THE RADIO AND ELECTRONIC ENGINEER

The Journal of the Institution of Electronic and Radio Engineers

FOUNDED 1925 INCORPORATED BY ROYAL CHARTER 1961

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

VOLUME 30 SEPTEMBER 1965 NUMBER 3

COLLABORATION WITH GOVERNMENT

THE Institution's membership is drawn from every organization employing radio and electronic engineers. Membership is, however, an individual matter, not only of qualification, but personal choice, since there is not in Great Britain compulsory registration of professional engineers. Without statutory obligation, however, the Institution does in fact collaborate as much as possible with educational authorities, industry and government, in securing the development of radio and electronic engineering.

It is therefore of historical importance to record that the Symposium just completed in Durham on " Electronic Control Systems for Industry " was the very first ever to be arranged jointly between the Institution and a government Ministry. The Symposium was officially opened by the President of the Institution, Colonel G. W. Raby who, in his Address, stated:

" Not only is the subject matter of national importance, but so is the fact that this is the first joint effort between the Ministry of Technology (or any other Ministry for that matter) and a learned Institution. We are, therefore, taking part in something of an historic occasion and for my part I hope it is a forerunner of many combined efforts in the future.

"The Government's initiative in founding a Ministry of Technology was widely welcomed as an excellent first step in securing better co-ordination of the country's scientific and technical effort which hitherto has been dreadfully patchy and isolated. The new Ministry has already undertaken parliamentary responsibility for the British electronics industry. Moreover, it is evident that the new Ministry will not be content with a superficial administrative responsibility, but will, in fact, afford active Government help in encouraging research and development in those techniques which will materially help the country to achieve efficiency in manufacture, and reliability of product, both of which are so essential for our export trade."

The President thanked those who had contributed to the planning and organization of the Symposium—officials of the Ministry of Technology's Northern Region, and the members of the Institution's North Eastern Section Committee. He also paid tribute to the hospitality and co-operation of "this ancient and honoured University of Durham." The President continued:

"Electronics touches almost every phase and activity of our life today. Here in the North Eastern Section there is an admixture of industries some old, some very old, but even the oldest of them are being modernized with the aid of electronics. Within the limit of the three days which we are giving up to this Symposium, the choice of subjects covers a very wide range of industrial application. In this way, the Committee responsible for the programme has tried to cover the variety of situations in which the electronic methods of control could be applicable.

"The attractiveness of increasing productivity and reducing costs needs no stressing from me; and it is vital that British industrialists if they are to retain and increase their competitive power, must show at least an equal concern for the maintenance and the improvement of the quality and reliability of their products. This is a gospel which the Ministry of Technology is very rightly preaching in season and out of season, and the tremendous importance of this policy has been kept well in mind in formulating the programme of this Symposium. This area of quality and reliability is eminently one in which electronic control and measurement systems can play a big part and with far greater effectiveness and efficiency than by any other means."

This practical collaboration with government in disseminating new ideas and the discussion of new applications is a venture widely welcomed by British radio and electronic engineers.

G. D. C.

INSTITUTION NOTICES

The Secretary to Visit North America

The Secretary of the Institution, Mr. Graham D. Clifford, will be in North America for about four weeks during October and November.

He expects to arrive in Toronto on 26th October and will meet members in the area. He will also inaugurate the Institution's permanent office in Canada—in the Royal Trust Building, Ottawa.

Subsequently Mr. Clifford will visit Montreal, and he will then go to the United States, visiting Los Angeles, San Francisco and New York.

Details of Mr. Clifford's itinerary will be sent to members in Canada and the U.S.A. and he may be contacted through the Secretary to the Canadian Division, Colonel J. A. D. Craig, Room 1304, 330 Bay Street, Toronto 1.

Conference on V.H.F.–U.H.F. Mobile Communication Systems and Equipment

The Institution is joining the Electronics Division of the Institution of Electrical Engineers in holding a conference on the above subject in London, from 12th to 13th January, 1966.

The scope of the Conference, as it concerns systems, will cover both civil and military mobile communication services and will include, in the civil category, area cover for fire and police services, private and public mobile telephone networks (similar to that which started recently in the London area), selective calling and paging systems, and the aeronautical and maritime services. The military aspects will include systems for tactical use in military operations. In that part of the programme dealing with equipment it is hoped to highlight the latest developments in the application of new components and techniques to equipments (including those carried by a man or mounted in a fighting vehicle) and in the applications to aerials and power supplies for fixed and mobile terminals.

