.[20] Therefore, it does not follow, as one can see from Fig. 21, that it is always an advantage to use the smallest diameter powder, as this will only give large surface gains at low sintering temperatures. Figure 22 shows the effect of sintering time on the capacitance and it can be seen that this also will affect the value of capacitance of the final unit. Figure 23 shows the variation of the major parameters, tan δ, leakage current, and capacitance with degree of sintering. High temperatures or long times can give a reduction in leakage current and also a reduction in capacitance because of the efficient sintering; there will also be a rise in tan δ due to the very fine narrow channels that will result from a high degree of sintering.

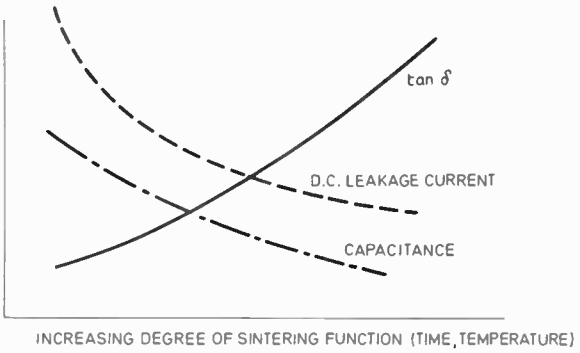


Fig. 23. Solid tantalum capacitors. Summary of variations of major parameters with degree of sintering.

One of the ways of showing up defects in tantala films grown on tantalum is by anodizing them so as to obtain field crystallization.[21] This extreme treatment shows up weak points in the oxide by causing crystallization where high current flows and Fig. 24(a) illustrates the results of such a treatment. In this sample, part of the sintered block was stuck to an adjacent block during the sintering process and when the blocks were separated a surface was thereby exposed from which impurities had not been removed during the sintering. Subsequent treatment to give field crystallization has shown up a large number of defect sites on the originally unexposed surface. Figure 24(b) shows two such sites in more detail. These points would be very low resistance paths through the dielectric and would thereby cause complete shorts in a practical capacitor.

Just as in the case of etched and anodized aluminium foil, it is necessary to connect to the opposite side of the dielectric from the original metal to form a complete capacitor. This can be done in two ways; either by immersing the sintered anodized body in a liquid electrolyte forming the so-called 'wet' tantalum capacitor or, by forming a layer of manganese oxide in the pores in the sintered body giving the 'solid' tantalum capacitor.[16, 22] The layer of manganese oxide can be prepared by immersing the porous body in a solution of manganese nitrate, removing the body and pyrolysing the solution that remains trapped in the pores by heating to approximately 400°C. The result of this is to give a thin coating of manganese oxide over the whole of the inside of the sinter. To obtain best characteristics however, it is necessary to repeat this process a number of times. The number of times varies between 6 and 20 depending on the type of capacitor being prepared. The effect of the number of pyrolyses or impregnations is shown in Fig. 25. It can be seen that the loss falls off as the number of impregnations and hence the thickness of the manganese oxide layer is increased but the d.c. leakage increases mainly due to the effect of the temperature cycling to which the dielectric has been subjected. Manganese oxide has a self-

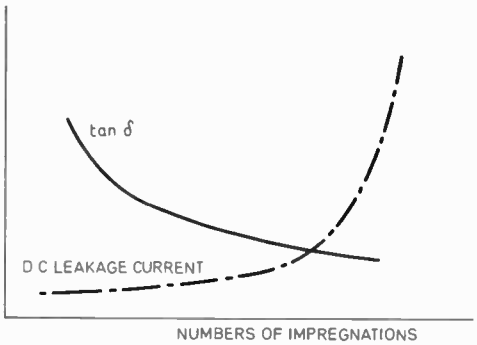
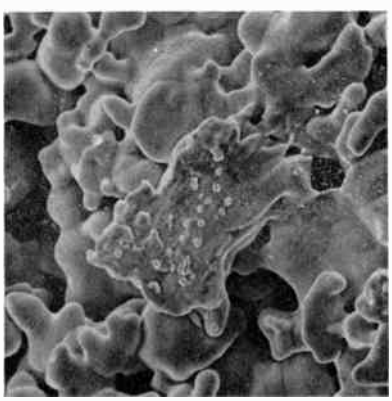


Fig. 25. Solid tantalum capacitors. Variation of d.c. leakage current, and loss with number of impregnations.



(a) Showing field crystallization on surface unexposed during sintering. Length of edge 50 μm.

(b) Showing detail of field crystallization. Length of edge 10 μm.

Fig. 24. Scanning electron micrographs of sintered tantalum.

healing effect similar to the electrolytes used in aluminium, tantalum foil and wet tantalum capacitors. It is not exactly clear what actually happens during the healing process but it is suggested that oxygen is transferred from the semi-insulating manganese dioxide which has been formed, to the oxygen-deficient tantala giving a fully insulating tantala film and at the same time a fully insulating manganese oxide film on top.[17]

Figure 26 shows a scanning electron micrograph of a complete sintered tantalum capacitor. Three separate regions can be distinguished. Firstly, the tantalum metal particles themselves, secondly the coating of oxide formed by anodization and finally a thick layer of manganese oxide built up on the outside of the grains.

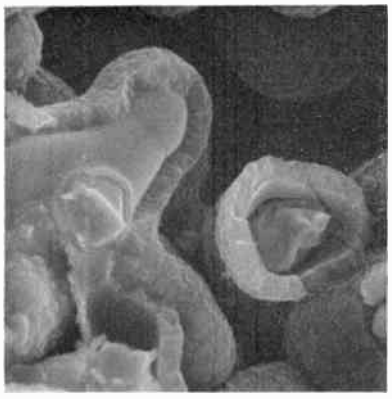


Fig. 26. Scanning electron micrograph of sintered tantalum showing the oxide layer and the manganese oxide layer in a completed capacitor.

The complete sinter is connected to the cathode by conducting paste and solder. Figure 27 shows a diagram of a typical capacitor.

Tantalum capacitors exhibit the same characteristics as aluminium ones. Figure 28 shows the ripple rating or ripple voltage permissible as a function of frequency and Fig. 29 shows an impedance-frequency characteristic for a solid tantalum capacitor. As in the case of aluminium the minimum is obtained when the capacitive and inductive reactance cancel out. However, the inductance is usually much lower than in the case of aluminium as there is no wound construction. The inductance will be considerably affected by the positioning of the leads and the variations shown in Fig. 29 can be due to slight changes in construction of this sort. The temperature coefficient of capacitance when measured on a series bridge is found to be approximately +1000 parts in 10^6 deg^{-1} C corresponding to a dielectric loss of 3%.[9]

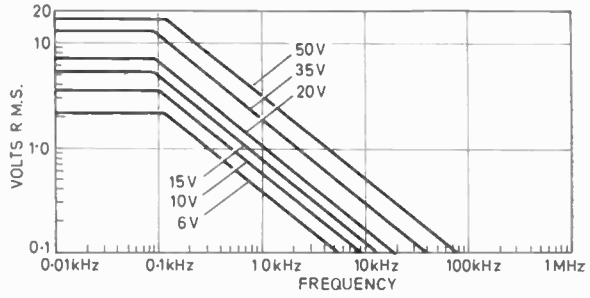


Fig. 28. Solid tantalum capacitors. Maximum permissible ripple voltage as a function of frequency.

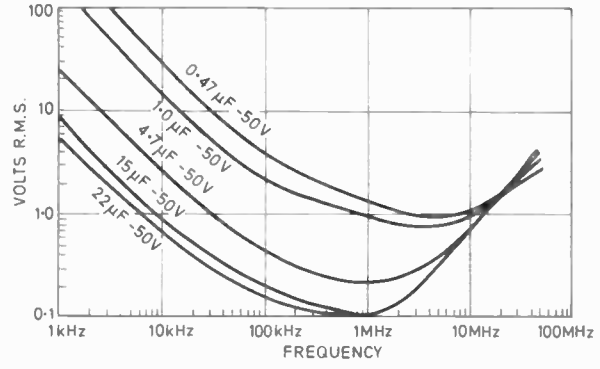


Fig. 29. Solid tantalum capacitors. Typical curves of impedance as a function of frequency at 25°C.

Figure 30 shows some typical tantalum capacitors. The top row, Nos. 1–6, shows various foil types, including an example of a capacitor made using the new high-gain foil developed at Bathgate (No. 6) (3000 μF). This may be compared with a similar construction using ordinary foil (No. 5) (580 μF). The second row shows a wet tantalum type. Various plastic-cased versions are shown in the next row, including recently developed flip-chip capacitors for use in microminiature applications (Nos. 10, 11 and 12). Metal case and tear drop capacitors are also shown, and at the bottom of the Figure are two small capacitors used in hearing aids.

4. Future Work

In terms of the new work that is going on to improve both aluminium and tantalum capacitors four things can be mentioned. In the case of aluminium there is constant striving to improve the surface gain so that capacitors with larger capacitances in smaller volume may be obtained. Secondly, work is continuing on the obtaining

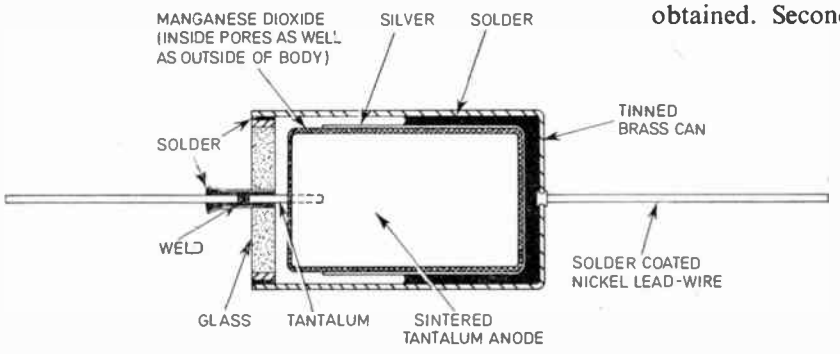
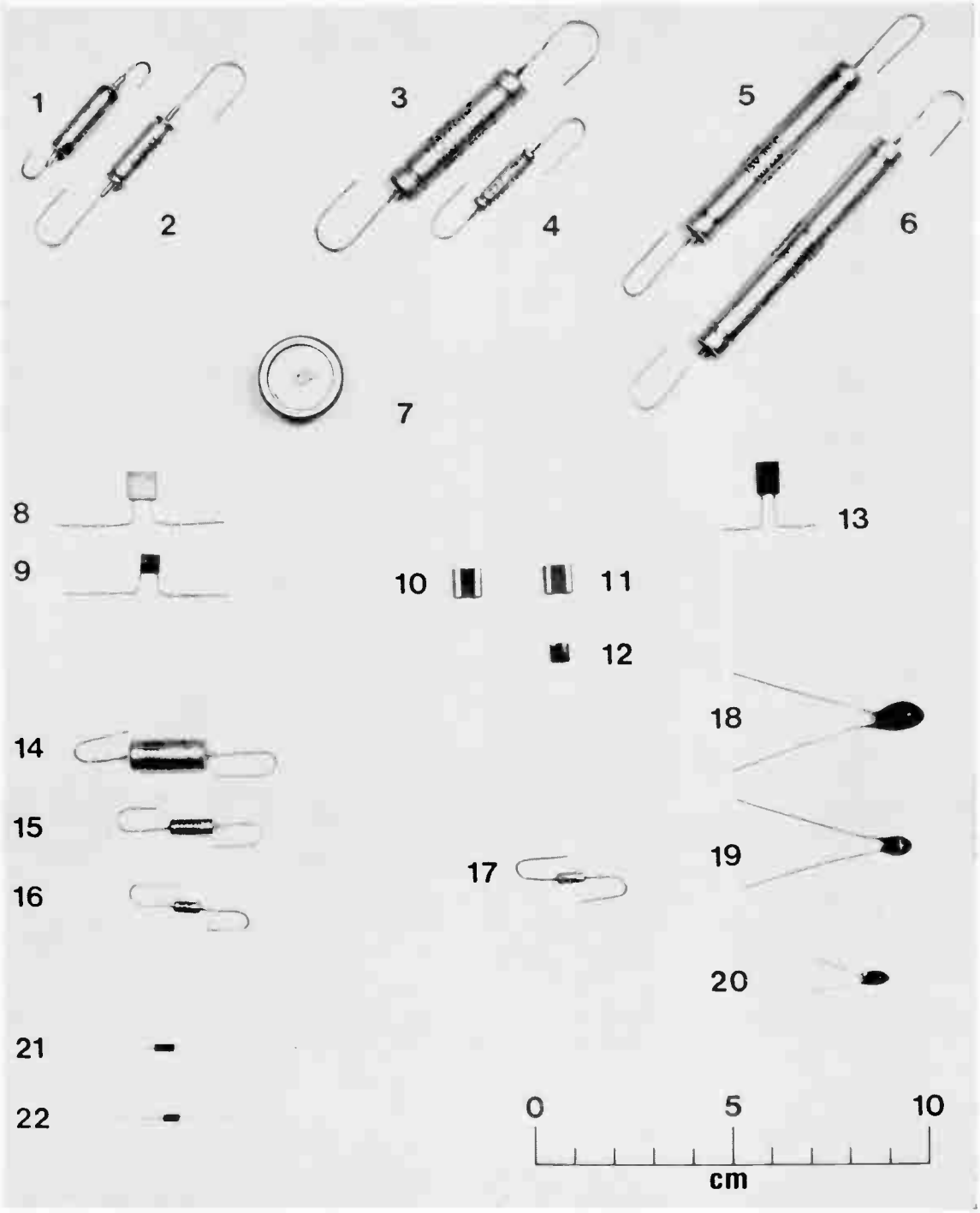


Fig. 27.
Solid tantalum capacitors. Diagram of the construction of a typical capacitor. (See Nos. 14, 15 and 16 in Fig. 30.)



1. Plain foil 16 μF, 6 V. 85°C.
2. Plain foil. Reversible. 8 μF, 6 V. 85°C.
3. Etched foil. 5 μF, 150 V. 125°C.
4. Etched foil. 0·22 μF, 160 V. 125°C.
5. Etched foil. 580 μF, 10 V. 125°C.
6. High gain etched foil. 3000 μF, 10 V. 125°C.
7. Wet electrolyte. 50 μF, 70 V.
8. Plastic case, solid. 3·9 μF, 10 V.
9. Plastic case, solid. 4 μF, 20 V.
10. Flip-chip, solid. 6·8 μF, 20 V.
11. Same type as 10—reverse side.
12. Flip-chip, solid. 1 μF, 50 V.
13. Transfer moulded plastic case, solid. 6·8 μF, 20 V.
14. Metal case, glass seal, solid. 15 μF, 35 V.
15. Metal case, glass seal, solid. 1·5 μF, 50 V.
16. Metal case, glass seal, solid. 0·068 μF, 35 V.
17. Metal case, resin seal, solid. 0·068 μF, 35 V.
18. Tear drop, solid. 68 μF, 15 V.
19. Tear drop, solid. 3·3 μF, 33 V.
20. Tear drop, solid. 1 μF, 35 V.
21. Hearing aid, solid. 4·7 μF, 6 V.
22. Small hearing aid, solid. 3·3 μF, 6 V.

Fig. 30. Tantalum electrolytic capacitors. Complete units.

of electrolytes for use in capacitors over a wide temperature range; figures from −55°C to +125°C are in fact quoted.

In the case of tantalum, we have already seen that increases of surface gains obtained by tantalum etching are being obtained. High-gain etches for tantalum foils will result in a much wider application of what is basically a very expensive material. In the case of sintered tantalum, different shaped powders are being investigated that will give high surface areas. New sintering methods are also under study.

In summary the basic aim of all new work is to give a larger capacitance in a smaller space and also to give a capacitor that is more reliable both on shelf-life and in actual use.

5. Acknowledgments

The author would like to thank his staff in the R. & D. laboratories of The Plessey Co. Ltd., at Bathgate for help in preparing this paper, in particular Mr. M. Bruce in relationship to tantalum and Dr. D. Eastham in relationship to aluminium. He would like to acknowledge the work of the materials research group at The Plessey Co. Ltd., Allen Clark Research Centre, Caswell, and in particular Mr. Norman Jackson who was instrumental in preparing the scanning electron micrographs of etched aluminium and tantalum foils and sintered tantalum powders.

Finally the author would like to thank The Plessey Co. Ltd., for permission to publish this paper.

6. References

1. Young, L., 'Anodic Oxide Films'. (Academic Press, London, 1961).
2. Hoar, T. P., 'Modern aspects of electrochemistry', Vol. 2, p. 262 (Butterworth, London, 1959).
3. Campbell, D. S., in 'The Use of Thin Films in Physical Investigations', p. 11. (Ed. Anderson, J. C.). (Academic Press, London, 1966).
4. Stirland, D. J. and Bicknell, R. W., 'Studies of the structure of anodic films on aluminium', *J. Electrochem. Soc.* **106**, p. 481, 1959.
5. Lamb, D. R., 'Electrical Conduction Mechanisms in Thin Insulating Films', (Methuen, London, 1967).
6. Simmons, J. G., in 'Handbook of Thin Film Technology', p. 141, (Ed. Maissel, L. I. and Glang, R.). (McGraw Hill, New York, 1970).
7. Harrop, P. J. and Campbell, D. S., in 'Handbook of Thin Film Technology', p. 16.1, (Ed. Maissel, L. I. and Glang, R.). (McGraw Hill, New York, 1970).
8. Cockbain, A. G. and Harrop, P. J., 'The temperature coefficient of capacitance', *J. Phys., D (Appl. Phys.)*, **1**, p. 1109, 1968.
9. Harrop, P. J. and Campbell, D. S., 'Selection of thin film capacitor dielectrics', *Thin Solid Films*, **2**, p. 273, 1968.
10. Jonscher, A. K., 'Electronic properties of amorphous dielectric films', *Thin Solid Films*, **1**, p. 213, 1967.
11. Bakish, R., Kornhaas, R. J. and Borders, E. Z., 'Etching of hard aluminium foil', *Electrochem. Tech.*, **6**, p. 192, 1968.
12. Dunn, C. G. and Bolon, R. B., 'Technique for a scanning electron microscope study of etched aluminium', *J. Electrochem. Soc.*, **116**, p. 1050, 1969.
13. Jackson, N. F., 'Improvement of Wet Aluminium Capacitors', Plessey Co. Ltd., Internal Reports, 1968–1970.
14. Morley, A. R., 'The effect of the electrolyte resistance in porous anodes', *Proc. Instn Elect. Engrs*, **117**, p. 1648, 1970.
15. Jenny, A. L. and Ruscetta, R. A., 'Electrolytic etching of dense tantalum', *J. Electrochem. Soc.* **108**, p. 442, 1961.
16. Jackson, N. F., 'Improvement of Solid Tantalum Capacitors', Plessey Co. Ltd., Internal Reports 1968–1970.
17. Martin, G. L., Fincham, C. J. B. and Chadsey, E. E., 'A study of factors affecting the electrical characteristics of sintered tantalum anodes', *J. Electrochem. Soc.*, **107**, p. 332, 1960.
18. Nichols, F. A., 'Coalescence of two spheres by surface diffusion', *J. Appl. Phys.*, **37**, p. 2805, 1966.
19. Nichols, F. A. and Mullins, W. W., 'Morphological changes of a surface of revolution due to capillarity—induced surface diffusion', *J. Appl. Phys.*, **36**, p. 1826, 1965.
20. Fincham, C. J. B. and Martin, G. L., 'Purification of tantalum anodes during sintering', *J. Electrochem. Soc.*, **107**, p. 658, 1960.
21. Vermilyea, D. A., 'The crystallization of anodic tantalum films in the presence of a strong electric field', *J. Electrochem. Soc.*, **102**, p. 207, 1955.
22. Taylor, R. L. and Haring, H. E., 'A metal-semiconductor capacitor', *J. Electrochem. Soc.*, **103**, p. 611, 1956.

Manuscript received by the Institution on 30th July 1970. (Paper No. 1359/CC94.)

U.D.C. 621.376.223

Multi-frequency Analysis of Switching Diode Modulators under High-Level Signal Conditions

By

A. M. YOUSIF, B.Sc., M.Sc.†

and

J. G. GARDINER, B.Sc., Ph.D.†
(Graduate)

Distortion products of intermodulation, cross-modulation and harmonic types can be calculated in switching modulator circuits by evaluating the coefficients of the multiple Fourier series resulting from the interaction of many signals of differing frequencies with a bi-linear diode characteristic.

An analysis of this type is developed in order to establish the range of signal levels for which approximate predictions of distortion product magnitudes are accurate and to demonstrate that under conditions where approximate techniques cannot be justified, satisfactory and rapid calculation of distortion performance is possible using straightforward computer procedures without recourse to general non-linear circuit analysis techniques.

1. Introduction

The problem of evaluating the magnitudes of low-level distortion products generated in severely non-linear circuits under high-level drive conditions remains acute owing to the very large number of components in the output spectrum of such circuits which are at a higher level or at best comparable with the intermodulation, cross-modulation or harmonic distortion products of interest. Recent work of Neill[1] indicates that computation of general non-linear circuits under these conditions is possible but in a number of specialized circuits, analytical evaluation of distortion levels yields valuable insight into the mechanisms by which these unwanted effects are produced and, in consequence, some indication of how designs may be improved. Ring, shunt, series, hybrid and star modulators using Schottky-barrier diodes are typical cases in point.