Requests for further information and registration forms, which will be ready a few weeks before the Conference, may be sent to the Institution at 9 Bedford Square, London, W.C.1.

Symposium Papers

The volume of digests of the six papers presented at the Joint I.E.E.-I.E.R.E. Symposium on "Electronics in Industry—The Next Five Years", which was held in Birmingham on 6th April last, is still available. Copies may be obtained at a charge of 7s. 6d. each from G. K. Steel, Electrical Engineering Department, College of Advanced Technology, Gosta Green, Birmingham 4, or from the I.E.R.E.

Subscriptions to *The Radio and Electronic Engineer*

With effect from the January 1966 issue the rates for subscriptions by libraries, etc. for *The Radio and Electronic Engineer* will be increased as follows:

1 year	..	£6 10s. 0d.
2 years	..	£12 0s. 0d.
3 years	..	£17 0s. 0d.

The charge for single copies to non-members will be 12s. 6d.

The cost to members of single copies of the *Journal* remains at 5s. 0d.

Change of Address—An Urgent Request

Members are requested not to delay in advising the Institution at 9 Bedford Square, London, W.C.1 immediately they change their address.

It is not necessary to notify local section secretaries in Great Britain of changes of address. Such notifications are automatically sent from Head Office to the local sections. Because of delays in post, however, overseas members should advise their local secretary at the same time as they advise London.

Failure to notify the Secretary at 9 Bedford Square causes considerable delay in members receiving *Journals*, notices of meetings and other separate communications. In addition, extra postal charges fall on the Institution for *Journals* which have to be sent out a second time.

Reprints of Papers

Members are reminded that a reprint of any paper published in *The Radio and Electronic Engineer* is usually available from the Institution within a fortnight of publication, price 3s. 6d. per copy. Orders should be addressed to the Publications Department at 9 Bedford Square, London, W.C.1, and be accompanied by the appropriate remittance.

Stocks of reprints for previous years are fairly extensive and details of these are shown in the publication, "Abstracts of Papers published in the Journal of the British Institution of Radio Engineers: 1952 to 1963 inclusive". This 108-page book, which includes nearly 900 titles classified according to subject, is also obtainable from the Publications Department, price 10s. 6d.

Correction

The following correction should be made to the paper "A High Speed Tunnel Diode Counter", published in *The Radio and Electronic Engineer* for April 1965.

Page 202, equation (10) should read:

$$\text{If} \quad v_3 > v_y$$

The Fortieth Anniversary Celebration

At the Institution's Fortieth Anniversary Dinner held in London in June (and reported in the August issue of *The Radio and Electronic Engineer*) the Toast of the Institution was proposed by Mr. Justice Ungoed-Thomas, one of Her Majesty's High Court Judges, who, before his elevation to the Bench, was well known to members as a leading Queen's Counsel and Member of Parliament. Sir Lynn's speech took a typically dry look at radio and electronic engineering and extracts are included here to complete the record of a notable occasion in the Institution's life.

"I have a very distinguished list of guests to represent. We are all particularly delighted to see here on this occasion the Italian Ambassador. Two heroes of my boyhood were Garibaldi and Marconi. Garibaldi was, of course, an obvious hero, a fighter and a guerrilla fighter at that, a founder of modern democratic Italy. And Marconi? I remember now how dramatically electronics broke through into my small boyhood world. Dr. Crippen was on the run. For a time even Garibaldi yielded pride of place to Marconi. And this, observe, was electronics in the service of the law.

"Invidious as it is to name any other guests you must allow me to mention two old political enemies—Mr. Marples and Sir Ian Orr-Ewing. For years we sat together, cheek by jowl—or, more precisely, two swords' length apart, across the floor of the House of Commons. Sir Ian is here not only in virtue of himself but also of the post to which we are delighted he was recently appointed—President of the Society of Electronic and Radio Technicians.

"When I was in the House electronics was called in to help. Microphones were unobtrusively inserted into the benches. They were inserted to enable members to hear one another speak. The trouble was we could on occasion also hear one another talk—and the talk was not all Parliamentary. That caused trouble and points of order. And when I was in the House my concern was not to hear other members speak. It was to catch the Speaker's eye and hear myself speak.