An elementary series modulator was considered many years ago by Bennett.[2] This consisted of a single diode in series with source and load resistances and driven by two input signals at different frequencies; the spectrum of current components in the loop was calculated by evaluating the coefficients of the resulting double Fourier series on the assumption that the diode characteristic was bi-linear. It was further demonstrated by Belevitch[3, 4] that the result obtained could be applied with only very minor modification to other modulator configurations. However, since Bennett's original analysis was restricted to two signals it was not possible to use it to predict third-order intermodulation effects arising when three or more signals are input to the circuit. An approximate analysis due to Tucker[5] overcomes this difficulty to some extent and the distortion properties of a variety of resistively terminated modulators are now well documented,[6, 7, 8] in situations where the input signal voltages at the diodes are known to be small relative to the local-oscillator voltage.

However, two important problems remain: (1) Over what range of input signal levels for a given applied local oscillator e.m.f. are the approximate results of refs. 6–8 accurate or at least an acceptable guide to the performance of a practical mixer? (2) When approximate techniques are clearly invalid, i.e. for input signal voltages approaching the local-oscillator voltage at the diodes, can switching function analysis provide a basis for the rapid computation of distortion levels using subroutines readily available in the great majority of computer facilities as an alternative to the general and sophisticated time-domain non-linear analysis procedures which are available in only a few specialized installations and which are for many practical problems slow and costly to run?

The idealized case of an exponential diode with zero bulk resistance has been treated by Pilyagin.[9] The present work is concerned, however, with the more practical case of a bi-linear diode and with demonstrating that it is possible to formulate the analytical expressions for the coefficients of the multiple Fourier series which result from a bi-linear diode model in such a way that very straightforward numerical integration procedures yield rapidly computed results.

2. Computation of Output Component Magnitudes

Considering the elementary series mixer of Fig. 1, we can represent the applied voltage as

$$V_{in} = \sum_{n=0}^{\infty} V_{sn} \cos \omega_n t. \qquad \ldots\ldots(1)$$

The current flowing in the circuit is a function of the applied voltages, i.e. $f(\omega_1 t, \omega_2 t, \omega_3 t \ldots \omega_n t)$ and may be expressed in the form of a multiple Fourier series in terms of these independent variables. Thus

$$f(\omega_1 t, \omega_2 t \ldots \omega_n t) = \sum_{m_1=0}^{\infty} \sum_{m_2=0}^{\infty} \cdots \sum_{m_n=0}^{\infty} A_{m_1 m_2 \ldots m_n} \times \times \{\cos(m_1\omega_1 t \pm \ldots \pm m_n\omega_n t)\} \qquad \ldots\ldots(2)$$

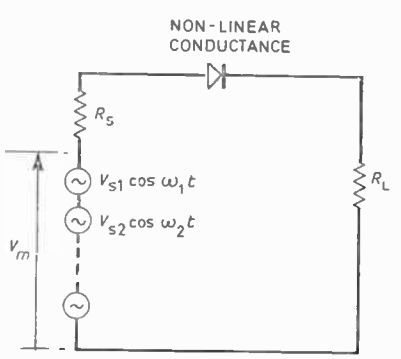


Fig. 1. Elementary series modulator.

† Postgraduate School of Electrical and Electronic Engineering, University of Bradford. (Mr. Yousif is on study leave from the Sudan Ministry of Communications.)

If the diode is assumed to be a bi-linear device then its voltage gain is a discontinuous function of the total applied voltage; high when V_{in} is positive (forward bias), low when V_{in} is negative (reverse bias). Thus if a suitable function, $\phi(t)$, takes the values 1 and 0 respectively for V_{in} positive and negative, then the output voltage V_L is obtained from

$$V_L = V_{in} \cdot K \cdot \phi(t) \qquad \ldots\ldots(3)$$

where K is an amplitude coefficient taking account of the relative values of R_s and R_L etc.

The coefficients of the series (2) are now determined in the usual way by solution of the integral

$$A_{m_1 m_2 \ldots m_n} = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi} f(\omega_1 t, \omega_2 t \ldots \omega_n t) \times \cos(m_1\omega_1 t \pm m_2\omega_2 t \pm \ldots \pm m_n\omega_n t) \times \mathrm{d}(\omega_1 t)\,\mathrm{d}(\omega_2 t)\ldots\mathrm{d}(\omega_n t) \qquad \ldots\ldots(4)$$

where $f(\omega_1 t, \omega_2 t \ldots \omega_n t)$ is $\phi(t)$.

A convenient discontinuous function exists in the form:

$$\frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{\sin(V_{in}\gamma)}{\gamma}\,\mathrm{d}\gamma = 1 \quad \text{for} \quad V_{in} \geqslant 0$$
$$= 0 \quad \text{for} \quad V_{in} < 0 \qquad \ldots\ldots(5)$$

This function by itself represents the output of an ideal 'hard limiter' and has been studied extensively in this context[10] so that only the principal steps in the mathematical argument are presented here. Extension to the switching-diode mixer/modulator is readily accomplished using (3) by making the substitution (5) in (4) to give

$$A_{m_1 m_2 \ldots m_n} = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi}\left\{\left[\frac{V_{in}}{2} + \frac{V_{in}}{\pi}\int_0^\infty \frac{\sin V_{in}\gamma}{\gamma}\,\mathrm{d}\gamma\right] \times \cos(m_1\omega_1 t \pm m_2\omega_2 t \ldots \pm m_n\omega_n t) \times \mathrm{d}(\omega_1 t)\,\mathrm{d}(\omega_2 t)\ldots\mathrm{d}(\omega_n t)\right\} \qquad \ldots\ldots(6$$

For the particular case of only two applied signals this integral takes a standard form, namely that of Weber and Schafheitlin. Solution of the three signal case has been set out by Watson[11] in terms of gamma and hypergeometric functions, but this result is not particularly convenient for either hand or computer evaluation when investigating practical mixer situations, nor is it practicable to obtain corresponding analytical solutions for higher numbers of input signals. However, by making use of standard Bessel function identities and interchanging the order of integration it may readily be shown that

$$A_{m_1 m_2 \ldots m_n} = \mathrm{Im}\left\{\frac{2}{\mathrm{j}\pi}\prod_{i=1}^{n}\mathrm{j}^{m_i}\int_0^\infty \frac{1}{\gamma}\frac{\mathrm{d}}{\mathrm{d}\gamma}\left[\prod_{k=1}^{n}\mathrm{J}_{mk}(V_{sk}\gamma)\right]\mathrm{d}\gamma\right\} \qquad \ldots\ldots(7)$$

$$= \frac{2}{\pi}(-1)^{\frac{\sum_{i=1}^{n}|m_i|-2}{2}} \times \int_0^\infty \frac{1}{\gamma}\frac{\mathrm{d}}{\mathrm{d}\gamma}\left[\prod_{k=1}^{n}\mathrm{J}_{mk}(V_{sk}\gamma)\right]\mathrm{d}\gamma \qquad \ldots\ldots(8)$$

for

$$\sum_{i=1}^{n}|m_i|$$

even and non-zero.

Components $A_{m_1 m_2 \ldots m_n}$ take the value zero when

$$\sum_{i=1}^{n}|m_i| - 2$$

is odd with the exception of the fundamental currents at impressed frequencies which give values determined by the d.c. component of (5).

3. Calculation of Harmonic and Intermodulation Distortion Components

The above result permits evaluation, in principle, of any order of harmonic and intermodulation distortion component arising from the application of any number of input signals. From a practical point of view, constraints are imposed, not only by available computation time, but also by consideration of realistic operating conditions for typical mixers and modulators. It would be unreasonable to expect a mixer to operate satisfactorily in the presence of a large number of input signals comparable in amplitude with the local oscillator; a more typical situation arises when the input consists of one or perhaps two high-level interfering signals together with a wide spectrum of low-level signals. We will, therefore, investigate the case of two-high-level inputs in order to establish, as indicated in the introduction, the range of signal levels over which the approximate techniques of Tucker are valid and the nature of departures from the approximate results as the signal levels approach that of the local oscillator.

Considering the input voltage to take the form:

$$V_{in} = V_{s1}\cos\omega_1 t + V_{s2}\cos\omega_2 t + V_{s3}\cos\omega_3 t$$

where V_{s3} would normally be a high-level local-oscillator supply and V_{s1} and V_{s2} test signals.

From equation (3)

$$A_{m_1 m_2 m_3} = \frac{2}{\pi}(-1)^{(|m_1|+|m_2|+|m_3|-2)/2} \times \int_0^\infty \frac{1}{\gamma}\frac{\mathrm{d}}{\mathrm{d}\gamma}\left[\mathrm{J}_{m_1}(V_{s1}\gamma)\mathrm{J}_{m_2}(V_{s2}\gamma)\mathrm{J}_{3m}(V_{s3}\gamma)\right]\mathrm{d}\gamma \qquad \ldots\ldots(9)$$

where $|m_1|+|m_2|+|m_3|$ is even and $\neq 0$, equation (9) may be evaluated by integrating by parts to give

$$A_{m_1 m_2 m_3} = \frac{2}{\pi}(-1)^{(|m_1|+|m_2|+|m_3|-2)/2} \times \int_0^\infty \frac{1}{\gamma^2}\mathrm{J}_{m_1}(V_{s1}\gamma)\mathrm{J}_{m_2}(V_{s2}\gamma)\mathrm{J}_{m_3}(V_{s3}\gamma)\,\mathrm{d}\gamma \qquad \ldots\ldots(10)$$

i.e. the large-signal sideband at $\omega_3+\omega_1$ is given by

$$A_{101} = \frac{2}{\pi}\int_0^\infty \frac{1}{\gamma^2}\mathrm{J}_1(V_{s1}\gamma)\mathrm{J}_0(V_{s2}\gamma)\mathrm{J}_1(V_{s3}\gamma)\,\mathrm{d}\gamma \qquad \ldots\ldots(11)$$

with a similar expression for the sideband at $\omega_3+\omega_2$. Harmonic distortion products of the form $\omega_3 \pm k\omega_1$ are obtained from

$$A_{k01} = \frac{2}{\pi}(-1)^{(k-1)/2} \times \int_0^\infty \frac{1}{\gamma^2}\mathrm{J}_k(V_{s1}\gamma)\mathrm{J}_0(V_{s2}\gamma)\mathrm{J}_1(V_{s3}\gamma)\,\mathrm{d}\gamma \qquad \ldots\ldots(12)$$

with similar expressions for products $\omega_3 \pm l\omega_2$, and intermodulation products of the form $\omega_3 \pm p\omega_1 \mp q\omega_2$ given by

$$A_{pql} = \frac{2}{\pi}(-1)^{(p+q-1)/2} \times \times \int_0^\infty \frac{1}{\gamma^2} J_p(V_{s1}\gamma) J_q(V_{s2}\gamma) J_l(V_{s3}\gamma)\, d\gamma \quad \ldots\ldots(13)$$

and so on.

The relative magnitudes of the components of interest are calculated by numerical integration of the above expressions. The computer program used in the present work employed standard subroutines available in the installation at Bradford (ICL 1909) which compute the Bessel function values to accuracies of parts in 10^8 and execute integration by Simpson's rule to better than parts in 10^5. The overall accuracy of the computation was, therefore, determined by the choice of upper limit for γ. A range of values for this was investigated and it was found that convergence was sufficiently rapid for a limit of 30 to be adequate—increasing this to 500 improved the accuracy by only about 1%. Taking a value of 30 still gives far greater accuracy than is usually required in practical situations but resulted in typical computation times of 30 seconds per product. This could clearly be greatly reduced by writing suitable reduced accuracy Bessel function and integration procedures if large numbers of products are of interest.

4. Application to Practical Balanced-Diode Mixers

As indicated in the Introduction, application of the foregoing results to practical modulators and mixers using balanced configurations of diodes, rather than a single non-linear element, is readily accomplished and results in the rejection by balance of some components[6] while leaving the unbalanced product levels unaltered from the elementary series mixer case, provided that the values taken for the signal input voltages accommodate the input transformer turns ratios. In the case of the ring, the local-oscillator voltage is taken as that existing at the local-oscillator port of the mixer, i.e. the centre taps of the transformers, and making allowance for the effects of the diode offset voltages as indicated in Ref. 7.

Experiments were carried out using a typical Schottky-barrier shunt mixer employing four diodes type HP 2900 in a bridge arrangement. The transformer turns ratios were 1 : 3 and the results are shown plotted in Fig. 2 in comparison with results predicted by both the foregoing analysis and the approximate analysis of Tucker. It is of interest to note that for the products evaluated (which are typical of the most important distortion products arising in communication applications) the Tucker analysis is valid to high accuracy for signal levels up to within 10 dB of the local-oscillator drive level.

The measuring equipment used in these experiments has been described elsewhere[8] and is shown diagrammatically in Fig. 3. The frequencies used were: for

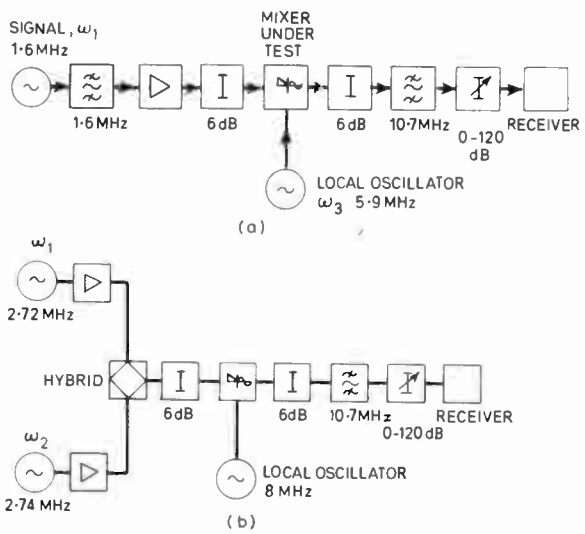


Fig. 3. Basic test set-ups for (a) harmonic distortion, and (b) intermodulation distortion.

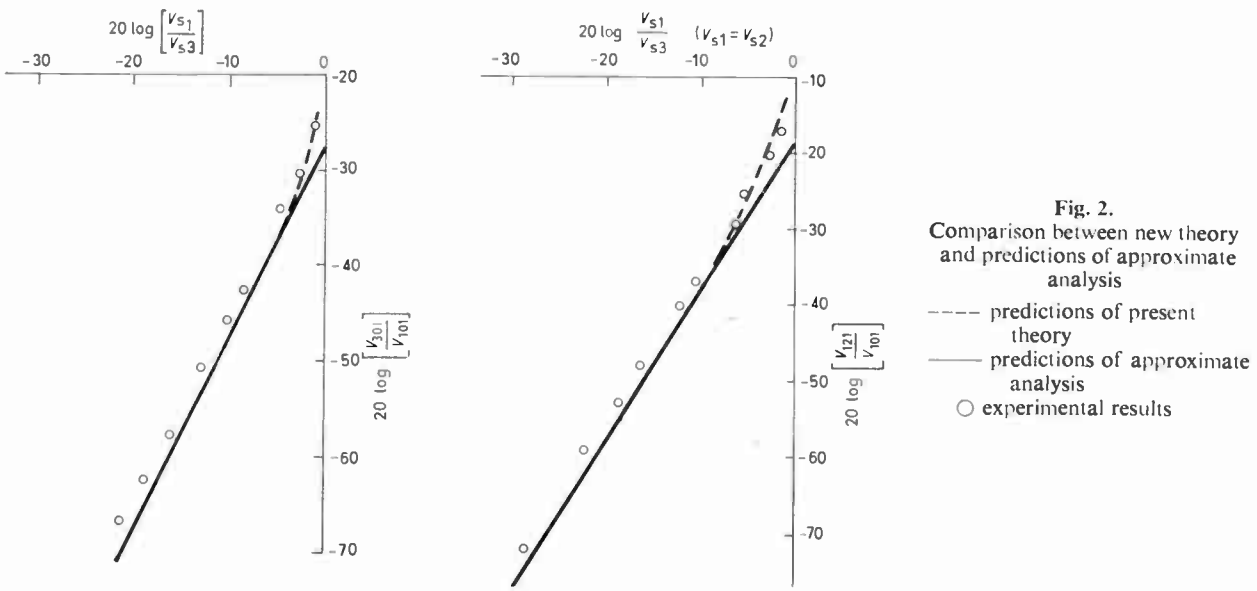


Fig. 2.
Comparison between new theory and predictions of approximate analysis
---- predictions of present theory
—— predictions of approximate analysis
○ experimental results

harmonic distortion, input at 1·6 MHz, with local oscillator at 5·9 MHz for third harmonic measurements so that the distortion product output was obtained at 10·7 MHz, and for intermodulation distortion equal strength input signals at 2·72 and 2·74 MHz with local oscillator at 8 MHz.

5. Conclusion

There are two major conclusions from the present work:

(a) The approximate analysis of Tucker gives useful predictions for much greater signal voltage levels than had previously been supposed, i.e. up to one third of the local-oscillator voltage at the diodes rather than one-tenth which Tucker had taken as an upper limit for his approximations.

(b) For larger signal voltages than this where significant departures from the approximate results are evident a straighforward computer program using standard Bessel function and integration subroutines gives rapid and accurate indication of intermodulation and harmonic distortion levels.

6. Acknowledgment

The authors are indebted to Mr. R. E. Scraton of the Department of Mathematics for some valuable suggestions.

7. References

1. Neill, T. B. M., 'Non-linear analysis of a balanced diode modulator', *Electronics Letters*, **6**, No. 5, pp. 125–128, March 1970.
2. Bennett, W. R., 'New results in the calculation of modulation products', *Bell Syst. Tech. J.*, **12**, pp. 228–243, 1933.
3. Belevitch, V., 'Non-linear effects in ring modulators', *Wireless Engineer*, **26**, p. 177, May 1949.
4. Belevitch, V., 'Non-linear effects in rectifier modulators', *Wireless Engineer*, **27**, pp. 130–1, April 1950.
5. Tucker, D. G., 'Intermodulation distortion in rectifier modulators', *Wireless Engineer*, **31**, pp. 145–52, June 1954.
6. Gardiner, J. G., 'The relationship between cross modulation and intermodulation distortions in the double-balanced mixer', *Proc. Inst. Elect. Electronics Engrs*, **56**, pp. 2069–71, November 1968.
7. Gardiner, J. G., 'An intermodulation phenomenon in the ring modulator', *The Radio and Electronic Engineer*, **39**, No. 4, pp. 193–197, April 1970.
8. Gardiner, J. G. and Yousif, A. M., 'Distortion performance of single-balanced diode modulators', *Proc. Instn Elect. Engrs*, **117**, No. 8, pp. 1609–14, August 1970.
9. Pilyagin, V. V., 'Theory of *N*-dimensional frequency converter with an ideal non-linear resistance', *Radio & Telecommunications Engng*, **29**, No. 8, pp. 37–42, 1966
10. Jones, J. J., 'Hard-limiting of two signals in random noise', *I.E.E.E. Trans. on Information Theory*, **IT-9**, pp. 34–42, January 1963.
11. Watson, G. N., 'A Treatise on the Theory of Bessel Functions' (Cambridge University Press, 1962).

Manuscript first received by the Institution on 23rd March 1970, and in revised form on 15th October 1970. (Paper No. 1360/CC95.)

U.D.C. 621.396.969.33:656.6.05

Electronic Engineering in the Solution to Harbour Approach Problems for Large Ships

By

T. W. WELCH,
C.Eng., F.I.E.R.E., M.Inst.Nav.†

Presented at a meeting of the Aerospace, Maritime and Military Systems Group in London on 18th March 1970.

The paper examines the range of electronic devices now coming into use, describes present methods of acquisition, display and dissemination of the data and suggests lines of future development. Outstanding problems to which no cost-effective solution has yet appeared are also discussed. The future problem is examined and a total system design approach suggested as most likely to provide effective operational usage of the information available in sophisticated installations.

1. Introduction

The introduction of ship-borne radar and hyperbolic position-fixing systems in the years immediately following the Second World War made it possible for ships to continue navigation in conditions of visibility which would have halted, or at best considerably delayed, their progress in similar conditions before the War. Because of this, harbour authorities soon found shipowners making it very clear that, having invested relatively large sums in the provision of ship-borne navigational aids, they expected some opportunity for their benefits to be retained in the terminal regions of voyages, i.e. in harbour authorities' areas. The accumulation of ships at Southend or at the Bar, for example, whilst awaiting conditions to clear sufficiently to enter harbour, would not be tolerated. Similarly, the owners and masters of ships, capable of continuing voyages once clear of the harbour authority area, did not take kindly to being 'held' in ports awaiting conditions suitable for them to be piloted out.

Some early efforts were directed at so improving the ships' own equipment as to make them suitable for the close-quarter work in harbour. Experiments were made with various means of providing the extreme resolution required; it was quickly shown, however, that this resolution would not be obtained with aerials of practical size other than by the use of the shortest available wavelength (Q-band, 8 mm) and at such wavelengths the heavy attenuation in rain and damp fog conditions made the apparatus unattractive. So did its price! Moreover, it was seen that merely providing a ship-borne radar of very high resolution was not going to solve the problems—much more information than that provided by any radar was required to give masters or pilots an adequate 'feel' for the situation when the effectiveness of their eyes was reduced by fog. Although work to improve the ships' own capability continued, there also grew up a certainty that good shore-based radar combined with radio-telephone communication would be more helpful.

The world's first true harbour radar system was set up in Liverpool in 1949. The fitting of v.h.f. radio-telephone in ships was at that time very little advanced, and communication was effected by way of shoulder-strap portable sets carried on board ships by their pilots.

† T. W. Welch & Partners Ltd., 64 Ash Hill Road, Ash, near Aldershot, Hampshire.

The successes of early systems soon indicated not only that this was a useful way forward but also that major ports would necessarily have to provide at least equal facilities, if they were not to lose valuable trade to competing ports. Thus began a series of developments and improvements in harbour information systems which is still going on today.

2. New Pressures

Current trends in shipping operations are bringing into use ships of very large size and of great capital value. Tankers of nearly a quarter of a million tons are already in operation and plans for ships of twice that size are well advanced. Larger bulk carriers are coming into use and specially designed, costly, container-freight ships are plying between even more costly container berths, frequently built and operated by the port authorities themselves. These trends bring several different, though closely related, pressures on both owners and harbour authorities. The enormous investments both in the ships and in the special terminals demand that the time spent by a ship in port should be minimized, if the invested capital is to show any reasonable return.

From the shipowner's viewpoint, the need is for more certainty than ever before that his ships will not lose time from restriction on port operations in bad weather, or from collision or stranding risks. Thus he will try to select, as operational bases, ports which are equipped to provide the necessary conditions for safe and expeditious entry, turn-round and departure.

The port authority also has its anxieties. Some of the new tankers already approach a quarter of a mile in length, which may be as much as three times the width of the fairway through which must pass *all* traffic using the port. The effects of the stranding of such a ship across the fairway would be to close the port entirely. No matter who was at fault in such a case, the effects of the consequent delays upon the usage, and hence upon the economy, of the port would be disastrous and could long outlive the period of the stranding itself, especially if other owners were thereby persuaded to shift their operations to other ports.

In some ports, situations can arise in which one large vessel, due to enter a lock at the seaward end of a large complex of docks, virtually blocks the fairway for considerable periods whilst manoeuvring off the lock

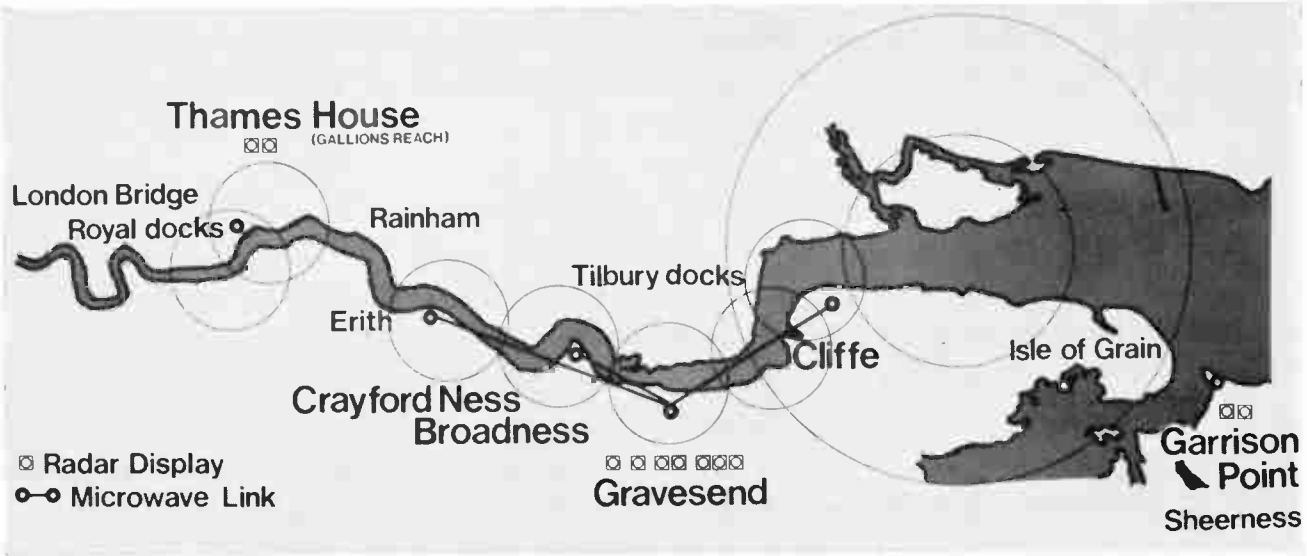
entrance. When, in such cases, a similarly large ship has to pass the area on passage to or from a dock further upriver, a major problem may arise in that the second ship has to be stopped in the fairway. When the length of the vessel exceeds the width of the fairway, a potentially dangerous situation exists.[2] Whatever aids can be applied to the avoidance of such situations will be welcomed by the port authorities. The public anxiety, only too keenly felt by all maritime interests, about the risks of an oil spillage from a tanker damaged in port approaches is too well known to need additional stress in this context.

All these considerations add to the growing interest in port information and planning systems of one kind and another. The problem is no longer merely one of providing aid in fog; the ability to keep ships moving in all conditions is today regarded as an aspect of port management no less important than wharves, warehouses, container facilities and port infrastructure.

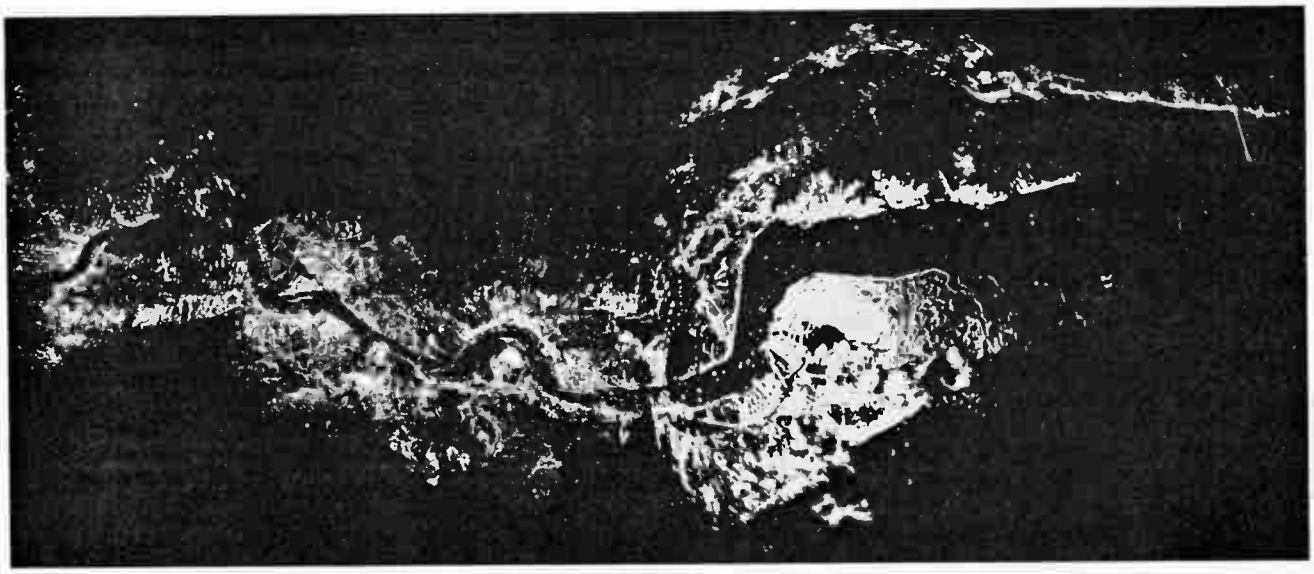
In parallel with this development, efforts to improve ship-borne equipment continue and considerable sums are being spent in research and development on ships' systems. It is the purpose of this paper, however, to discuss the development of shore-based services, the present position and the problems remaining to be solved.

3. Some Existing Systems

Within the United Kingdom, existing systems include those at Liverpool, Southampton,[3, 8] London, the Medway, Teesport and Great Yarmouth. Each is different in its organization from the others, though all rely basically on harbour radar systems and v.h.f. telephone communications. The most complex U.K. system is that at London, developed by the Port of London Authority under the title 'Thames Navigation Service' (Fig. 1). This now includes five radar stations, three of which are linked



(a) Radar coverage of the Port of London by the Thames Navigation Service.



(b) The River Thames from Gallion's Reach to Southend Pier. A composite picture from the five radars of the Thames Navigation Service. (*Decca Radar photograph.*)

Fig. 1.

by microwave and v.h.f. radio systems to an operations centre at Gravesend. This system, to which further reference is made later, has been almost continuously under development for over ten years and still further improvements are planned for the future. The original Liverpool system was first replaced about ten years ago and both it and the Southampton installation are currently in process of replacement and extension.

On the continent of Europe, systems already in operation include ones in the Elbe/Weser Rivers,[5] at Hamburg, Rotterdam,[6] Rostock and at Le Havre; several North American ports are equipped and one of the world's most advanced systems operates on the St. Lawrence Seaway. Several of the systems mentioned have been fully described in the literature and the list of references includes some examples.

The experience gained in the operation of the early systems has led to new recognition of requirements and to the identification of necessary lines of future development to cope with the new pressures noted above. Amongst the most important lessons learned is that, whilst most systems have been centred upon the employment of radar and radio telephone, the possession of these facilities does not, by itself, provide a solution to the operational problems. They must rather be included in a total system design, specifically oriented towards the provision of the proper total amount of information at each point in the port management complex where it is needed, in sufficient time for it to be of value. Electronic engineering plays a very major part in the acquisition, transmission, integration, filtering and dispersal of the information and will obviously continue to do so.

4. Current Requirements

The precise requirements at any port must depend on the nature of the approaches, the distribution of docks or other terminals in relation to the fairways, the degree of dependence upon tide and the numbers and types of traffic using the port.

In general, however, the requirements include:

Foreknowledge of the times and points of entry into, and departure from, the traffic pattern of as much as possible of the traffic. The 'warning time' may be anything from many weeks in the case of scheduled services to a very few minutes in the case of local operations.

Detailed knowledge of weather, tide and current conditions at all points within the port area, with reliable short-term forecasts of such matters as depth of water available at any point.

Fully up-dated knowledge of all obstructions, working operations and other hazards to, or limitations on, navigation within the port area.

Continuous surveillance of all traffic movements taking place in the port area, preferably with all traffic identified.

Short-term (tactical) forecasts of the interactions of traffic movements with one another or with the environmental conditions, aids to rapid formulation of advice to ships' masters or pilots, and means for passing such information to ships.

Presentation of integrated situation data to port control or port information officers.

Means to pass necessary information about traffic movements, particularly when these depart from announced intentions, to selected groups of addressees.

The equipment needed to satisfy these requirements may typically include:

Telephone and/or telex-links with dockmasters, agents, pilots, etc.

One or more radar head and aerial systems.

Remote-reading tide gauges.

Wind direction and speed indicators, possibly remote.

Remote current direction and speed indicators.

Remote visibility meters.

Television systems, local and remote.

Microwave links of wide bandwidth to carry radar or television video.

Display systems for all data at operations centre.

V.h.f. or u.h.f. links for narrow-band linking of command signals and wind, tide, current, etc. data from remote sites.

V.h.f. r.t. on internationally allocated marine channels for communication with ships, work-boats, patrol craft, etc.

5. Equipment Characteristics

5.1 *Communication Links*

The characteristics of the ship/shore v.h.f. radiotelephone and the telephone/telex information channels are the usual ones for such apparatus and require no further discussion. Careful siting of v.h.f. radio aerials is required to ensure good coverage free of blind spots in important regions, and it is sometimes necessary in large port areas to provide multiple remote stations.

5.2 *Radar Equipment*

The radar equipments used for harbour information services vary from simple ship-type equipment to highly specialized apparatus particularly developed for the role. Many, perhaps most, systems represent a compromise between these extremes, production-line units of ship systems being modified in relatively minor ways and combined with specially designed aerials, control systems and display ancillaries. This reduces the quite formidable cost of developing and producing entirely specialized systems and has consequent benefits for the user in initial cost and in assuring continuous availability of spare components and units, at prices appropriate to large-scale production.

Most harbour radar systems operate on special frequencies. In the U.K. these have been agreed between the G.P.O. and the Board of Trade so as to avoid mutual interference between the harbour equipment and shipborne or airborne radars, though within that part of X-band internationally allocated to the Radio Navigation Service. Aerials of up to 8 metres aperture are

typical, giving half-power beamwidths down to about $\frac{1}{3}°$. Pulse length is usually about 0·05 μs, at p.r.f.s around 2500 Hz; some installations include provision for switching to a longer pulse length of 0·1 or 0·15 μs where long-range performance is required. Receiver noise figures of 10 dB are typical, using balanced mixers. Bandwidths wide enough to accommodate the spectrum of the very short pulse are essential for preservation of the range discrimination.

Two complete transmitter-receiver systems are usually fitted, with remotely-controlled waveguide, video output and control line switching. The stations are designed for unattended operation and include reserve power supply arrangements brought into use automatically on failure of mains power. When high towers are used to avoid local shadowing, the radar transmitter-receivers are housed in a cab near the top of the tower to avoid both the attenuation and the range discrimination problems which arise with very long waveguide runs (Fig. 2(a)). A simple monitor p.p.i. and local control unit is usually fitted at remote stations to assist setting up and maintenance (Fig. 2(b)).

5.3 *Video Link Equipment*

The video link equipment associated with remote radar sites normally operates in the 7000 MHz band in the U.K. (Fig. 2(c)), though in some overseas installations it might be required by local regulations to be in the 4000 MHz band. Parabolic dish aerials of up to $3\frac{1}{2}$ m diameter are employed; the use of the wider apertures is not always merely a matter of obtaining additional loop gain, but is also aimed at reducing the risk of destructive interference by reflexion from obstructions a few degrees off the line of sight. Aerial heights required may considerably exceed the optimum height of the associated radar aerial, in order to ensure clearance of obstructions in the line of sight and of the Fresnel zone. Particular care must be taken to avoid interruptions of the beam by the passage of large ships and at the planning stage the range of movement of dockyard cranes may have to be taken into account. For long paths over tidal waters, vertical space diversity may be used to provide a sufficient signal margin at all states of the tide.

Long waveguide runs pose fewer problems for links than for radar and the transmitters and receivers are often fitted at ground level. The waveguide is sometimes pressurized with inert gas to prevent long-term deterioration within the guide. Duplicate channels are normally provided (Fig. 2(d)). Bandwidths approaching 20 MHz are required for the radar video; other information such as radar trigger, aerial rotational data, monitor signals indicating control states and a speech channel are often multiplexed with the video signals, but may sometimes be carried on a separate v.h.f. or u.h.f. link.

Command signals from the operations room for selecting the radar head, setting up the controls and other functions are passed on a v.h.f. or a u.h.f. link, again using duplicated equipment. In the U.K. the 460 MHz band has been used for the tele-command link in existing systems. It has been learned recently (March 1970), however, that this service will be shifted to the 1500 MHz band in future in cases where a bandwidth exceeding 7000 Hz is required. In one modern system a total loop integrity check is included, by routeing a test signal over the command-link, re-modulating it at the radar site on to the video link and re-transmitting it back to the operation centre. Alarms are signalled automatically when this integrity check reveals a failure in the link, following which simultaneous diagnostic and restoration routines can be fairly simply carried out by remote selection of duplicate units until the loop is restored.

Both microwave and v.h.f. or u.h.f. equipments are of conventional design for such systems, though the video link is necessarily somewhat modified from the forms developed for other purposes (e.g. television relay) because of the extreme demands of bandwidth.[14] Multiplexing equipment is also fairly conventional but it has been found helpful to develop special modems to ensure proper handling of the radar data.

5.4 *Tide Gauges*

Typically, tide gauges consist of a float moving within a restraining cylinder. In one remote reading version, the supporting cable is wound on a drum to which is geared a coder unit, which varies the mark space ratio of a train of tone-pulses as the drum rotates with up and down movements of the float. The tone-pulse data may be modulated on a v.h.f. or u.h.f. carrier for transmission to the operations room, where it is decoded to operate dial indicators and chart recorders. A 'datum' signal may be originated and transmitted by a separate tone whenever the tide reaches a pre-determined level, providing a system accuracy check twice in each tide.[11]

5.5 *Wind Force and Direction Indicators*

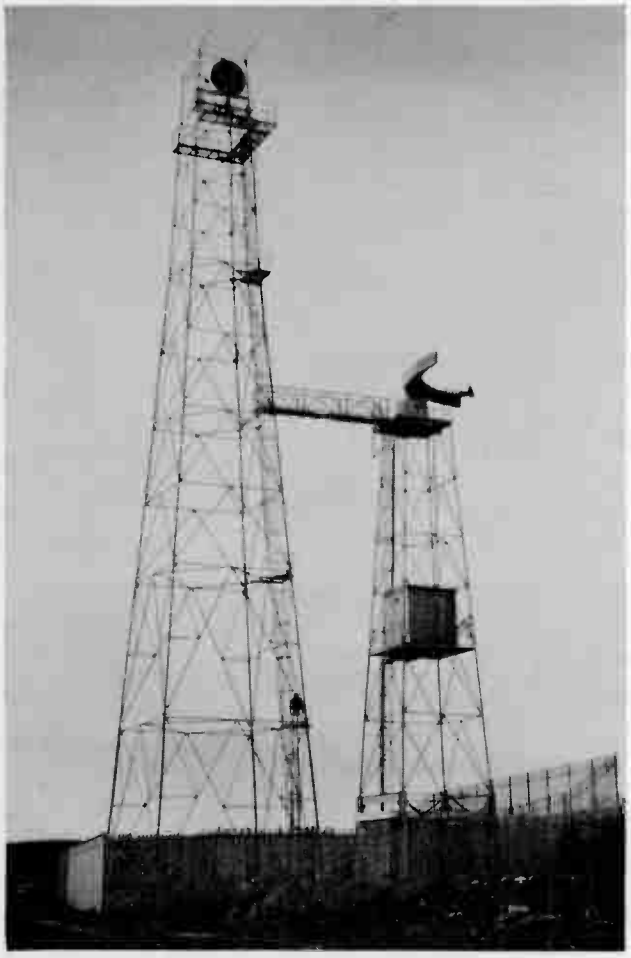
These are adaptations of conventional types used in meteorology. For remote indications over cable or radio links, analogue-to-digital convertors are added to both windspeed and direction shafts. Short-term oscillations in azimuth of the direction indicator are normally damped, but it is advantageous to preserve the fluctuations in wind speed so that at the receiving end both mean and gust-peak readings can be obtained. Dial indicators are commonly used in operations centres, chart recorders also being used in some instances for record-keeping purposes.

5.6 *Visibility Meters or Fog Detectors*

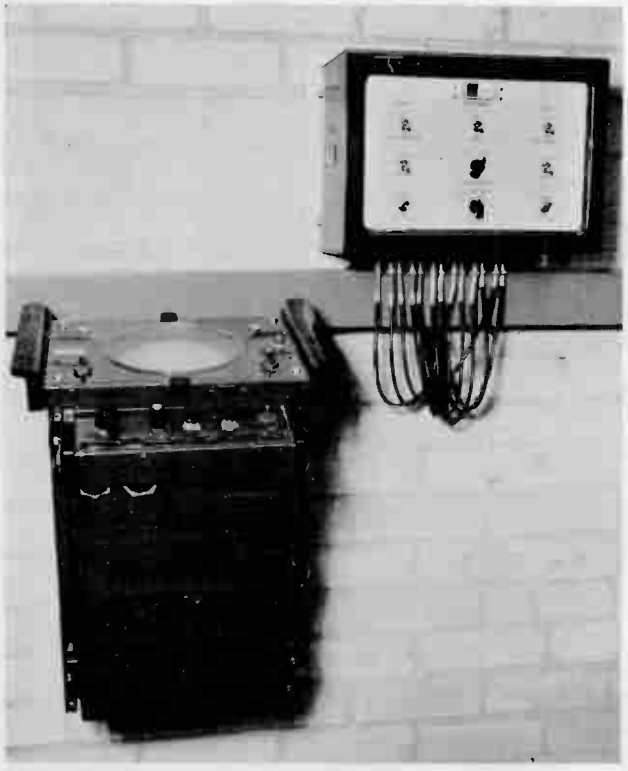
These instruments were originally developed for fully automatic control of unmanned light stations, radio-beacons, sound signals, etc. which may be required to be switched on at the onset of fog.

One version works by the detection of back-scatter energy from a modulated infra-red emission. The back-scatter is caused by the particles of condensed water vapour suspended in the air in fog conditions.

A threshold setting, normally calibrated in range, permits pre-setting the level at which a relay is closed by the received energy. For remote operation the relay may in turn be used to key a tone (or a d.c. level in the case of line connection) to a radio-frequency carrier. A signal is thus transmitted to the control centre whenever the visibility falls to or remains below the pre-set range.

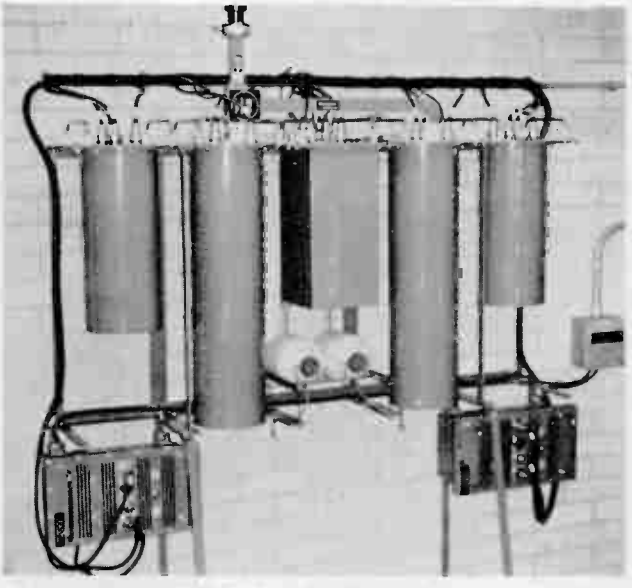


(a) Radar and video link aerials at Broadness.



(b) Remote station monitor display and command link servo unit.

(c) 7000 MHz receiving aerials at Gravesend, for radar data signals from Broadness and Crayfordness.



(d) Twin-channel 7000 MHz video link carrying radar data from Broadness to Gravesend.