"Yet I would welcome an extension of electronics in the House of Commons. What I strongly supported was the introduction of television into the House of Commons. It's not that I think it would provide good entertainment—though on occasion it can be extremely funny. It is simply that M.P.s act in the House as representatives of the people who send them and the people are entitled to see what they are up to there.

"You, Mr. President, suggested that I should pass judgment on the Institution. I know nothing at all about your science. It is wizardry and you are all wizards; it is dabbling in the occult; it is sheer mystery to me. Not that that would deter me from passing judgment on it. The less I know about a case the easier I find it to pass judgment on it. Nor would I boggle at its mystery. After all, it is part of my job to try income tax cases.

"Your Institution once went to law—once only—and that was twenty-five years ago. Speaking as a superannuated member of the Trade Union of the Bar, I will certainly pass judgment on that—a deplorable record. And that was an income tax case. So it was not a case in which reason might be expected to prevail. And, of course, your Institution with its command of the occult, succeeded.

"But there is one thing which I and all your guests do understand and indeed share with your members—your admiration for and devotion to your guest of honour, your Charter President. I remember as a small boy before the first World War knowing, what historians have recorded since, the great work done by your Charter President's father in preparing the Navy for war when Haldane was preparing the Army. And then came his son—young, gay, gallant, daring, dashing and naval. He has since become an historic figure not only in this country but in the Commonwealth and throughout the world—and in more than one field. But I feel that his supreme place in our romantic affection is with the Navy. He has wedded his work with the Navy with his work with this Institution. He has explored into the future and, by his leadership and inspiration, he has rallied and united, in this country and throughout the Commonwealth, services and peoples to his venture. His name will always be associated with this Institution. And in this Institution you, Mr. President, are in the forefront. Your Charter President has given high praise to your pioneering and to your magnificent achievements. All of us, members and guests, would, I know, wish me to subscribe our tribute too."

A photograph of Sir Lynn was published in the August *Journal*, p. 77.

Symposium on Microwave Applications of Semiconductors

Over 450 engineers and physicists took part in the joint I.E.R.E.—I.E.E. Symposium 'Microwave Applications of Semiconductors' held at University College London from 30th June to 2nd of July. The meeting must surely rank as one of the largest ever held in this country on a subject of such a relatively specialized nature. Nearly a quarter of those attending came from overseas and 10 of the 39 papers presented were by authors from the United States or Europe.

The theme of the Symposium was introduced by Professor G. D. Sims (University of Southampton) in a lecture surveying the different applications, and the first day's session then dealt with some of the basic microwave properties. Particular interest was aroused by a group of papers discussing bulk effect microwave generation phenomena, notably the Gunn effect. Subsequent sessions covered various classes of circuit configurations—varactor multipliers, tunnel diode circuits, p-i-n diode modulators, etc.—and concluded with discussing low noise applications and systems such as installations used in radio astronomy and satellite communications.

One of the most valuable contributions which the organizers of a symposium on a highly specialized subject can make for the benefit of those taking part is to provide ample time for informal discussion outside the sessions. Many such meetings held by the I.E.R.E. have been at Universities where college rooms or halls of residence have provided these opportunities, far into the night if need be. The advantages of a London meeting, attractive as they are both to those from overseas and to the organizers, do not readily include such congenial opportunities. However, by arranging for lunch to be taken at the College and for an informal reception on the first evening and a dinner on the second evening, the discussions of the sessions could be continued.

Further occasions for informal exchanges of opinions and experience in this rapidly expanding field arose through the facilities kindly provided by Professor H. M. Barlow, F.R.S., for seeing research work in progress in the Department of Electrical Engineering. A number of exhibits set up by authors of papers were also accommodated in his Department.

The Symposium Dinner was held in the Old Refectory of the College and was under the Chairmanship of the Immediate Past President of the I.E.R.E., Mr. J. Langham Thompson. The Toast of Authors and Overseas Delegates was proposed by Mr. O. W. Humphreys, President of the Institution of Electrical Engineers, who began by recalling that he had himself graduated—in physics—from University College just forty years previously. He suggested that the subject of semiconductor technology was particularly important because it required the closest collaboration of physicist and engineer. Indeed many physicists who began their careers with some years of original research work in this field would find themselves a second career in developing the technology and thus tending to become engineers!

In the past the discoveries in pure physics had often remained virtually hidden in the proceedings of learned societies and lacked practical application for many years. Today the engineer wanted to know what was being done elsewhere—whether this was down the road, across the Channel or across the Atlantic. Meetings such as this symposium were particularly important in enabling engineers to find out at first hand, and quickly, what their opposite numbers were doing.