Fig. 2. Radar and radio equipment of the Port of London Authority's Thames Navigation Service. *Decca Radar* (a) (b), *P.L.A.* (c) *and Ferranti* (d) *photographs.*

Determination of the actual range of visibility, up to a maximum of about three nautical miles, can be arranged by the installation of multiple sensors, each having its threshold set at a different level. Where the system layout allows, the visibility signal may be multiplexed on to one of the other radio-frequency carriers in a total harbour information system.

5.7 *Television Equipment*

Television systems, when used, are entirely conventional, with remote control of pan, tilt, aperture, and zoom. However, their 'exposure' (using the term in a meteorological context) often imposes severe demands on weather-proofing, both for camera equipment and for its panning mount. These demands, which are perhaps not peculiar to harbour services, call for special attention to the internal ventilation of cameras and prevention of condensation on lenses and windows. De-icing and 'windscreen wiper' arrangements for the viewing window are also required.

When remote radio-linking of television signals is required, separate link equipment is usually needed, because of the already severe demands on bandwidth of the link used for the radar. Monitors used in the operations room are usually entirely conventional.

Whilst the use of television for external viewing, as indicated above, has been specified for some harbour information systems, its use within the operations room complex for the local presentation of changing information to port control or information officers has not been exploited, though it may have applications in future.

6. Display Methods

6.1 *Radar Data Display*

The plan position (p.p.i.) display has to date remained the only satisfactory method of presenting the radar positional data to the control or information officers who use it.[7] The need to present remote areas to relatively large scales has led to the development of off-centring facilities, sometimes to much more than one radius, for both fixed-coil and rotating-coil types of display. As larger p.p.i. tubes have become available they have been adopted into harbour radar use to assist in the presentation of the information to suitably large natural scales. Although larger displays have been fitted in some installations, tubes of about 40 cm (16 in) diameter seem now to be preferred by most users, representing the best compromise between having to search a large area for information and achieving the desired open scale (Fig. 3).

In the past, rotating-coil displays have had advantages in cost and in simplicity of setting-up and maintenance. Fixed-coil displays have had advantages in the available degree of off-centring and in the provision of 'wandering interscan' measurement techniques. Currently, techniques are being developed which greatly extend the capability of rotating-coil displays whilst both the cost and stability of fixed-coil types have considerably improved.

Accurate measurements of the positions of ships, relative to one another or to navigational marks, were originally required to be made by direct methods on the display. To this end, techniques such as variable origin interscan lines were developed. However, measurements made in this way are often much less accurate than is the basic information available in the radar system of the range and bearing of each ship or navigational mark from the radar. Advanced techniques are available in which the range and bearing of one object from another are obtained by automatic computation, based on their respective positions relative to the radar. This method is both more accurate and quicker to use than the 'wandering interscan' and will probably find application in the harbour radar context as well as in other fields.

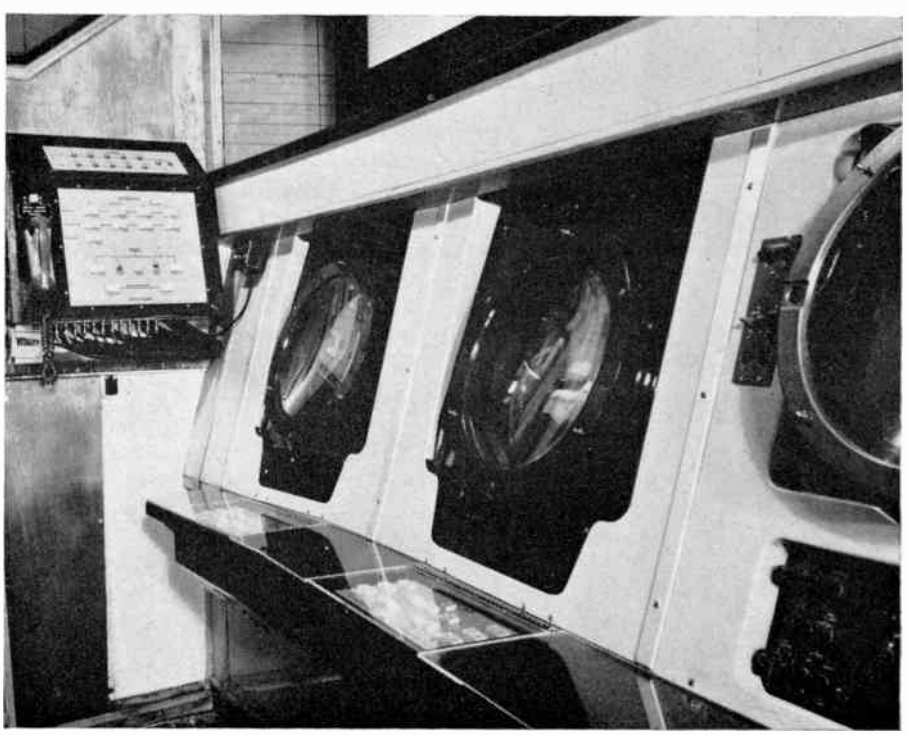


Fig. 3.
Modern 40 cm (16 in) p.p.i. displays and radar and link control units, PLA Thames Navigation Service, Gravesend (*Decca Radar photograph.*)

For many years, harbour authorities have fitted fairway buoys and beacons with corner reflectors to improve their radar responses on both ship and shore radars. In a crowded fairway, however, this has not always been sufficient to provide identification on the harbour radar display, and proposals have been made for the use of map techniques, either by engraving on the perspex implosion screens of the display tubes or by the electronic optical methods of the video map. These two devices, however, both produce some stability and registration problems. All points in the fairway can, of course, be defined by co-ordinates relative to the radar position, and represented by computer-generated point markers on the radar display, with inherently very high accuracy; this kind of technique is likely to be adopted for future systems.

6.2 *Operations Room Lighting*

Traditionally, port operations have been conducted from 'look-out' positions having extensive windows from which full views of the port approaches and manoeuvring areas could be commanded. In many installations the radar display system has been required to be incorporated within such a 'look-out' room, bringing the problem of viewing the p.p.i. displays in high ambient light.

Various solutions to this problem have been attempted. The simplest arrangement is the provision of blackout curtains or screens closely surrounding the radar displays but leaving free access to the remainder of the operations room.

Such arrangements have some disadvantage, in that officers obtaining information from the radar displays are isolated from other sources of important information, such as the tide gauges and the maps or charts on which the permanent information is plotted. Delays in appreciation of situations from ambient light adaption problems are also liable to occur.

High-brightness displays derived by television scan-conversion techniques have also been tried. Early equipment of this kind lacked the high resolution needed for the harbour radar function but equipment recently tested and shortly to be permanently installed at Le Havre shows a useful advance in this direction. The scan-conversion technique used in the trial had 1000-line resolution which appears from the photographs to be just adequate for the radar and display area shown; the radar in this case has a pulse length of 0·06 μs and a beamwidth of 1°. So far as is known the technique has not yet been applied to a radar of higher resolution (Fig. 4).

Another solution to this problem has been the provision of a nearly instantaneous photographic process, in which the radar p.p.i. picture is recorded on film, immediately processed and run directly through a back-projection system, to give a large photographic reproduction of the display area.[15] One advantage of this method, which is used at Teesport, is that the areas covered by two separate displays, even if they are derived from two separate radars, may be integrated. This imposes severe registration problems, however, on both the p.p.i. displays and the optics (Fig. 5).

In many modern installations, control officers are dealing with information derived by remote radars in areas of which no visual surveillance is possible from the operations centre. In such cases, there is little point in providing 'look-out' windows and the lighting problem is somewhat eased, though there remains the need for sufficient light to study charts and documents and to read the indicators of other data sources.

At the P.L.A.'s present centre at Gravesend, the operations room itself is brightly lit, and the radar displays are confined in a 'well' between the operational officers' desks and a large back-lit map display which carries the permanent information. This arrangement reduces the problem to some degree, though still leaving something to be desired in the full integration of data sources.

6.3 *Display of Other Data*

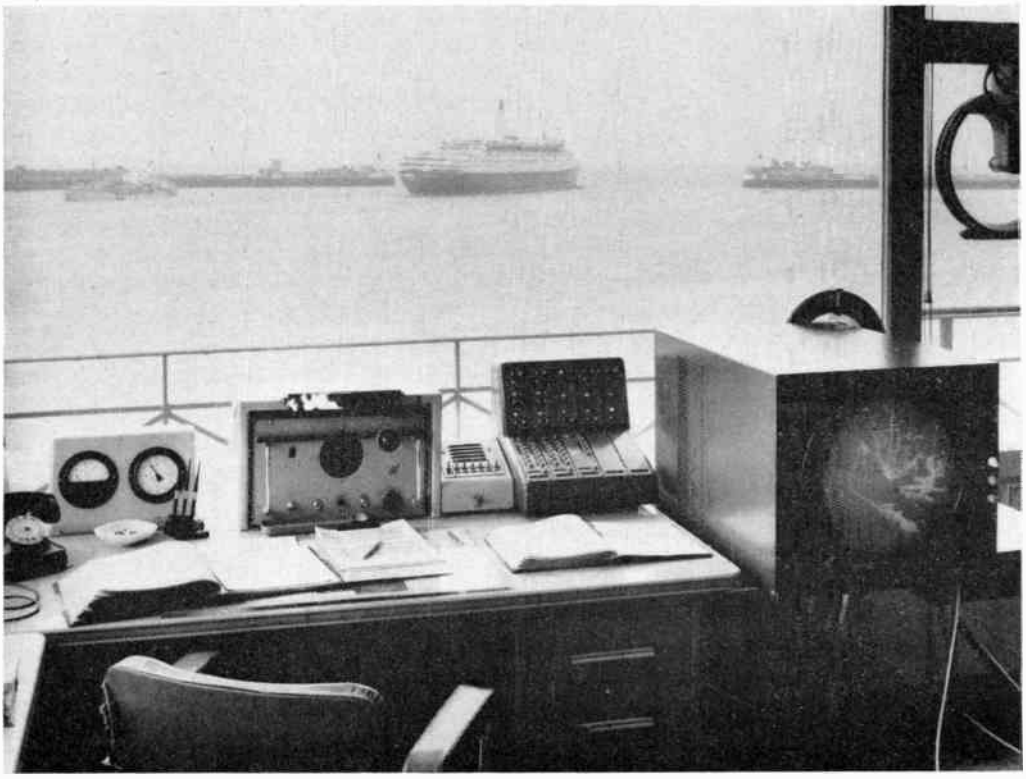
At present, display of the supporting information takes a variety of forms, but none is integrated directly with the display of the radar-derived information. Wind, tide and current indications are often read from 'dial' or 'thermometer' displays, whilst in some installations paper chart recorders constitute the only displays available to the operators.[10,11] These are necessarily installed in brightly-lit positions well separated from the radar displays.

The non-variable information (that is, non-variable during typical operating periods) is presented in various ways. At the P.L.A. Gravesend Centre the truly permanent information, such as the limits of the fairways, the positions of the banks, wharves, dock entrances and navigational marks is engraved on a back-lit map of the River. This map is also inscribed with chinagraph as to such matters as times of high and low water, dredging, survey work, wrecks or other navigational obstructions or hazards in the fairways, and the names and positions of ships moored at river berths.

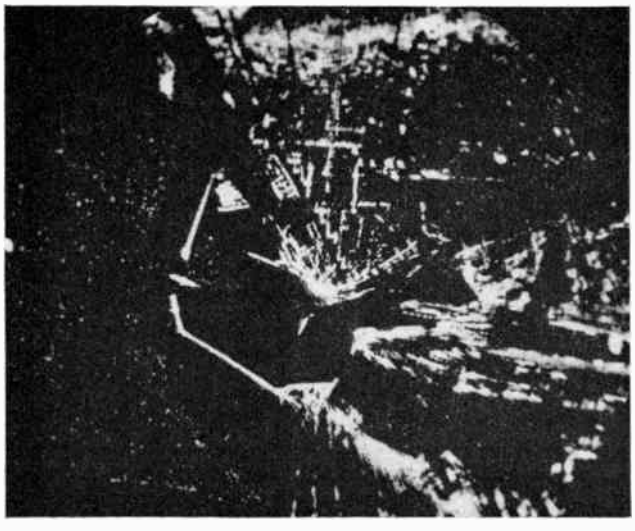
In other installations the permanent information is merely that carried on charts and plans fixed to the walls of the operations room (often where space can be found, rather than where they are operationally most useful) and information of a static but non-permanent nature (moored ships, workings, wrecks, etc.) either by pinned or magnetic labels on those charts or, in even less sophisticated systems, on handwritten lists kept by the control officers.

7. Unsolved Problems

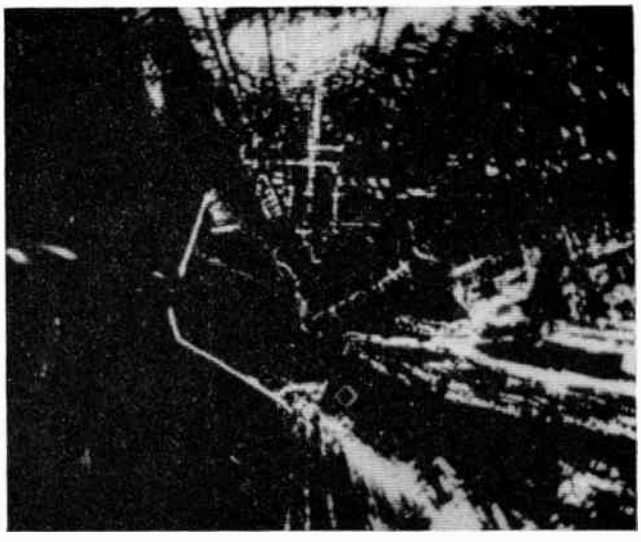
By far the most urgent of the problems remaining unsolved is that of maintaining a positive identity between ships under way within the port area and their echoes on the p.p.i. display. At present, this is achieved either by having an operator continuously tracking important vessels or by periodically attempting re-identification from the position or from the echo characteristics (size, brightness, speed, etc.). In the former case, the tracking operator cannot normally be the information officer himself, since the other duties of that office distract too greatly from the continuous attention necessary for positive tracking in busy waters. In the latter case, the dangers of a mis-identification are great because of the extreme way in which radar echo characteristics change with aspect, shadowing by other ships, etc.



(a) Operations room at Le Havre, showing experimental scan-conversion display.



(b) 'Raw' radar p.p.i. display.



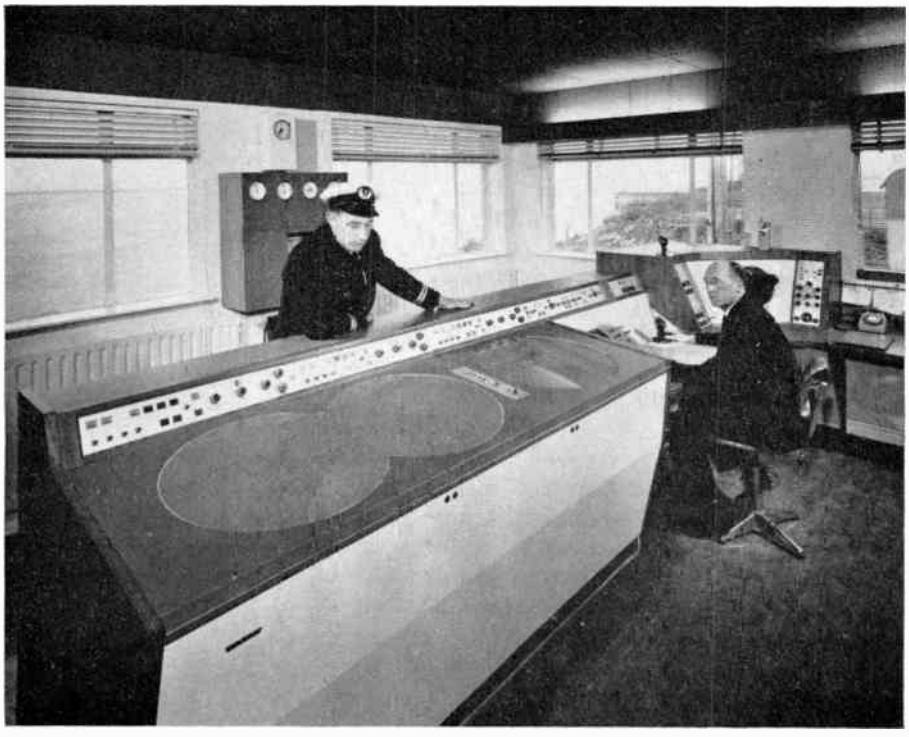
(c) 'Scan converted' p.p.i. display on standard television monitor

Fig. 4. The scan-conversion technique. (*Thomson CSF photographs.*)

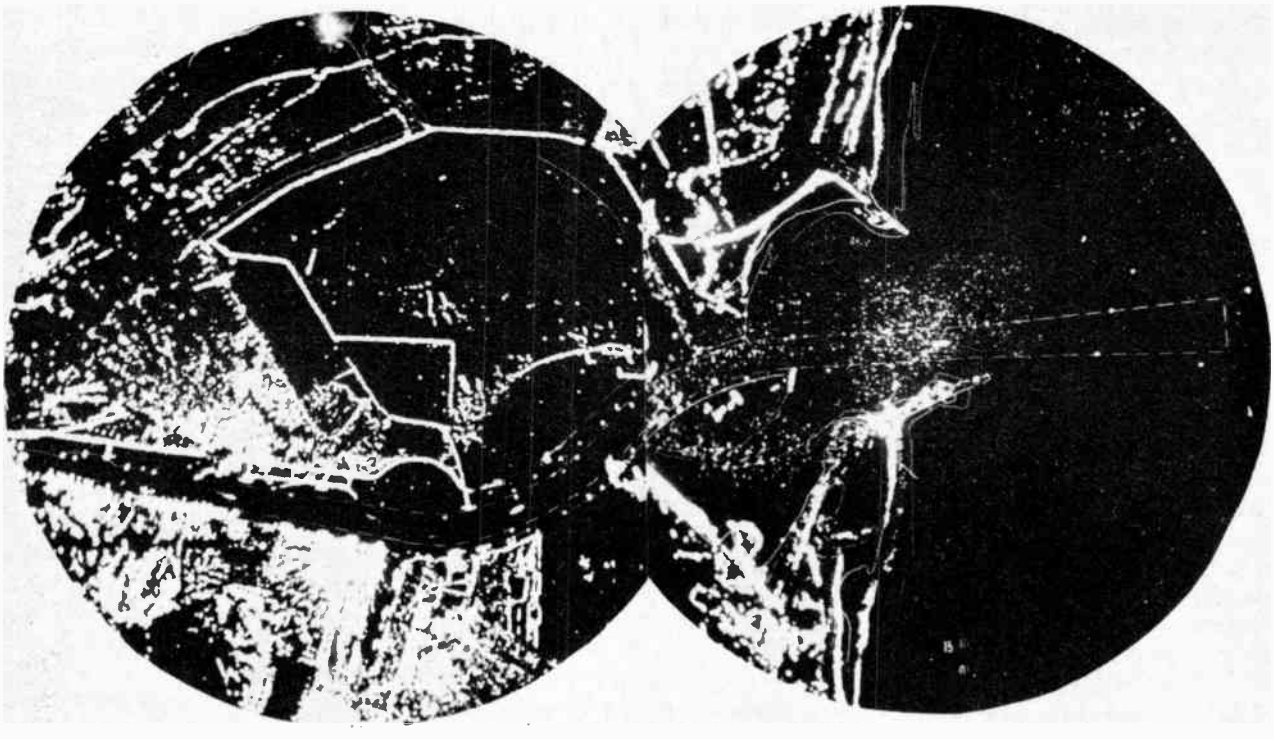
In a particularly straight-forward situation (from the information and control standpoint) such as the St. Lawrence Seaway, identity, once established, is relatively easy to maintain because ships proceed in an entirely ordered manner along the length of a single fairway. The same can never be true in a typical sea-port area where, especially in the outer approaches, ships overtake and pass one another and where traffic is bound to or from large numbers of different terminal points within the port complex.

From time to time, some form of secondary radar has been proposed for the solution of this problem and a new system has been described for the Hook-of-Holland/Rotterdam/Europort area which relies upon an X-band secondary radar system to provide a *two-way* data link between ship and shore.[17] It is intended that the equipment required in the ship shall be carried aboard by the pilot, the secondary radar aerial being hoisted by a signal halyard to an operational height. There are evident technical and operational problems with such a system but something of the kind will be required in the future.

An acceptable alternative might be found in a highly accurate local area navigation network, such as Decca Hi-Fix, and a data link to the shore operations centre.



(a) Operations room, Teesport harbour radar with 'Photo-plot' displays. (*Ferranti photograph.*)



(b) P.p.i. presentations from two separate radars (Tees Dock, left and South Gare, right), integrated on 'Photoplot' displays at Teesport Operations Centre. (*Kelvin-Hughes photograph.*)

Fig. 5.

Equipment akin to this has already been extensively developed for air traffic control use and could no doubt be applied readily to the port operations requirement. The data link itself might be an adaption and extension of the vehicle automatic state indicator (v.a.s.i.) developed for radio-telephone fitted vehicle fleets in police, ambulance and similar operations on shore. This system has the merit of requiring no alteration to the existing v.h.f. radio-telephone equipment, a small unit being plugged in between the microphone of that equipment and the normal microphone input jack. Voice communication remains unaffected. The equipment, which is fitted at Liverpool and in Guernsey for the control of road vehicles, has not yet been described in any formal paper.

By whatever means the 'identity' code is transmitted to the shore, provision should be made to include the other 'vital statistics' of length and draught of the reporting ship.

Another major problem lies in the relatively unsatisfactory nature of radar p.p.i. pictures. That these are acceptable at present merely reflects the basic fact that they represent the best that can be achieved in the current state of technology, at a cost which is deemed reasonable. From a good aerial position extremely realistic and informative plan data about the fairway and the land boundaries is obtainable. However, the lack of realism in the presentation of ship echoes, 'shadowing' of small ships by large ones, the uncertainty about just which parts of a large ship are producing the echo as seen on the display and the limitations of resolution with even the finest types of radar demand that an alternative means of display should be developed for the future.

To this particular problem no evidently acceptable solution is in sight. It might be hoped that, in the future, some form of dynamic map presentation will become possible, with ships and other objects represented true-to-scale; the radar data would then be used only to position the moving objects shown on the map.

Present display arrangements for the other data are often unsatisfactory, in that the many different sources of information which have to be consulted are not usually grouped in such a way as to bring them all simultaneously under the surveillance of the control officers. Further, the presentation of data in many different analogue forms (linear or circumferential displacement of a pointer or rotation of a drum), does not always lend itself to instant appreciation. In the St. Lawrence Seaway scheme, much of this information is presented digitally, on numeric indicators, at points where it falls naturally into place with other information required. This will, without doubt, become the normal practice for the future.

8. Uses of Computers

It is not at present entirely possible quickly to derive from the typical port information system the answers to some important operational questions. For example, it is becoming not uncommon for ships to enquire how much water there will be at a particular point at a future time of arrival there, in a time varying from perhaps half-an-hour to several hours. Conversely, a ship may announce its draught and ask at what earliest time it may expect to be able to negotiate a particular section of a tidal fairway.

Such questions, when asked today, are answered by *ad hoc* computation by the information officers, employing tide gauges or other appropriate sources of 'present' data, a graphical or tabular representation of tidal variations, a local knowledge of the probable influence of wind (taking account of its behaviour for several days past) and a slide-rule. The answers cannot come quickly—that they come within the typical short periods now taken, with a sufficient degree of accuracy, is a tribute to the usually very busy officers who supply them.

Port authorities are no strangers to the world of computers, which are now widely applied to other aspects of port management such as the scheduling of container terminal operations. It is therefore to be expected that they will soon be specifying computerized solutions to the operational problems of their information services and some of the possible applications are discussed below.

The computer may also in the future assist the communications problem. It is not uncommon in adverse weather conditions, just when the port officers are at their busiest, analysing incoming data and passing information and advice to ships navigating in the harbour area, for messages to arrive amending the schedules of several ships in quick succession. Each such message may require that the new schedule be passed to groups of up to six or seven addressees, of whom perhaps only two or three are common. At present, except for the relatively few cases in which telex facilities are available, this must be done by successive telephone calls, demanding almost the full attention for a considerable period of one or more of the officers. Extension of teletypewriter facilities amongst the likely addressees and computer-controlled message formatting, routeing and acknowledgment, will in future do very much to relieve this problem.

9. Systems Design for the Future

Most port information services which exist today have grown up piecemeal, with facilities added as the need arose or as they became available. They have been primarily aimed at providing tactical solutions to operational problems at the moment of detection. In the future, much more pre-planning of traffic movement, both long-term (several weeks ahead) and short-term (a few minutes ahead) will be undertaken and the information systems will need to be designed to cater for this change. Much of the planning can and will be done by computers, but it will be essential to retain the judgement of skilled seamen, as information officers are required to be, in the solution of immediate tactical situations. The computers, therefore, will also be used to organize the presentation of tactical information in the ways best suited to aid the officers in making measurements and appreciations, in performing calculations and in reaching sound decisions swiftly.

This is not merely a hope for the distant future. The new techniques of speedy, accurate measurement and position-definition on p.p.i. displays, mentioned above, require computers for their operation, albeit quite small and relatively inexpensive ones. By using devices only a

little more capable, in speed and capacity, than are needed for just the p.p.i. organization, additional calculations will be possible. Time and position of closest approach of a ship to a buoy or beacon, calculations relating to depth of water at any designated position within the area and the time at which a narrow fairway will be cleared by one large vessel so that another proceeding in the opposite direction may be admitted, are all relatively simple examples of extension of the display computer capability at almost negligible extra cost.

Display methods themselves will improve in the long-term. At present, the only foreseeable satisfactory display will be one in which the permanent features of the harbour area are shown on one or more large-scale maps and the movements of ships indicated in true scale, with identity and other relevant information, and with depth of water, wind and visibility continuously displayed at salient points along the fairways. A long approach, such as that of the Thames, may need several overlapping 'sector' maps to cover the whole area on increasingly open scales as the fairways become narrower. Computer monitoring of the progress of traffic will set off alarms whenever an undesirable situation is 'seen' to be developing and tactical action is needed. Video terminals and keyboards will enable the information officers to utilize the capability of the computer in reaching decisions regarding that action, or to obtain answers to questions posed by ships moving within the area. Stereo-typed message formats and computer-routeing will permit the necessary information about ship movements and service requirements to be passed quickly and automatically to those needing it. Relief of radio-telephone channel loading will be brought about by digital data exchanges between ship and shore.

In some ways, the problems are analogous to those of the air traffic control situation in terminal areas. Great caution is necessary, however, in any tendency to assume that any particular a.t.c. technique is applicable to the port operations context. The problems are different, traffic loadings and patterns are different, freedom of manoeuvre is relatively restricted, range and azimuth 'cells' for data storage are many times smaller and therefore may be much more numerous in a typical port area. Moreover, although some of the problems and situations are common to all ports, many others are specific to particular areas, with their own specialized types of traffic and their own particular distribution of fairways and terminals. Again, whilst it may be possible to designate some fairways as reserved for 'scheduled traffic' in 'instrument conditions' it seems unlikely that any organization of 'seaspace' similar to that of 'airspace' would ever be generally acceptable or enforceable.

Every future installation, therefore, will need to be the subject of its own systems study, its operational needs precisely defined and the equipment performance exactly specified. Much more attention will have to be paid to bringing together the designers of the various kinds of data-collecting apparatus, from radar to current-meters, so that true integration is achieved. Operational techniques and routines much in advance of present ones will have to be developed, though they must continue to be based on the practices of good seamanship.

In all these activities, electronic engineers from many branches of the industry will need to work closely with port operations staff and system designers to ensure that adequate but not over-designed systems emerge, with the proper capability for expansion or extension as the port grows or new techniques are developed.

10. Effect of Ship Problems

It must never be forgotten by the systems designers that, no matter how good the flow of data from the shore may be, the final success or failure of ship operations will depend upon the 'feel' the master or pilot has for the particular situation of his ship at each moment of time. Totally blind landing of aircraft, without pilot intervention, is already in operation but it seems likely to be a very long time indeed before automatic operation of ships, particularly as to manoeuvres in narrow waters and final berthing, becomes possible. Because this is so, much parallel work is going on to improve ship-borne methods of providing the 'feel' needed on the bridge. The tremendous part played (and to be played) by electronic engineering in that field is another story.

11. Electronics Applied to Older Navigation Aids

Finally, it must not be overlooked that, amongst all the sophisticated activity of electronic engineers in modern maritime systems, electronics is also playing a valuable part in improving the operation of older and more traditional navigational aids. Electronic flasher units have been developed, for example, to provide the identification characteristic for an electrically-lit buoy or beacon. This device is saving a great deal of electrical power and of maintenance effort compared with older motor-driven commutating devices, an important contribution for battery-operated lights in inaccessible places. Remote control and monitoring of lighthouses, foghorns and similar coastal or off-shore navigational aids is now almost commonplace and new electronic devices to solve both old and new problems emerge regularly.[9]

12. Conclusion

This paper has attempted to review the part played by electronics in the service of port management. The references will supply much of the detail of equipment design which has necessarily been omitted. The subject is one of growing interest and importance throughout the world, since the economies of many countries, developed and developing, depends so largely on efficient transport systems. It is hoped that the paper has shown that electronic engineers do not think only in isolated parochial terms about narrow equipment specializations but can also take the broad, system-design approach to the study, understanding and solution of the problems facing those it is their professional privilege to serve.

13. Acknowledgments

The ready assistance given by the under-mentioned in the preparation of this paper, and the provision of illustrations, is gratefully acknowledged:

Cdr. R. B. Richardson, Harbour Master, P.L.A.;
Mr. J. Rees, Port Telecommunications Officer, P.L.A.;

Capt. J. Andrew, Deputy Dock & Harbour Master, Southampton; Cdr. Knight, Mersey Docks & Harbour Board; Mr. N. S. Kirby, AGA (UK) Ltd.; Mr. A. Harrison, Kelvin Hughes Division, S. Smith & Sons; Mr. G. Wanless, Ferranti Ltd.; Mr. A. P. Tuthill, Decca Radar Ltd.; Mr. G. Hyatt, Radio Services Division, Post Office Telecommunications Headquarters; M. B. Cambier, Thomson-CSF, France.

14. References and Bibliography

The extensive literature on particular aspects of the port control and information problem, or describing equipment and systems used in that service, includes the following:

1. Wylie, F. J. (Ed.), 'The Use of Radar at Sea' (Hollis and Carter, London, 1952).
2. Tani, H., 'The reverse stopping ability of supertankers', *J. Inst. Nav.*, **21**, p. 119, April 1968.
3. Andrew, J., 'The port arrival and departure of ships', *J. Inst. Nav.*, **14**, p. 22, January 1961.
4. Webster, E. M., 'A Proposed Program for Maritime Research' RTCM Symposium, San Francisco, April 1961.
5. Hilke, Otto, 'Safety Radar for Elbe and Weser Rivers', RTCM Symposium, San Francisco, April 1961.
6. Frijlink, C. H., 'Some Aspects of Waterway Radar', RTCM Symposium, San Francisco, April 1961.
7. Benjamin, R., 'Man and machine in the extraction and use of radar information', *The Radio and Electronic Engineer*, **26**, No. 4, pp. 309–16, October 1963.
8. Andrew, J., 'The Southampton Port Operation and Information. Service', VIIth IALA Conference, Rome, May 1965.
9. MacKellar, A. C., 'The remote control of lighthouses and beacons', *The Radio and Electronic Engineer*, **34**, No. 3, pp. 175–82, September 1967.
10. Woods, A. J. and MacMillan, D. H., 'A new development in current meters', *Dock and Harbour Authority*, November 1959.
11. Balestrini, P., 'PLA's tide gauge system', *Dock and Harbour Authority*, May 1968.
12. Oudet, L., 'The crisis in the increase of tonnage', *J. Inst. Nav.*, **21**, p. 305, July 1968.
13. Graham, P. W. W., 'Operational aspects of v.h.f. communication and radar surveillance by port operations centres', *The Radio and Electronic Engineer*, **36**, No. 3, pp. 149–52, September 1968.
14. Wanless, G., 'Microwave-link characteristics for a harbour radar surveillance system', *The Radio and Electronic Engineer*, **36**, No. 3, pp. 153–60, September 1968.
15. Harrison, A., 'A display centre for harbour surveillance and control', *The Radio and Electronic Engineer*, **36**, No. 3, pp. 161–9, September 1968.
16. Richardson, R. B., 'Some profiles for the future in coastal and port approaches', *J. Inst. Nav.*, **21**, p. 465, October 1968.
17. Schimmel, N., 'Safety of Shipping in Harbours and Waterways: Control and Surveillance'. Norwegian Navigation Conference, 1969.

Manuscript received by the Institution on 15th September 1970.
(Paper No. 1361/AMMS35.)

STANDARD FREQUENCY TRANSMISSIONS—December 1970

(Communication from the National Physical Laboratory)

Dec. 1970	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)		Dec. 1970	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	— 300·2	0	0	671	623·5	17	— 300·1	0	+ 0·1	675	627·9
2	— 300·0	0	+ 0·1	671	623·7	18	— 300·0	0	+ 0·1	675	628·1
3	— 299·9	0	+ 0·1	670	623·5	19	— 300·0	0	+ 0·1	675	627·9
4	— 300·0	— 0·1	+ 0·1	670	624·3	20	— 300·0	0	+ 0·1	675	627·7
5	— 299·9	0	+ 0·1	669	624·1	21	— 299·9	0	+ 0·1	674	627·8
6	— 300·2	— 0·1	+ 0·1	671	624·7	22	— 300·0	0	+ 0·1	674	628·2
7	— 300·1	0	+ 0·1	672	624·4	23	— 300·0	— 0·1	+ 0·1	674	628·7
8	— 300·0	0	+ 0·1	672	624·9	24	— 300·0	— 0·1	+ 0·1	674	629·7
9	— 300·0	0	+ 0·1	672	624·9	25	— 299·8	— 0·1	+ 0·1	672	630·4
10	— 300·1	+ 0·1	+ 0·1	673	624·3	26	— 298·9	— 0·1	0	661	631·1
11	— 299·9	0	+ 0·1	672	624·7	27	— 299·8	— 0·1	0	659	632·2
12	— 300·1	— 0·1	+ 0·1	673	625·7	28	— 299·4	0	0	657	632·3
13	— 300·0	0	+ 0·1	673	625·9	29	— 300·2	0	0	651	630·5
14	— 300·0	0	+ 0·1	673	626·1	30	— 300·2	— 0·1	0	653	631·2
15	— 300·0	— 0·2	+ 0·1	673	628·3	31	— 300·0	— 0·2	0	653	635·3
16	— 300·1	+ 0·1	+ 0·1	674	627·7						

Note: The frequency offset for 1971 will be -300×10^{-10}.

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.

CONFERENCE REPORT

Laboratory Automation

London, 10th to 12th November, 1970

Automation is usually thought of as being the concern of industry, in the processing of material, or in manufacture, rather than the laboratory where information is the end-product. In industrial fields, its successful application had been uneven according to Dr. I. Maddock, Controller Industrial Technology at the Department of Trade and Industry, who opened the Conference. He regretted that the application of automation techniques in the laboratory had been even slower for he believed that it could provide opportunities to rethink a whole procedure. The need to measure indirectly what had previously always been measured directly called for careful consideration of what was being measured and why, so that the end result was meaningful.

Dr. Maddock drew attention to the fact that about two-thirds of the papers came from organizations which were government-funded and he hoped that industry would not be too slow in further exploitation of the undeniable richness of innovating ability and research and development capability which existed in this country. The apparent reluctance of industry to take up new ideas and its frequent alacrity in exporting advanced equipment at the expense of building up the important solid home markets were dangers to be guarded against. However, he also warned industry against being too carried away by the cleverness of a subject, again at the expense of marketing its products.

The thirty papers which were presented in the following three days were notable for the wide range of subjects covered under the deliberately comprehensive title chosen for what is believed to have been the first conference with this theme. The full programme of the Conference was published in the October issue of the *Journal* and it is intended to reprint a selection of papers from the Conference Proceedings in forthcoming issues.†

The Conference papers were reviewed and a thought-provoking postscript given on the lessons which could be learned, by the Chairman of the Joint Organizing Committee, Mr. G. S. Evans, in a closing address. After humorously referring to the difficulties of trying to summarize a conference of this kind, he continued:

'We have covered a considerable amount of ground in the last three days and probably this Conference is unique in terms of the width of the field that we have covered. We have seen glimpses of the worlds of the chemist, the physicist, various engineering disciplines, and the manufacturer. Perhaps the thought that has passed through my mind once or twice is that "I wish he wouldn't assume that I knew so much about his job, and I wish that he wouldn't assume that he knows so much about mine". One picture which has been kept before me throughout is shown opposite.

'The diagram can be debated in detail, and one can re-draw it in many ways, but using this diagram, there are one or two things that echo back. I looked particularly to see if speakers had really defined their objectives, really fixed their cost targets when they started, and then looked to see if they have achieved them. This has not always been easy to see.

'Mr. De Vries[1] made the point that his objective was to reduce the wasted time of the analyst rather than to increase his throughput. Mr. Gilford[2] made, I thought, a very good point when he said that if you really look at your objectives, you sometimes just end up with improving your manual methods or your re-scheduling.

'Perhaps the diagram most nearly suited Mr. Collyer's paper.[3] However, a quick cost estimate on his system which he bought for £31 000, shows by my reckoning that the supplier was very lucky if he managed to keep his works cost down to that figure.

'The thing that struck me most during the Conference was how many schemes were apparently evolved by hammering the overheads. We have seen some examples of very highly qualified, competent, experienced people in their own field who have, dare I say, dabbled in systems engineering, electronics and programming activities. Not only have they managed to get away with it, but they have found their financial structures to be such that it was easier to fiddle a do-it-yourself job on the overheads than have it done professionally by systems specialists. I suggest that this can lead to methods and techniques which may be well within their limited knowledge, but they are not necessarily the best techniques available. Perhaps research environments are more guilty of this than others, and I wonder sometimes what it is costing the country in squandered expertise which should more rightly be devoted to defining and evaluating. How often do we see such dabbling resulting in the inevitable bizarre spectacle of the innumerable set zeros, obsolete semiconductors, coupled with costly C-core or oil-filled transformers. Or alternatively, cascades of paper-tape punches and readers feeding on each other like parasites. This sort of thing, I believe, is fostered by financial control systems which don't allow realistic appraisal of the true in-house costs. I wonder if there is a clue here to Dr. Maddock's comment that two-thirds of the contributions were from Government Departments, and only one-third from Industry.

Were objectives defined at the outset?

↓

Was a cost target set to cover the completion of the work?

↓

Were costs collected in a realistic manner?

↓

Were the objectives met?

↓

Did the system do more than merely replace the existing method. If so, in what way?

↓

Was the project cost effective?

↓

Was the final actual cost known and how did this compare with the target?

---

† The complete set of Conference papers is available in a single volume from the Publications Department, I.E.R.E., 8–9 Bedford Square, London WC1B 3RG, price £6.00 (6) post free.

'The above philosophy is probably most pronounced—and it has certainly shown up at this Conference—in references to software. I think that in this sphere there are some who are perhaps wiser now than they were early on Tuesday morning, but I sense that there are some who are more confused, and perhaps more disillusioned. To an engineer, for example, and this is an engineering Institution, such terms as program failure, program reliability and program development that we have heard, are anathema.

'To begin with, programs don't fail, they have errors. I sometimes wonder if less reputable software houses introduce the term to reduce their responsibility to the client, or if the computer manufacturer wished to mask their hardware shortcomings. If a program is right it will not fail, which is the beauty of the general-purpose computer approach anyway. That is not to say, however, that a program may not contain latent errors, which, particularly with a complex system, may only show themselves after a considerable time, and also as Mr. Wade pointed out,[4] later changes to program or peripherals have to be watched carefully. I believe that this stresses the need for professional, competent, and above all, experienced programming, rather than a "do-it-yourself-quick-programming-course-and-anyone-can-do-it-providing-it's-on-the-overheads" approach. This may work for the simplest system, and I think we have seen some examples where it certainly has, but it is quite a different matter when the program has different priority levels and is on a real-time basis, or where teams of programmers are involved. You may have to pay for experience, but you pay a lot more for the outcome of inexperience.

'Next, "program development", heaven forbid! Program definition yes, put that on the overheads if you must, but then let the professionals do a fixed price job on the program production. You may argue, "but we don't always know what we want when we start". My answer to that is, "boy, you've got a problem!" Did you notice the number of times that hardware costs were given throughout the Conference? The software was, I think, always stated in terms of man-months, sometimes with no assurance that the work had really been completed. I would suggest that program times alone are meaningless. I have met ratios of 10 : 1 in time scales for rates of production of fully de-bugged programs, depending on who did it, how many times it was done, how many times programmers had left the company during the project, the word length, the store occupation efficiency, the program complexity, the calibre of the program management control, and particularly the reliability and availability of the system during the program test period. The computer itself may be extremely reliable, but when you come to put all your programs together with all its peripherals, sometimes newly developed peripherals, then you can be in trouble.

'It is my firm hope that engineering institutions will soon actively recognize this form of software for what it is, an engineering function, and give a lead in the field. References during the Conference to the need for high-level engineering based languages support this view. During informal discussions there have been requests made to me for exchanges on this subject. I think personally we should support this, but have several reservations. I don't think the high-level language is the answer to all problems, and there will always be many applications where we just cannot afford the luxury.

'At the outset of the Conference, the hope was expressed that we should break down the word barrier, and some of us indeed have learned a number of new words in the last three days, and I hope that we are better people for it. The bouquet for the word of the Conference must surely go to "automation overkill" (or was it "technical overkill"?)!

'In many ways I think this has been an extremely useful and a good Conference. Perhaps there are two highlights. One is the question of CAMAC. The concept of CAMAC as a stand-alone, yet computer-compatible data logger must surely merit close attention, even if as we were warned, the standard CAMACs are different! Another highlight of the Conference was perhaps the way in which A.W.R.E. presented their case—I think it was described as the A.W.R.E. onslaught—and how, rather than be pre-occupied with tomorrow's computer, they have got on with the job and have used one which was available.

'Finally, it has been mentioned that the Conference appeared to be pre-occupied with computers, and speakers have asked more than once "why the G.P. computer approach?" I think that this needs very careful and unprejudiced consideration, and of course, the answer is . . . because it was the right way to do it. Perhaps one last point which I think we should bear in mind is that these systems take a fair time to develop, the technology is growing very rapidly and what we have seen are really the concepts that were thought up perhaps two, three or four years ago. From the speed with which the minicomputer is evolving, and as seen in Mr. Marson's paper,[5] the rate at which the price is dropping, I am sure many of the authors would not start the same way if they had the chance right now.'

References

1. De Vries, G. W. and Verhoeven, L. S., 'Automation of routine analyses in works laboratories', pp. 47–62.
2. Gilford, S. R., 'Productivity in the analytical laboratory—a rational approach to automation', pp. 401–9.
3. Collyer, L. M., Hawkins, L. H. C. and Thomson, G. H., 'Some hardware aspects of computer-aided gas chromatography', pp. 181–95.
4. Wade, B. O., 'Exploiting small computers for on-line applications', pp. 411–20.
5. Marson, G. B., 'Computer control of laboratory experiments and test rigs', pp. 365–86.

Proceedings of Conference on Laboratory Automation, 1970.
(I.E.R.E. Conference Proceedings No. 20)

U.D.C. 621.317: 621–52

The Economics of Operator Participation in Automatic Testing

By
R. CROSHER,
B.Sc.(Eng.)†

Reprinted from the Proceedings of the Conference on Automatic Test Equipment held in Birmingham from 10th to 14th April 1970.

After briefly considering some of the factors which should be considered when installing automatic test equipment, the paper gives detailed consideration to ways in which the degree and nature of operator participation affects the cost of operation. The area of machine interface is significant and a field test console developed for testing of Army radio equipment is quoted as an example where a well designed interface significantly reduces the overall test time, lowers the level of skill required to operate the equipment and generally increases the integrity of testing.

1. Introduction

For some time now certain products have been obvious subjects for automatic testing. The testing procedure was a straightforward routine and there was a large throughput of units to be tested. Economic justification was immediately evident and little effort was required to prove this in financial terms. Slowly it is being realized that more and more complicated test routines can be automated, but often with the added complication that it is more difficult to justify in economic terms. As a test routine becomes more complicated, so inevitably does the test system.

In order to increase the viability of applying automatic test equipment to more sophisticated test procedures, it is clearly necessary to investigate ways of decreasing testing time, improving test integrity, reducing numbers of skilled operators required, improving documentation and records, etc. In this paper it is the intention to elaborate on some of the ways in which overall test time may be reduced by examining the role of the operator in testing relatively complex units. Before it is possible to evaluate the most suitable method of test for any given unit-under-test it is necessary to understand the factors which can influence operator participation. Decisions taken with respect to these factors will define the type of operator that will be required to use the automatic test equipment which in turn will influence the cost of running the installation.

The prime objective of a systems engineer when laying down the basis for an automatic test system is to meet the test requirement of the proposed unit-under-test, bearing in mind that the system chosen should not only be economically viable in initial capital outlay but also cost effective in operation. Besides the obvious necessity to meet the test specification, the most important single factor in this assessment is the cost of labour. Therefore, it is in the interests of the systems engineer to keep the actual operation of the equipment as simple as possible in order to be able to use the cheapest labour rate. This is known as 'de-skilling' the test operation.

For the purposes of this paper it is assumed that an unskilled operator is one who has no knowledge of either the unit-under-test or the test routine used to test it. A skilled operator is taken to be someone who has had some education in electrical engineering, understands the unit-under-test and its test routine and is therefore capable of interpreting test results intelligently.

2. Factors Affecting Operator Participation

Five main factors affect the degree of operator participation. They are:

(i) type and nature of unit-under-test,
(ii) nature of tests and type of test routine,
(iii) type of labour available,
(iv) design of the automatic test equipment,
(v) design of machine interfaces.

A derivation relating the overall test time for any given unit-under-test and the degree of operator participation is given in Appendix 1.

2.1. *Type and Nature of the Unit-under-test*

An automatic test equipment may be used for testing a wide variety of electrical equipment. This may vary from simple continuity checking of cable looms to complex testing of advanced avionics systems. More obvious subjects for testing are products that are manufactured or repaired in quantity, thereby providing a more continuous flow of test subjects; printed circuit boards, plug-in modules, radio equipment, domestic radio and television receivers, are some typical examples.

The type of unit is going to dictate whether test points are easily accessible and can be connected to the automatic tester by a simple operation, e.g. an edge connector on a printed circuit board. For some test routines an edge connector may provide sufficient test points, but for others further test points may be required. In the latter case some kind of jig arrangement may be required to provide probe contacts on to the copper track or soldered joints.

Thus the potential user of an automatic tester has a choice;

(i) either he keeps test routines down to a minimum by employing easily accessible test points only, or
(ii) he makes greater use of the automatic test equipment capability by making more elaborate jigs and test programs, thereby enabling much more thorough testing of the unit-under-test.

† Test Systems Division, Honeywell Ltd., Hemel Hempstead, Hertfordshire.

2.2. *Nature of the Tests and Type of Test Routine*

Inherent in writing a test specification for an automatic test equipment will be the type of testing to be performed on the unit-under-test. Testing falls into three basic categories, namely functional testing, diagnostic testing and alignment testing.

In the case of *functional testing*, it is required to know whether the unit-under-test performs within the specification without attempting to find out what may have gone wrong should it fail to meet the specified tests. In this case operator participation merely entails connecting and disconnecting the unit-under-test. The overall test time is usually relatively short. Routines which consist of plain functional testing are usually the easiest to which unskilled labour can be applied.

Diagnostic testing is where it is required to ascertain a faulty module or component when functional testing has shown that a fault exists. In this instance, if unskilled or semi-skilled labour is to be used, it may be necessary to provide detailed instructions to the operator or diagnostic charts so that he can ascertain the faulty component. If, however, a skilled operator is employed then these instructions are unnecessary and the operator may locate the faulty component by virtue of his knowledge of both the unit-under-test and the test routine. The usual procedure is to combine these two alternatives. An unskilled operator is used on the automatic test equipment who then passes faulty units only together with their print-outs to the skilled man. In this way the more highly paid skilled man is employing his skills for a much greater proportion of his time, and not spending time doing the simple manual operations associated with testing an electrical circuit.

Alignment testing, on the other hand, is where the automatic test equipment is used to assist the operator in adjusting variables in the unit-under-test in order that it should meet a given specification. In this instance it may be necessary and practical to employ the skilled man, i.e. test engineer, as the operator of the automatic test equipment. He can employ his knowledge to manipulate the tester to suit the alignment having to be carried out. This is especially true where the alignment necessitates the adjustment of a number of inter-dependent variables, e.g. the i.f. strip of a radio set.

If the same automatic test equipment is used for a range of products it may be that a variety of operators will use it, the operator being chosen according to the degree of knowledge required for performing tests of any given unit-under-test. This means that each individual person has a change during the course of a day's work and is not tied to the machine all the time.

2.3. *Type of Labour Available*

To a certain extent the degree of operator participation is going to depend on the grade of labour available for operating the automatic test equipment. If the availability of skilled labour is at a premium then the design of the automatic test equipment, interfaces and software must be for use by an unskilled operator. This may necessitate the design of more elaborate interfaces, more expensive and detailed displays and consequently a higher initial capital outlay. However, this must be offset against the running expenses by employing a cheaper grade of labour. Inherent in this 'simplification' of the testing procedure, it is more than likely that the number of operator interventions programmed onto the test tape will be increased since whereas the skilled operator could probably go direct to a likely cause of failure by intuition, the unskilled operator must follow pre-programmed procedures.

The ideal is to simplify the actual testing operation to such an extent that unskilled or semi-skilled personnel can be used without increasing the operator participation. This presumes that no decision may be made by the human operator during the testing process and that the unit-under-test is passed, together with a print-out of test results, to another area where more skilled staff may analyse results and carry out a repair if necessary.

2.4. *Design of the Automatic Test Equipment*

Associated with every operator intervention is the necessity for the operator to manipulate the controls of the automatic test equipment and read the indicator display. For unskilled operation few controls are required. Sometimes it may be convenient to duplicate these on a remote control panel placed near to the unit-under-test. On the other hand, for more sophisticated testing or when proving new programs, the whole of the control panel should be within easy reach of the operator. Illuminated push buttons, thumb-wheel switches and neon numerical tube displays are usually preferred for use on control panels.

Another important consideration, to minimize the time spent by the operator when carrying out a manual operation during a test routine, is the method of conveying suitable instructions to the operator. There are various ways of doing this, of which the commonest is probably a list of written instructions itemized by test numbers. This method is relatively simple and cheap, but not the quickest from the point of view of overall operator-participation time.

An alternative to having the instructions written out separately is to program them direct on to the program tape and use a teletypewriter on-line. In this way the automatic test system can be made to select the instruction depending on the test result. For example, a test might be to check the output state of an amplifier. If there is insufficient output, the instructions could be 'Select Continuous Encode and adjust RV3 for maximum reading'. If, however, the test passed in the first place, the test routine will pass on to the next test in sequence skipping over the instruction. The main advantage of this method is that the operator does not have to waste time scanning a list for the appropriate instruction, since it is presented directly to him when it is required.

Visual projection methods may also be used in the interests of greater speed, though these tend to increase cost. Two such methods are:

either an automated method of selecting and projecting normal photographic slides on a frosted glass screen,

or an alpha-numeric display on a cathode-ray tube.

The above considerations help the designer of an automatic test equipment to produce the best ergonomic design, bearing in mind two factors, firstly, the initial capital outlay, and secondly the cost of operation after installation.

2.5. *Design of Test Interfaces*

Having designed the test equipment to minimize operator time, it follows that equal care must be given to the ergonomic design of interfaces for use with the automatic test equipment.

The test interface, often commonly referred to as a jig, is the special-to-type item necessary for connecting the unit-under-test to the automatic test equipment. The design of the interface is probably the most significant area where care can keep operator time to a minimum. Operator time is the most important single factor where bad design can radically increase overall test time. A bad design will mean the operator spends a long time loading the unit-under-test into the interface, a long time removing it after testing is complete, and a long time manipulating any controls during operator interventions. However, in some cases it may be unavoidable that this is a lengthy process by virtue of the design of the unit-under-test. In this instance, care must be taken to write the test routines in a manner to permit maximum testing once the unit-under-test has been connected.

To illustrate the importance of minimizing the operator intervention time, an example is given in Appendix 2 of testing radio transceivers during manufacture. The automatic test equipment was required to test thin-film modules, sub-units, sub-assemblies and the final assembled radio. In the case of the completed radio it was necessary to be able to test in either the cased or uncased condition, and in addition alignment tests, a complete set of production 'A' tests or a quick overall functional test was required. In all there were 36 units to be tested, plus the completed radio assembly. A sample of 9 units were chosen and analysed for test requirements. The end result is to show a £5000 saving per annum for a 50% reduction in mean operator intervention time for just one radio.

Having considered some of the factors which determine the degree of operator participation in automatic testing this paper continues to examine the area of machine interfaces in more detail. In addition to providing interconnection functions, an interface may provide dummy loads, special monitoring facilities, remote switching or multiplex facilities. It is felt that this area represents one of the most significant aspects of modern automatic test equipment design and will dictate to a large extent the degree of automation which is achieved in automatic testing over the coming years. The following section considers the various approaches to the design of interfaces.

3. Design of Interfaces

In general the experience has been that the more complex the unit-under-test becomes (i.e. the more complex the test routine) then the greater is there the need for skilled operators and the more complex interfaces become. Obviously as interfaces become more complex their cost increases, together with the cost of preparing programs for them. Inevitably there is a break-even point, but it would be difficult to define this in general terms. Every unit-under-test has to be treated on its merits, and factors such as quantity throughput, depth of diagnosis required and economics of alternative methods must be considered.

In Section 2 above it has been shown how important it is to minimize operator participation time when testing complex units. The design of the interface to be used between the automatic test equipment and the unit-under-test will influence the connecting-up time, any operator action time required during the test routine and disconnecting time at the end of the test routine. In order to make best use of an automatic test equipment these times need to be kept to a minimum, thereby increasing the proportion of the overall test time when the machine is in use. Some examples of interfaces which take into account the above consideration are given in the following Sections.

3.1. *Multi-head Interfaces*

It may be possible, in order to increase test equipment usage, to have a facility whereby two or more units-under-test can be connected simultaneously, so that whilst the automatic test equipment is testing one unit the operator may be carrying out a manual operation on another. A simple example of a multi-head interface is a double-headed jig for testing thin-film circuits. Within the base of the jig there could be a switching arrangement which will route the test lines to either one of two positions, A or B. A lamp for each position may be used to warn the operator that the unit is in the middle of a test routine. A unit is placed in position A and the test routine commenced. As tests are being carried out in position A the operator can be loading a unit into position B so that as soon as the automatic test equipment has tested the unit in position A it can be immediately switched over to B, thus eliminating loading time.

Multi-head interfaces can take on basically two forms. Either there is only one set of connexions to the automatic test equipment, as described above, and switching is done within the interface or, alternatively, the switching is done within the machine and the interface connecting panel accommodates connexions to a number of similar interfaces. Decisions as to which method to use will depend on a number of factors. Amongst them no doubt, will be:

(i) space required for accommodating the switches and other components,

(ii) whether the switching is to be manually or automatically controlled,

(iii) the suitability of connecting other interfaces should other units have to be tested on the same automatic test equipment.

3.2. *Variable Size Printed Circuit Board Interface*

In the case of printed circuit boards fitted with edge-connectors through which all electrical connexions are

made when in normal operation, it may be possible to functionally-test a board using the edge connector alone. In this case the interface required is simple to make and easy and speedy to use. However, if diagnostic testing is required, particularly if the printed circuit board is known to be faulty, then it is probable that some kind of probing to the printed-circuit track will be necessary. Jigs using spring-loaded gold-plated probes that are capable of doing this are not uncommon. To date they have usually been developed for use on mass-production lines where the throughput of the same printed circuit board has warranted a special interface and they are commonly referred to as a 'bed of nails'. In these interfaces both the probe positions and the size of the printed circuit board are fixed.

Using the 'bed of nails' principle an interface has recently been developed which can accommodate printed circuit boards of different sizes up to a maximum of 25 cm × 30 cm (10 in × 12 in), and provides probes in any required position over the surface area of the board on either side. A photograph and diagram of the interface are given in Fig. 1. The technique is to use a perspex sheet cut to the same size as the board-under-test. The spring-loaded probes are mounted on this perspex sheet. A light illuminates the printed circuit board track so that the operator can line up the probes with the track to ensure that correct contact is made. The clamps holding the boards are mounted on slides to permit testing of printed circuit boards of different sizes. The perspex board can usually be prepared quicker than the time it takes to write and prepare the program for testing it.

3.3. *Interfaces Requiring Probe Control*

The interfaces described in the above Section, using probes to collect test information from the printed circuit track, are usually acceptable when the measurements are of a d.c. or low-frequency nature. Large numbers of probes can be applied to the circuit at the same time without affecting the performance of the circuit. This is not the case when testing has to be carried out at high frequencies, and problems can be caused by many factors. Typical problems are the probes themselves becoming radiators, a probe may upset the impedance balance of the circuit, etc. This is not to say, though, that every time high-frequency testing has to be carried out many probes are required or that, when doing diagnostic testing, the circuit must be operated at the high frequency in order to find the faulty components.

Where high-frequency test points exist a fairly simple and cheap solution may be to use the interface described above but employing two different perspex boards, one with a few probes for a high frequency test routine and another with a larger number of probes if a diagnostic routine is required. Alternatively, an interface may be designed with which, by operating a simple lever system, the operator may apply as few or as many probes as are required. These interfaces may be considered as static interfaces, inasmuch as there is no control of the mechanical operation by pre-programmed command from the automatic test equipment. Dynamic interfaces, on the other hand, are controlled from the program.

3.4. *Dynamic Interfaces*

A certain amount of development has already been undertaken into dynamic testing using probes. The outcome, a 'programmable interface', has been in use for testing television printed circuit boards as part of a production flow process. The interface is designed so that, from an instruction in the program, it is possible to call one or more probes as and when they are required. When a probe is called it physically moves to make contact with the circuit, and when the instruction is cancelled the probe is withdrawn. In this way only the probes required for any particular test need be in contact with the circuit at any one time. Thus problems of high frequency testing can in most cases be overcome.

Another dynamic interface which has recently been developed is the electro-mechanical interface or automatic 'knob-turner'. The particular electro-mechanical interfaces mentioned in later Sections have been developed primarily for use with modern radio transceiver equipment, but their construction is such that they can usually be adapted for the testing of any unit-under-test where it is desirable to automate the pre-setting of the controls on it.

3.4.1. *Electro-mechanical interfaces*

Some units-under-test require a considerable degree of operator intervention during the test routine. Presetting of switches and knobs, adjustment of variable resistors, inductors, capacitors and similar variables are some typical examples. It is sometimes feasible to automate the control of these in order to reduce operator intervention time and thus eliminate causes of potential unit-under-test rejection due to unreliable operator action. In the past however economic justification for the use of electromechanical interfaces was difficult to appreciate due to the lack of technical development in this area and the associated scepticism on the part of the user.

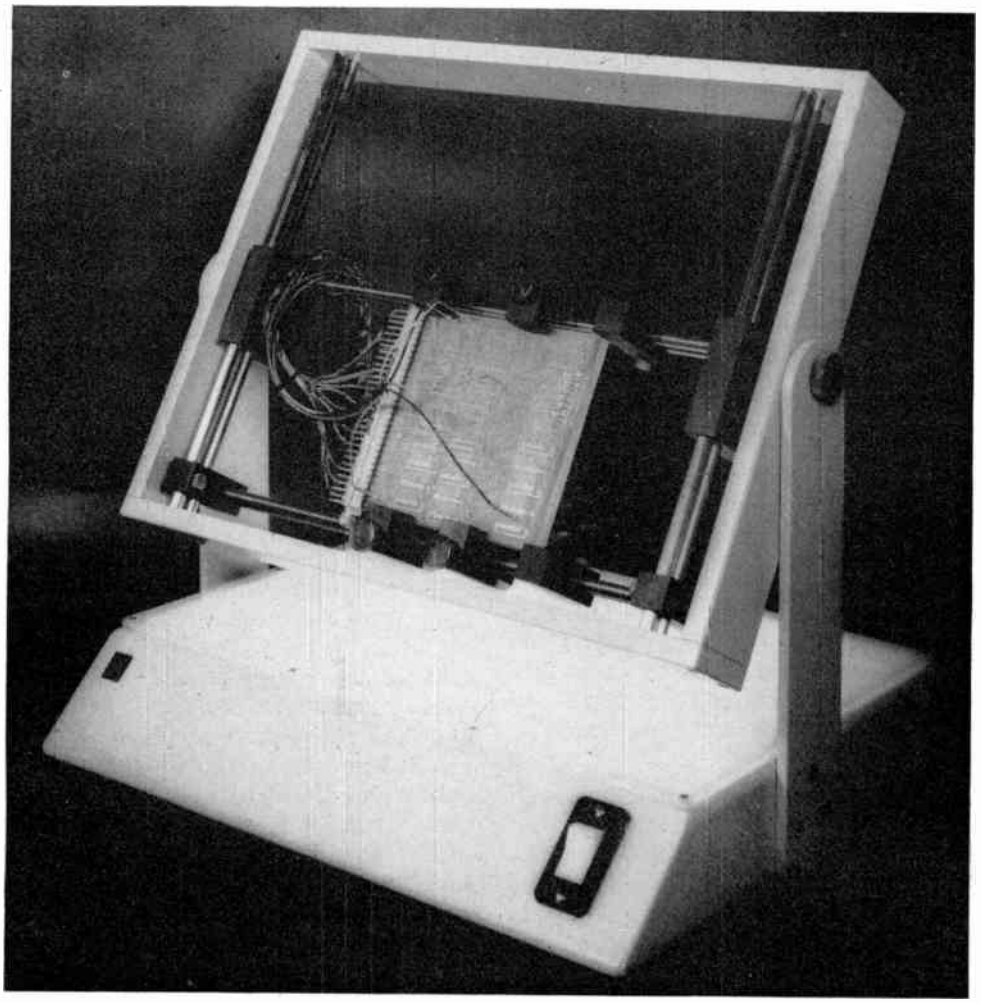
To illustrate the economic advantage of using an electromechanical interface the following example is quoted in Appendix 3. A study was carried out by Honeywell in the U.S.A. on a range of seven radios to ascertain relative test-time to operator action time. It is shown that a saving of 76% over the equivalent manual method can be achieved. This is reflected directly into monetary terms, namely the operator's wages and the overhead cost for operating the automatic test equipment.

An example where theory is being put into practice is the interface used as part of the *Clansman* field test console and this is discussed in some detail in the following paragraphs.

4. The Clansman Field Test Console

The *Clansman* field test console is mentioned briefly in this paper since, not only is it the first automatic test system to be used for testing military radios in the field, but it makes extensive use of electro-mechanical interfaces to minimize operator intervention during testing.

The requirement was for an automatic test system capable of testing radios. The radios range from the small man-packs used by foot soldiers to mobile radios installed in soft-skinned and armoured vehicles.



(Courtesy of British Steel Corporation)

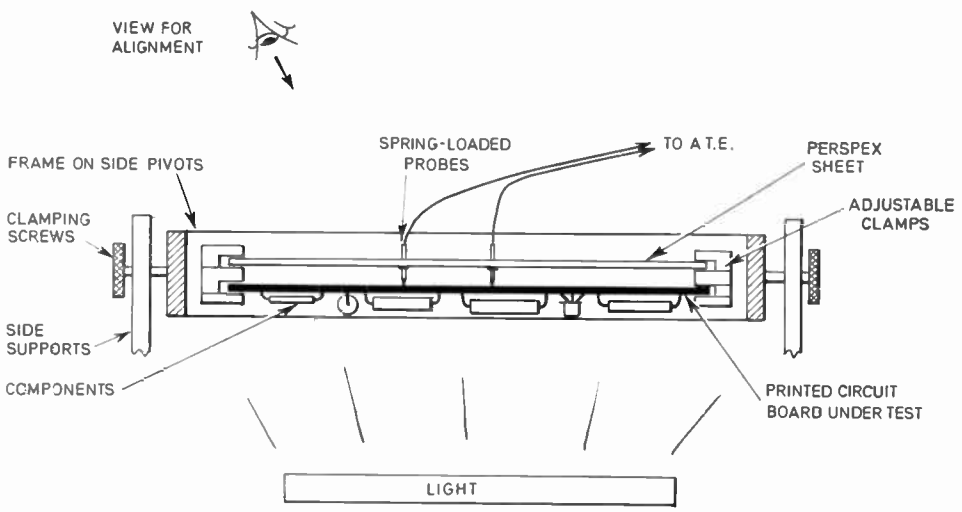


Fig. 1. Universal test jig for picking up test points on printed circuit boards of different sizes. This interface demonstrated the ability to route high-frequency signals through spring-loaded pins to the automatic test equipment and at the same time use connexions available at the edge connector.

In examining the test requirements to support the *Clansman* radios, the British Army carried out an investigation of test times and concluded that the amount of operator intervention that would be required precluded the use of conventional manual test methods, and that some form of automatic test equipment using automatic knob setting would be required. An eighteen-month research and development programme was undertaken to investigate the optimum method of producing an electro-mechanical interface. This programme examined all aspects of the problem including methods of coupling to the knobs, allowance for angular inaccuracies, variations in torque from radio set to radio set, methods of drive, variations in detent pull-in, switch position monitoring, control of drive speed to alleviate wafer damage, spring position holding and safety interlocking, etc. Having regard to the above factors, and the requirements for high reliability, it was found that a number of possible solutions such as rotary solenoids, Ledex switches, stepping motors, printed circuit motors, and torque-sensing motors were unacceptable. The optimum solution required the use of d.c. motors, clutches and digitizer modules. The mechanical approach was to produce a modular device, which would be capable of adaption to other sets by straightforward re-configuration of the basic modules.

The result of the above development was to provide a set of highly precise but reliable, interfaces for the range of *Clansman* radios. The electro-mechanical interface illustrated in Fig. 2 is used for pre-setting the controls on a *Clansman* s.s.b. radio and is typical of all interfaces in the range. The interfaces are controlled by the programmed tape and are capable of rapid attachment and detachment to the radios. They also contain safety interlock circuits and are capable of manual control if necessary.

The programming language used on this tape-controlled test equipment is a mnemonic language, based on the ISO code (or ASCII code). A two-letter address is used to call a particular instrument or interface and this is followed by a four-digit command. The address letters give the language its mnemonic form by being selected as initials of the instrument or interface being addressed; for example, the letters DV are used to call the digital voltmeter and MU, ML are used to address the mechanical interface upper and lower switch decades respectively.

The *Clansman* field test console enables a quick turnround in the repair of an expensive piece of equipment, and at the same time reduces the administration necessary for returning a radio to a base work-shop. In conclusion, the electro-mechanical interface enables a much improved integrity of testing and a more efficient use of existing manpower.

5. Future Trends

The 'de-skilling' operation in automatic test systems will become increasingly more important when considering future designs. Already automatic testers are being used for interpreting physical quantities so that testing can be carried out on mechanical equipment. The ability to do this today primarily depends on the development of suitable transducers for interpreting a physical quantity into an electrical signal. In time this will lead to the automatic testing of complete mechanical systems such as jet engines, cars, lorries, bus fleets, which today is becoming an ever more urgent requirement because of threats of stricter legislation and the consequent need for more accurate methods of testing.

In this case the operator will be a normal service mechanic and not likely to be familiar with the detailed operation of the automatic test equipment designed to help him with his diagnosis. Therefore, the necessity is to look now for ways in which automatic test equipment of the future will be easy to operate.

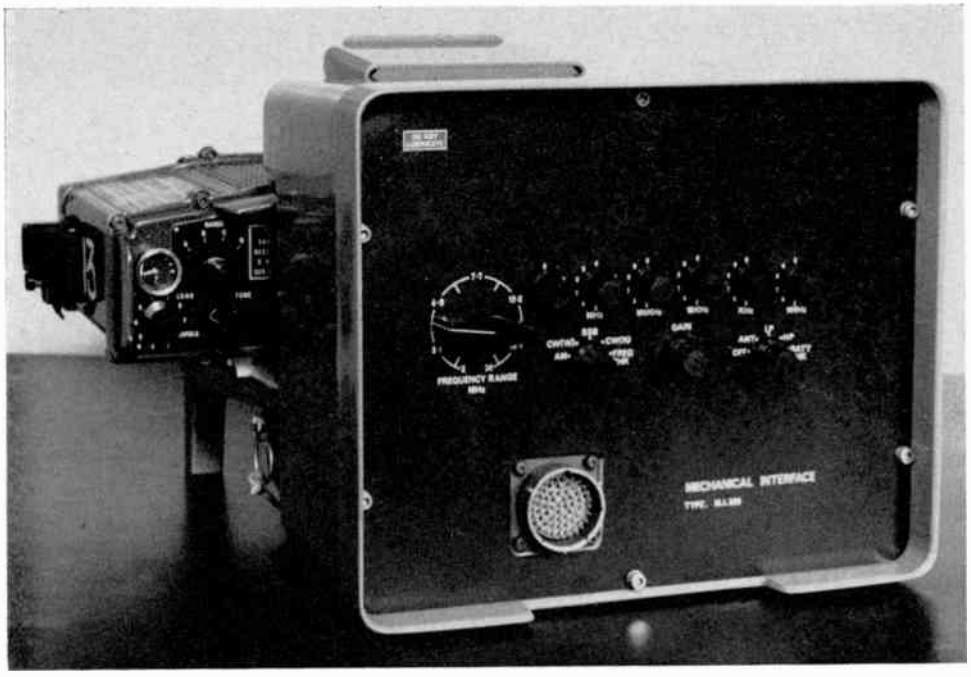


Fig. 2.
One of the electro-mechanical interfaces to be used as part of the *Clansman* field test console.

Operator intervention in automatic testing will decrease as techniques of interfacing are developed and connecting methods standardized. Accompanied by an appreciation on the part of the test subject designer to eliminate the need for manual intervention during testing, automatic test equipment may one day become truly automatic whilst remaining capable of testing a variety of products.

6. Conclusions

The cost and availability of labour suitably qualified to carry out routine manual testing on modern electronic equipment is today forcing management to consider automatic methods as an alternative means of testing. Using today's instrumentation this is feasible from an engineering point of view, but has to be proved as economically viable before capital sanction is granted. As part of the exercise for justifying the purchase and installation of an automatic test system care must be taken to engineer the system so that its operation is 'de-skilled' as much as possible. At the same time care must be taken to prevent design engineers who, in their efforts to 'de-skill' testing cause the operator action time to be increased so much that the system becomes uneconomic to operate. A compromise has to be achieved between a short repayment period of invested capital on the one hand and a system that is economic to run on the other.

Associated with the 'de-skilling' of the test operation is the correct method of interfacing to the unit-under-test and the choice of the test routine when preparing the programs. Careful examination of the cost of achieving a reduction in operator participation against eventual savings is an essential part of establishing a cost-effective automatic test system.

7. Appendix 1: Relation between test time and operator participation

Let y be the overall test time for any given unit-under-test.

Let m be the average time per operator intervention. In practice this is not usually too difficult to work out but may differ considerably from one unit-under-test to another.

Let c be the total number of tests to be performed.

Let n be the average time per test.

Let x be the number of operator interventions.

Then,

$$y = mx + nc$$

nc is the time that the automatic test equipment is actually performing a test operation whilst mx is the time it is waiting for the operator to perform some manual intervention.

For any given automatic test equipment n (the average time per test) is known and so nc may be taken as constant. Therefore there is a direct linear relationship between the test time for a unit-under-test and the number of operator interventions. The slope of the graph y against x is determined by the average time for each operator intervention: hence it is important to keep this to a minimum.

8. Appendix 2: An analysis of a radio prior to production testing

The automatic test equipment was required to test thin-film modules, sub-units, sub-assemblies and the final assembled radio. In the case of the completed radio it was necessary to be able to do alignment tests, a complete set of production 'A' tests or a quick overall functional test. Where practicable tests had to be carried out on the completed radio in either a cased or an uncased condition. In all there were 36 units to be tested as well as the completed radio assembly. A sample of 9 units were chosen and analysed for test requirements, and the results obtained are given in Table 1.

Table 1

Results of an analysis carried out on a modern radio set

	Tests	Operator interventions
9 sub-units:		
Alignment testing	99	29
Functional testing	118	14
Complete assembly:		
Alignment testing	125	18
Production 'A' testing	99	6
Functional testing	8	0
TOTAL	449	67

Experience to date has shown that when carrying out automatic testing on radios an average time of one second per test can be achieved ($n = 1$ in equation, Appendix 1). However, the time per operator intervention is not so definable, but if an average time of 15 seconds was achieved, then total operator intervention time for the example quoted would be $(15 \times 67) = 1005$ seconds, i.e.

$$\frac{1005}{1005+449} \times 100\% = 69\% \text{ of total test time.}$$

A reduction of 50% in this average time for operator intervention will give

$$\frac{502{\cdot}5}{1005+449} \times 100\% = 34{\cdot}5\% \text{ reduction in overall test time.}$$

If it is assumed that the cost of running the automatic test equipment is £2·50 per hour including overheads, then 34·5% would give a saving of

$$\frac{34{\cdot}5}{100} \times 50 \text{ (weeks)} \times 40 \text{ (hours)} \times £2{\cdot}50 = £1725$$

spread over a year of operation.

In other words a reduction from 15 seconds to 7½ seconds in mean operator intervention time by careful design of interfaces would lead to an approximate saving of £1725 per annum on 9 sub-units and the completed radio set. If this figure is amortized over the production testing of the whole radio a saving of £5000 is probable. Taking this a logical step further to cover the testing of a number of different radios being manufactured in the same factory and more than one automatic test equipment in use, much higher savings may be achieved.

9. Appendix 3: An analysis of a radio series to justify the use of an electro-mechanical interface during automatic testing

A study was carried out by Honeywell in the U.S.A. on a range of seven radios to ascertain relative test-time to operator-action time. The seven radios are made up of three basic types: three type A, one Type B and three Type C. The results of this investigation are given in Table 2.

In general an operator can only turn one knob at a time which means that it is necessarily a sequential operation. An operator familiar with the test program will take an average of 15 second per operator action, including the time taken to read an instruction chart, if necessary. On the other hand an electro-mechanical interface will turn several knobs at once as a parallel operation. During trials it has been shown that an average time of approximately 3 seconds per operator action can be achieved. Using these figures the overall test time for the radios can be worked out as shown in Table 2. Summarizing these results it can be seen that the test times for the 7 radios are as follows:

time using a manual knob-turning method

$$= (3 \times 5840) + 4506 + (3 \times 5445) \text{ s}$$
$$= 10 \text{ h } 39 \text{ min } 21 \text{ s}$$

time using the electro-mechanical interface knob-turning method

$$= (3 \times 1328) + 1206 + (3 \times 1401) \text{ s}$$
$$= 2 \text{ h } 36 \text{ min } 33 \text{ s}$$

The saving in time is 8 hours approximately, i.e. 76% of the manual method.

Table 2

Analysis of a series of 7 radios prior to automating their test routines

Radio	Test time (seconds)	No. of operator actions	Manual operator time (seconds)	E.M. interface time (seconds)	Overall test time (manual) (seconds)	Overall test time E.M. interface (seconds)
	a	b	$c = (15 \times b)$	$d = (3 \times b)$	$e = (a+c)$	$f = (a+d)$
Type A	200	376	5640	1128	5840	1328
Type B	381	275	4125	825	4506	1206
Type C	390	337	5055	1011	5445	1401

Notes

1. E.M. interface stands for electro-mechanical interface.
2. There may be a number of operator actions (e.g. 4 knobs may require turning) for one test to be carried out.
3. The 7 radios are made up of three Type A, one Type B, three Type C.
4. Assumptions based on field trials
   (i) The mean time for setting knobs manually per operator action was taken to be 15 seconds.
   (ii) The mean time for setting knobs by electro-mechanical means per operator action was taken to be 3 seconds.

Manuscript first received by the Institution on 12th January 1970 and in final form on 4th December 1970. (Paper No. 1362/IC38.)

U.D.C. (621.376.2+621.394.44)(091)

The Early History of Amplitude Modulation, Sidebands and Frequency-Division-Multiplex

By

Professor D. G. TUCKER,
D.Sc., C.Eng., F.I.E.E., F.I.E.R.E.†

It is shown that the ideas of f.d.m. originated with Alexander Graham Bell around 1870 and were formulated as an f.d.m. telephone system by Leblanc in 1886. Amplitude modulation of a carrier by speech probably originated also with Leblanc in 1886. The existence of sidebands (or sidetones) was demonstrated experimentally by Mayer in 1875 and theoretically by Rayleigh in 1894, but was not known to the early radio and telephone engineers, being apparently rediscovered in 1915. The main developments up to about 1920 are briefly discussed.

1. Introduction

The early history of carrier telephony as such is, on the whole, reasonably clearly presented in the literature; and in particular the long paper by Colpitts and Blackwell[1] of 1921, although it is misleading in one or two respects concerning the earlier history, includes an excellent historical survey from about 1890 onwards. There is little need, therefore, to repeat this part of our history in anything but outline form. When, however, we look at the way the basic ideas developed—that is, the ideas of frequency-division-multiplex, of amplitude modulation, and of sidebands—we find no very clear picture. Indeed, the story of sidebands is most remarkable. Writers on the history of modulation, such as Heising,[2] state that the ideas of sidebands were developed around 1915; certainly radio and electrical workers before that time appeared completely ignorant of sidebands. Yet sidebands were experimentally demonstrated by Mayer[3] in 1875 and theoretically and experimentally demonstrated by Rayleigh[4] in 1894, in both cases in the context of acoustics.

In this paper we set out to describe the early history of the subject in reasonable perspective, and the description will be aided by the diagrammatic summary of Fig. 1.

2. Early Ideas of Multiplexing in Telegraphy and their Influence on Telephony

Before starting on our main subject, it is interesting to note that several methods of multiplexing telegraph channels were proposed and/or developed, and three of these proved to be the forerunners of important telephone multiplexing methods. 'Duplex' telegraphy was the name given to a system permitting messages to be sent simultaneously both ways over a single line; it was invented in 1853 by Gintl, but there were many other similar inventions around the same time, including one by Lord Kelvin using capacitors to enable duplex working to be achieved on cables. A good account of this piece of history is given by Bright.[5] The essential principle was a differential coil, or later a balanced bridge, to prevent a transmitted signal from disturbing the receiver at the transmitting station. The same principle was later used in amplified telephone lines, using what became known as the 'hybrid' coil to separate 'go' and 'return' channels at the terminals and at intermediate repeater stations.[6]

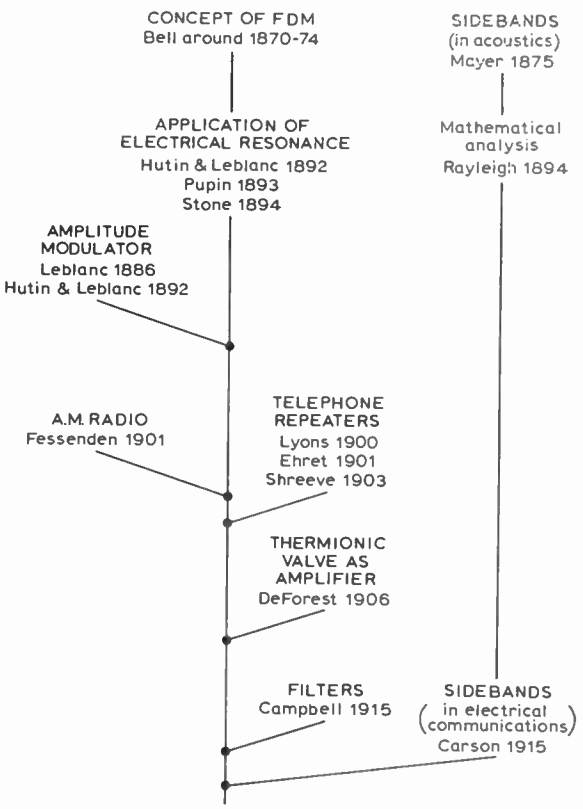


Fig. 1. Diagrammatic summary.

'Quadruplex' telegraphy,[7] developed by Edison in 1874, although first proposed in 1855, was a method of transmitting two messages each way over a single line, and was achieved by adding to the duplex system a method of distinguishing two simultaneous messages by what amounted to putting a d.c. bias on one of them. There was no telephone application of this.

'Multiplex' telegraphy,[7] successfully developed by Baudot in 1875, used a mechanically-driven rotary switch to sample a number of messages successively at an appropriate speed, and a synchronous switch at the receiving end distributed the samples to the correct receiving channels. This was what is now called 'time-division multiplex' (t.d.m.). After some misunderstanding of its possible application to telephony by people such as Leblanc[8] who failed to appreciate the need for a very high speed of sampling, it was Miner[9, 10] who in 1903 showed that it could work on speech signals if the

† Department of Electronic and Electrical Engineering, University of Birmingham.

sampling frequency was made higher than the frequencies contained in the speech. In his own experiments he used a sampling frequency of 4320 Hz. He did not fully understand the sampling requirement, for his patent claim No. 1 was:

> 'The herein-described improved art of multiplex telephony consisting in synchronously closing the connection between the line and corresponding branches or subcircuits with a frequency corresponding to the frequency of the tones and overtones characterizing speech.'

According to Black,[11] it was left to Carson in 1920 to develop the true sampling theorem on which modern t.d.m. systems are based.

Finally, a system of multiplex telegraphy was proposed by Bell[12] sometime during the period 1870–1874, using the principle we now recognize as 'frequency-division multiplex' (f.d.m.). As this is the real starting-point of our main study, it is dealt with fully in the next section.

3. The Development of Frequency-division Multiplex

3.1 *Bell's Concept of Multiplex Telegraphy using F.D.M.*

Alexander Graham Bell (born in 1847) had been interested in acoustics and resonance and had studied Helmholtz's work[13] at quite an early age, and by 1870 was experimenting with electrically-maintained tuning forks. He studied electricity and telegraphy, and was struck with the fact that the Morse code could be read by sound.[12]

> 'Instead of having the dots and dashes recorded upon paper, the operators were in the habit of observing the duration of the click of the instruments, and in this way were enabled to distinguish by ear the various signals.
>
> 'It struck me that in a similar manner the duration of a musical note might be made to represent the dot or dash of the telegraph code. . . . It seemed to me that in this way a number of distinct telegraph messages might be sent simultaneously from the tuning fork piano to the other end of the circuit, by operators each manipulating a different key of the instrument. These messages would be read by operators stationed at the distant piano, each receiving operator listening for signals of a definite pitch, and ignoring all others. In this way could be accomplished the simultaneous transmission of a number of telegraphic messages along a single wire, the number being limited only by the delicacy of the listener's ear. . . .'

This is clearly a proposal for an f.d.m. system. Of course, Bell had no idea of the factors limiting the number of channels he could obtain. He refers to the 'delicacy of the listener's ear' as the limitation. Later on in the paper cited[12] (p. 400) he says, referring to the use of undulatory currents (i.e. sine waves) as what we would now call carriers:

> 'Hence it should be possible to transmit as many musical tones simultaneously through a telegraph wire as through the air.'

He goes on to describe the 'electric harp' as an application of the principle and describes it as 'my first form of articulating telephone'; so we might say that his first telephone proposal was derived from f.d.m. telegraphy. He also proposed the application of f.d.m. to provide one axis of scanning for picture telegraphy (see p. 395 of paper cited[12]).

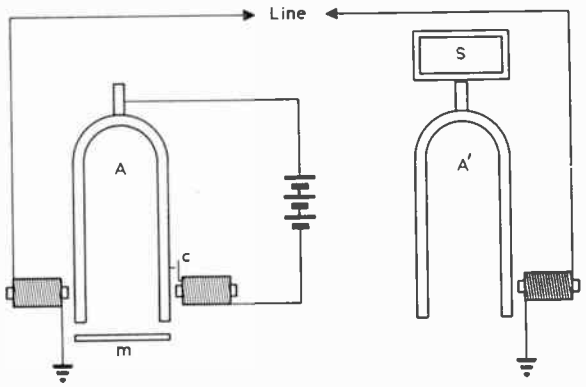


Fig. 2. Leblanc's amplitude-modulation system of 1886.
A transmitting tuning fork
A′ receiving tuning fork
S sounding box
c make-and-break contact in maintaining circuit
m modulating plate or diaphragm

The application of these ideas of f.d.m. to telephony appears to have originated with Leblanc,[8] who in his paper of 1886 describes very clearly a system of amplitude-modulation and its application to f.d.m. telephony. Figure 2 is redrawn from his paper. An electrically-maintained tuning fork (A) generates the supersonic carrier-current wave, and the amplitude of this in the line is modulated by varying the mean (or biasing) magnetic field at the pick-up coil by vibrating an iron plate (m) in the vicinity of the tines. This iron plate can be made the diaphragm of a microphone. Leblanc actually draws (in another figure) the waveform of the amplitude-modulated signal. Reception is effected by a tuning fork (A′), not self-maintained but driven by a coil carrying the line current, and connected to a sounding box (S). By using several such arrangements, each with a different frequency, multiplex telephony can be obtained.

Leblanc was evidently aware of the difficulty of getting a highly-resonant system such as a tuning fork to respond fast enough to reproduce the envelope of a speech-modulated wave, for he says the fork A′ must be very light:

> 'Le diapason A doit être très lourd, et le diapason A′ au contraire très léger.
>
> 'En effet, il faut que l'inertie du diapason A soit assez grande pour qu'une fois ébranlé, il continue à vibrer pendant la durée d'une conversation, d'un autre côté le diapason A′ doit avoir aussi peu d'inertie que possible pour que l'amplitude de ses vibrations soit à chaque instant proportionelle à l'intensité des courants phoniques de pèriode convenable qui parcourent la ligne.'†

† 'The tuning-fork A should be very heavy, and A′ on the contrary very light.

'Actually the inertia of fork A must be great enough to keep it vibrating, once started, for the duration of a conversation; on the other hand fork A′ must have as little inertia as possible so that the amplitude of its vibrations may be proportional at each instant to the intensity of the phonic currents of suitable period which traverse the line.'

The frequency of the fork vibration was to be chosen to be above the audible range, which Leblanc understood to extend up to 8 kHz. There is no sign, however, that he was aware of the idea of sidebands.

The f.d.m. systems of Bell and Leblanc clearly depended on mechanical resonance for separating the channels. The use of electrical resonance for this purpose, discovered around 1892–4, was a big step forward. There seem to have been several almost simultaneous proposals, by Hutin and Leblanc[14] in France, and by Pupin[15, 16] and by Stone[17] in the U.S.A. The idea of the complete f.d.m. telephone system with electrical tuning appeared in Hutin and Leblanc's proposals. However, the fact that they thought a 2 kHz spacing of channel centre frequencies to be suitable suggests that they made no practical trials of their system.

3.2 *The Beginning of Amplitude Modulation in Telephony*

Clearly the application of f.d.m. to telephony depended on the development of ideas of amplitude modulation of a carrier wave and the corresponding process of detection. We have mentioned Leblanc's tuning-fork method above. The first such development using electrical methods appears to be by Hutin and Leblanc[14] in 1892. Their system is shown in Fig. 3. Here G is the h.f. generator (a commutator device), T is the microphone used as a modulating element, modulating the h.f. current in the same way as it would normally modulate a direct current in an ordinary telephone, and DR is a dynamometer receiver which by virtue of its square-law action virtually rectifies the a.m. wave and produces an acoustic output derived from the envelope. This appears to be in all essentials a workable a.m. system.

The application of amplitude modulation to radio seems to have been an independent but later development and here Fessenden[18] seems to have been the pioneer. He proposed (in 1901) basically two ways of obtaining a.m. One was the use of a magnetic or dielectric modulator to detune the aerial by an amount dependent on the speech current and thus reduce (i.e. modulate) the proportion of the generator current which flowed in the aerial. The other, which seems to have been used in all his effective radio-telephone experiments, was to use the resistance of a carbon microphone to modulate the aerial current directly in more-or-less the same way as used by Hutin and Leblanc for line telephony.

The idea of using a rectifier of some sort as a detector for a.m. signals appears to have been the invention of

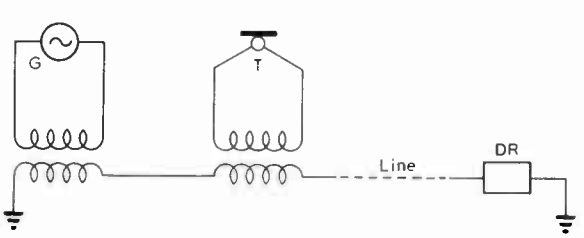


Fig. 3. Hutin and Leblanc's system of amplitude-modulation and detection

G h.f. generator T microphone DR dynamometer receiver

Pupin[19] in 1898. The rectifier was an electrolytic cell, and it was telegraph rather than telephone signals that Pupin was concerned with.

There seems to have been some interaction of radio and line telephony in the period following Fessenden's early work, although it is clear that the correct attribution of ideas was not always made. It is interesting, for example, that Ruhmer[20] said in 1907:

> 'At the same time it is possible that wireless telephony will have a considerable influence on the development of wire-telephony. We may notice, for instance, the problem of multiplex telephony, the solution of which may lie in the adoption of wireless methods.'

Ruhmer himself took out a Belgian patent[21] in 1910 of which the official abridgement is:

> 'On emploie des courants alternatifs à haute frèquence et de frèquence différente, qui se superposent sur la ligne et ne sont séparés qu' á la station réceptrice en agissant là sur différents circuits oscillatoires accordés avec les circuits oscillatoires de transmission correspondants.'†

This seems hardly an advance over the concepts of 1892. But his experiments show a big advance in detailed technique through the adoption of radio methods.[22]

Ruhmer also discusses in his book[23] the possibility of using f.d.m. telephony over a light-beam link, using a modulated arc-light for transmitting and a selenium cell as detector, with light filters for the various channels.

An important discussion of f.d.m. on lines appeared in 1911 in the paper by Squier,[24] who also held some patents on f.d.m.[25] He was concerned with working on cable, superposing carrier channels on the existing 'low frequency battery system', i.e. on the existing audio channel. He used tuned circuits for filtering, and gave many experimental measurements of the response of tuned lines.

The techniques of modulation changed considerably when the thermionic valve became available. The early triode valve, or 'audion', could be used for modulating a carrier wave, and a good account of the development of modulation using valves is given by Heising.[26] Similarly the diode valve could be used as a detector, although for some time after its invention the crystal detector was preferred as being more reliable and much cheaper; it was discovered by Dunwoody[27] in 1906 and then based on carborundum (silicon carbide). Pickard[28] followed this up rapidly with a choice of other materials, and many more were subsequently used as crystal detectors.[29, 30] The modern semiconductor rectifiers and transistors are directly derived from the crystal detector.

3.3 *The Influence of Telephone Repeaters*

The influence of telephone repeaters (or amplifiers) on the development of f.d.m. telephony was for a long time negligible. A good history of the telephone repeater was given by Gherardi and Jewett[6] in their paper of 1919.

† 'High-frequency alternating currents are used, of different frequencies, which are superposed on the line and are separated only at the receiving station. There they act on different oscillatory circuits lined up with the corresponding oscillatory circuits at the transmitting station.'

They suggest that the first practical telephone repeater was a receiver/transmitter type (i.e. an electrodynamic repeater based on telephone instruments) developed by H. E. Shreeve in 1903, and they also mention that in 1900 J. Lyons disclosed to the Bell System a proposal for a repeater based on an asynchronous generator, i.e. an induction motor driven above synchronous speed by a local prime mover. Such an arrangement is, in effect, a negative-resistance amplifier. Probably neither of these repeaters would have been suitable for f.d.m. telephony on account of frequency range. It is interesting though that Ehret,[31] who also invented frequency-modulation,[32] proposed in 1901 an f.d.m. telephone system involving line repeaters of the asynchronous-generator type.

The invention of the triode thermionic valve by De Forest[33] in 1906, following earlier work by others on the diode, opened the door to more effective and useful telephone repeaters. The trouble was that, although the frequency range could now be obtained for f.d.m. working, non-linearity in the response prevented satisfactory working of multiple channels because of the intermodulation-crosstalk produced. Early carrier (or f.d.m.) telephone systems operated on open-wire lines, without intermediate amplification. Some success was obtained with two-channel systems with intermediate repeaters in which the non-linearity was nominally cancelled by the insertion of a complementary non-linearity produced by rectifiers.[34] However, it was not until Black made his ideas on the negative-feedback amplifier[35] generally available in 1934 that multi-channel systems could operate really successfully over amplified lines, and it was from that date that the really important commercial development of f.d.m. telephony began.

4. The Curious History of Sidebands

4.1 *The Radio and Line Telephony Era*

As stated briefly in the Introduction, the story of the discovery of sidebands (or sidetones) is very curious. Radio workers were apparently quite unaware of the concept of sidebands and of the bandwidth requirements of a modulated signal for practically two decades from the first successful radio experiments of Marconi and others. Heising,[2] who was himself a pioneer of radio technology, said:

> 'No equation for a modulated wave is found in the literature prior to World War I. So far as the author has any direct knowledge, the equation was first set up and side frequencies discovered by Carl Englund in 1914. The disclosure to the armed forces during World War I and to the public afterward stemmed from Englund's discovery. It is true that some engineers of the American Telephone and Telegraph Company seem to have had earlier knowledge of the sidebands from their study of carrier-system theory for wire lines, but the perusal of their memoranda and other Company writings has not turned up a prior discoverer. Dr. G. A. Campbell (A. T. & T. Company) believed their knowledge was derived from mathematical expressions in Rayleigh's "Theory of Sound" involving periodically interrupted sound tones.'

If Heising is right in stating that Englund was the first to discover sidebands in the radio context, then Englund was slow to patent a use for his discovery. The first disclosure by Englund that the author can trace is a U.S. Patent[36] filed on 29th March 1916, which mentions that a carrier wave of frequency C modulated by a signal of frequency S gives rise to three frequencies $C-S$, C, and $C+S$, and mentions carrier suppression and re-insertion at the receiver. But in the meantime, Carson had filed a patent[37] on 1st December 1915 which gives a full mathematical theory of the production of sidebands of various types in a non-linear device (i.e. a modulator) and also shows the benefit, not only of carrier-suppression but also of single-sideband operation. It looks very much to the present author as though Carson was the real discoverer (or re-discoverer) of sidebands in radio. Carson also later published a paper[38] on this work. It was Carson who also first demonstrated the sideband structure of frequency-modulation.[39]

An interesting demonstration of how little understanding of sidebands there was in the period prior to 1920 is a mistaken idea of them in the book by the very reputable radio engineer Goldsmith.[40] He says:

> 'If a 100,000 cycle sustained wave be modulated by a 1000-cycle note, both theory and practice agree as to the propriety of regarding the modulated wave as the resultant of *three* separate waves: namely, one corresponding to the frequency of 100,500, one corresponding to the frequency of 99,500, and one corresponding to the frequency of 100,000. All three, being physically present, are detectable with a wave meter, and this has a certain bearing on the selectivity in radio telephony, particularly at very long wave lengths, corresponding to low radio frequencies.'

4.2 *The Real Discovery of Sidebands, 1875 onwards*

It is probable that the first reference to sidebands (although not by that name) is in the paper by Mayer[3] published in 1875. He describes experiments in acoustics in which he modulates or interrupts the sound from a tuning fork by means of a rotating screen with holes in it. As the speed of rotation is increased from zero he notices two additional sounds appear:

> 'On revolving the perforated disk, two additional or secondary sounds appear—one slightly above, the other slightly below the pitch of the fork.'

This is a very clear picture of sidetones.

Lord Rayleigh did some further work on this subject, and set out the theory clearly in the second edition of Vol. 1 of his famous book[4] in 1894. He gives in effect the equation

$$2(1+\cos 2\pi mt)\cos 2\pi nt = 2\cos 2\pi nt + \cos 2\pi(n+m)t + \cos 2\pi(n-m)t$$

and explains that this is only a particular case. As an example of a more complex modulation he expands

$$4\cos^4 \pi mt \cos 2\pi nt$$

showing that this has four secondary sounds.

Rayleigh also gives a very satisfying physical explanation of the secondary sounds. He describes first of all an experiment in which a tuning fork of frequency 128 is driven by a current which is interrupted at frequency 128

by a fork-driven interrupter. This current can also be interrupted by another independent interrupter of frequency 4. When the second interrupter was inoperative, the fork had a strong response in its normal tuning of 128, but scarcely any when tuned to 124 or 132. When the second interrupter was working, however, the fork would respond powerfully when tuned to 124 or 132 as well as when tuned to 128, but not when tuned to intermediate pitches, such as 126 or 130.

The physical explanation which Rayleigh then gives is:

> 'When a fork of frequency 124 starts from rest under the influence of a force of frequency 128, the impulses cooperate at first, but after 1/8 of a second the new impulses begin to oppose the earlier ones. After 1/4 of a second another series of impulses begins whose effect agrees with that of the first, and so on. Thus if all these are allowed to act, the resultant effect is trifling; but if every alternate series is stopped off, a large vibration accumulates.'

This is a very helpful way of looking at sidetones.

It is thus more than likely that Campbell, quoted by Heising, was right in believing that the A. T. & T. Company's knowledge of sidebands was derived from Rayleigh. What is surprising is that so few radio workers were aware of Rayleigh's work.

4.3 *Filters*

The re-discovery of sidebands led to proposals for single-sideband working and to the desire for better methods of selecting frequency bands than the simple tuned-circuit arrangements used up till then. Thus gradually the art and science of filter design developed. Probably the first filter invention was that by Campbell[41] in 1915. The invention of *m*-derived filters by Zobel[42] in 1923 was a most important step, and led to very refined filter design.

5. References

1. Colpitts, E. H. and Blackwell, O. B., 'Carrier current telephony and telegraphy', *Trans. Amer. Inst. Elect. Engrs*, **40**, pp. 205–300, 1921.
2. Heising, R. A., 'Modulation methods', *Proc. Inst. Radio Engrs*, **50**, pp. 896–901, May 1962.
3. Mayer, A. M., 'Researches in acoustics, Part 6', *Phil. Mag.*, **49**, pp. 352–365, May 1875.
4. Rayleigh, Lord, 'The Theory of Sound', Vol. 1, Second edition (Macmillan, London, 1894).
5. Bright, C., 'Submarine Telegraphs', see p. 122 and pp. 635–43 (Crosby, Lockwood & Son, London, 1898).
6. Gherardi, B. and Jewett, F. B., 'Telephone repeaters', *Trans. Amer. Inst. Elect. Engrs*, **38**, pp. 1287–1345, 1919.
7. Denman, R. P. G., 'Line Telegraphy & Telephony', Catalogue of the Collections in the Science Museum, South Kensington (H.M.S.O., 1926).
8. Leblanc, M., 'Etude sur le téléphone multiplex', *La Lumière Électrique (Paris)*, **20**, pp. 97–102, 17th April 1886.
9. Miner, W. M., U.S. Patent No. 745,734, filed 26th February 1903, issued 1st December 1903.
10. 'Recent developments in multiplex telephony,' *Electrical World & Engineer (U.S.A.)*, **42**, p. 920, 5th December 1903.
11. Black, H. S., 'Modulation Theory', see p. 5 (Van Nostrand, New York, 1953).
12. Bell, A. G., 'Researches in electric telephony', *J. Soc. Telegraph Engrs*, **6**, pp. 385–421, 1877.
13. Helmholtz, H. L. F., 'Die Lehre von dem Tonempfindungen als physiologische Grundlage für die Theorie der Musik', 1863. English translation by A. J. Ellis, 'Theory of Tone'.
14. Hutin, M. and Leblanc, M., British Patent No. 23,892; Application date 27th December 1892.
15. Pupin, M. I., 'Practical aspects of low frequency electrical resonance', *Trans. Amer. Inst. Elect. Engrs*, **10**, pp. 370–94, 1893.
16. Pupin, M. I., U.S. Patent No. 707,007, filed 23rd February 1894, issued 12th August 1902.
17. Stone, J. S., U.S. Patents 726,368; 726,476; 729,103; 729,104; all filed 4th April 1894, issued 1903.
18. Fessenden, R. A., U.S. Patent No. 753,863, filed 28th September 1901, issued 8th March 1904.
19. Pupin, M. I., U.S. Patents Nos. 713,044 and 713,045, filed 4th January 1898, issued 4th November 1902.
20. Ruhmer, E., 'Wireless Telephony, in Theory and Practice', published in Germany 1907; English translation by J. Erskine-Murray, published by Crosby, Lockwood & Son, London, 1908. See p. 194.
21. Ruhmer, E., Belgian Patent No. 224,008, dated 16th March 1910.
22. Ruhmer, E., 'Multiplex telephony', *Electrical Review and Western Electrician*, **59**, pp. 28–29, 1st July 1911.
23. Ruhmer, E., as ref. 20, p. 73.
24. Squier, G. O., 'Multiplex telephony and telegraphy by means of electric waves guided by wires', *Trans. Amer. Inst. Elect. Engrs*, **30**, pp. 1617–65, 1911.
25. Squier, G. O., U.S. Patents Nos. 980,356; 980,357; 980,358; 980,359, all issued in 1911.
26. Heising, R. A., 'Modulation in radio telephony', *Proc. Inst. Radio Engrs*, **9**, pp. 305–352, 1921.
27. Dunwoody, H. H. C., British Patent No. 5332, 23rd March 1906.
28. Pickard, G. W., U.S. Patent No. 836,531, filed 30th August 1906, issued 20th November 1906 (silicon crystal); U.S. Patent No. 886,154, filed 30th September 1907, issued 28th April 1908 (zincite).
29. Pierce, G. W., 'Principles of Wireless Telegraphy', see pp. 160–1 (McGraw-Hill, New York, 1910).
30. Palmer, L. S., 'Wireless Principles and Practice', see pp. 302–308 and many references quoted on pp. 332–3 (Longmans, Green & Co., London, 1928).
31. Ehret, C. D., U.S. Patent No. 789,087, filed 3rd December 1901, issued 2nd May 1905.
32. Tucker, D. G., 'The invention of frequency-modulation in 1902', *The Radio and Electronic Engineer*, **40**, pp. 33–37, July 1970.
33. DeForest, L., 'The audion', *Trans. Amer. Inst. Elect. Engrs*, **25**, pp. 735–63, 1906.
34. Ryall, L. E., 'A Few Recent Developments in Telephone Transmission Apparatus', Inst. P.O. Elect. Engrs Printed Paper No. 155, 1934, see pp. 18–20.
35. Black, H. S., 'Stabilized feedback amplifiers', *Bell Syst. Tech. J.*, **13**, pp. 1–18, 1934, or *Electrical Engr*, **53**, pp. 114–20, 1934.
36. Englund, C. R., U.S. Patent No. 1,245,446 filed 29th March 1916, issued 6th November 1917.
37. Carson, J. R., British Patent No. 102,503, Convention date (U.S.A.) 1st December 1915, issued 30th November 1917.
38. Carson, J. R., 'A theoretical study of the three-element vacuum tube', *Proc. Inst. Radio Engrs*, **7**, pp. 187–200, 1919.
39. Carson, J. R., 'Notes on the theory of modulation', *Proc. Inst. Radio Engrs*, **10**, pp. 57–64, 1922.
40. Goldsmith, A. N., 'Radio Telephony', p. 181 (The Wireless Press, New York, 1918).
41. Campbell, G. A., U.S. Patents Nos. 1,227,113 and 1,227,114, filed 15th July 1915, issued 1917; or British Patent No. 142,115, issued 1921.
42. Zobel, O. J., 'Theory and design of uniform and composite electric wave-filters', *Bell Syst. Tech. J.*, **2**, pp. 1–46, 1923.

Manuscript first received by the Institution on 26th March 1970 and in final form on 12th June 1970. (Paper No. 1363/Com.35.)

Contributors to this issue

Mr. A. M. Yousif obtained a B.Sc. degree in electrical engineering from the University of Khartoum in 1964. Subsequently he held an appointment with the Department of Posts and Telecommunications in the Sudan Ministry of Communications. After a short period at Philips International Training Centre at Hilversum, he undertook the information systems engineering course at Birmingham University, obtaining his M.Sc. in 1968. He then joined the University of Bradford where his research work for Ph.D. is nearing completion; he has been working on non-linear distortion phenomena in diode mixers and modulators.

Dr. J. G. Gardiner (G. 1963) is a lecturer on the staff of the Post Graduate School of Electrical and Electronic Engineering at the University of Bradford. A fuller note on his career was published in the May 1969 issue of the *Journal.*

Professor D. G. Tucker (F. 1953) has been Head of the Department of Electronic and Electrical Engineering at the University of Birmingham since 1955. He is particularly interested in the history of science and industrial archeology. A fuller note of Professor Tucker's career was printed in the July 1970 issue of *The Radio and Electronic Engineer.*

Mr. R. H. Crosher obtained his technical training as a student apprentice with British Thomson-Houston in Rugby (later A.E.I. (Rugby) Ltd). He subsequently studied at the Polytechnic, Regent Street, and Enfield College of Technology for the B.Sc. (Eng.) degree of London University, which he obtained in 1967. He has held appointments as technical assistant with Elliott-Automation, as an applications engineer with Ether Engineering Ltd and since September 1968 he has been sales engineer with Honeywell Limited, being concerned with automatic test equipment.

Dr. D. S. Campbell is Technical Manager of the Plessey Company's capacitor plant at Bathgate, West Lothian, Scotland, where he has been since 1968. Previous to this he was initially with the then Ministry of Supply at S.R.D.E,. Christchurch, Hants, and then for 15 years at the Allen Clark Research Centre of the Plessey Company. During the latter part of his time at Caswell he was group leader in charge of work on dielectric, metallic and magnetic properties of materials with particular emphasis on the properties of thin films. He was also a Senior Visiting Lecturer in Materials Science in the Electrical Engineering Department of Imperial College, London. Dr. Campbell studied at Woolwich Polytechnic for his honours degree in physics of London University. Subsequently he was granted a D.I.C. by Imperial College for work on electrical conduction processes in thin films and a D.Sc. by London University for work in materials science with particular reference to the properties of thin films. He has over 60 publications to his credit, including contributions to several books, and he is at present on the Editorial Boards of both *Thin Solid Films* and the *Journal of Material Science.* He holds an Honorary Fellowship at Edinburgh University.

Mr. T. W. Welch (Fellow 1966, Member 1952) spent ten years in the Electrical Branch of the Royal Navy, specializing in radio communications, radar and navigational electronics. From 1950 to 1954 he was with the Radio Advisory Service of the Chamber of Shipping and he then joined Decca Radar Ltd to conduct field studies in new applications of radar, subsequently becoming manager of the radar applications group. In 1965 Mr Welch formed a specialist company, T. W. Welch & Partners Ltd, to provide consultant services in the formulation of operational requirements, feasibility and cost studies and system design for navigational systems, including those employing radio and radar aids in air traffic control, port information and hydrological and flood control systems. This work has extended from the U.K. and Europe to the Far East. Mr Welch has been a member of the Institution's Aerospace, Maritime and Military Systems Group Committee since 1964; he is a Governor of both Guildford County and Farnborough Technical Colleges and chairman of the Engineering Advisory Committee at the former.