Mr. Humphreys coupled his Toast with the name of Professor H. H. Meinke, Director of the Institut für Hochfrequenztechnik at Munich Technical University, who, he said, was one of the pioneers of microwave engineering.

Replying, Professor Meinke related some of his experiences from the earliest days of microwave engineering and aroused some amusement in the audience by speaking of the one signal generator operating on one frequency which had been in use at the Telefunken laboratories when he worked there 32 years ago. Referring to the very rapid progress in the field of microwave semiconductor devices he said that the authors were now talking about new work which they had carried out in the months since they submitted their papers earlier this year.

As an educationalist he hoped that the rate of progress in microwave electronics would not continue to increase. In order to teach for the future, the educationalists had some difficulty in knowing what should be taught since the direction of developments with which students would have to deal was often completely new.

Professor Meinke concluded by thanking the Organizing Institutions for arranging the Symposium. His Toast was acknowledged by Mr. Thompson who reminded those present of the close co-operation between the I.E.E. and the I.E.R.E.: this was the second of four joint conferences to be held in London during 1965. Mr. Thompson then expressed the appreciation of all concerned to the authorities of University College and especially welcomed Professor A. Lewis, Head of the Department of Botany where the main meetings were being held.

The Institutions' thanks to University College were again referred to by Professor W. A. Gambling when, as Chairman of the Joint Organizing Committee, he formally closed the Symposium on the following afternoon. He made particular mention of the provision by the Department of Botany of a closed circuit television link between the nearby Engineering Theatre with the larger main Botany Lecture Theatre and which enabled over 400 persons to take part in the Symposium simultaneously.

All the papers and the full proceedings including discussion reports will be published as a complete volume in the Autumn. Certain of the papers are being considered for reprinting in *The Radio and Electronic Engineer*.

F. W. S.

U.D.C. 621.317.757 : 621.3.018.783.3/4

Synchronously Tuned Methods of Harmonic and Intermodulation Distortion Analysis

By

D. E. O'N. WADDINGTON

(*Associate Member*) †

Presented at a meeting of the Electro-Acoustics Group in London on 28th April 1965.

Summary: The need for an instrument to simplify the measurement of distortion as a function of frequency is discussed and existing systems for making these measurements are reviewed.

The new system described in this paper uses a heterodyne wave analyser as the basic unit for carrying out the measurements. By means of a 'beat frequency oscillator' system using the local oscillator from the wave analyser, crystal-controlled oscillators, frequency dividers, mixers and filters, test signals with the requisite purity of wave-form are generated. Switching permits the selection of harmonics up to the 8th for single signal tests, side bands up to the 4th for the S.M.P.T.E. test and intermodulation components up to the 8th for the C.C.I.F. test. The frequency at which the test is carried out is controlled by the wave analyser tuning. Automatic level control ensures that the level of the test signals is held constant throughout any series of tests.

Interesting aspects of the circuits (which employ transistors) used to achieve the necessary signal purity and stability are described and the results on an experimental model are given.

1. Introduction

Non-linear distortion occurs to a lesser or greater extent in every link of an audio frequency reproducing system. In order to ensure that the quality of reproduction is sufficiently good, it is necessary, among other tests, to measure the distortion introduced by the system. The distortion generally manifests itself in two forms, i.e. harmonic distortion and intermodulation distortion. Each of these is frequency dependent and, in order to make an accurate prediction of the performance or to discover what steps should be taken to improve it, it is desirable to investigate the distortion over the whole frequency range. Point by point measurement is, of necessity, very tedious and there is always the possibility that an important frequency may be missed.

2. Current Methods

The need for an apparatus to measure distortion as a continuous function of frequency has been recognized for some years and several equipments have been developed for this purpose. They divide up into the two main distortion groups, harmonic and intermodulation. A brief description of these methods follows.

2.1. *Harmonic Distortion*

This may be measured either as a whole, as in a distortion factor meter, or the individual harmonic components may be measured separately, as by the use of wave-form analyser. Both methods have been applied to automatic distortion analysis.

2.1.1. Distortion factor method

In essence, this method is very simple.[1] The signal to be measured is fed via a high-pass filter with a very fast cut-off rate to a meter (See Figs. 1 and 2). As the signal will be in the stop-band region of the filter while

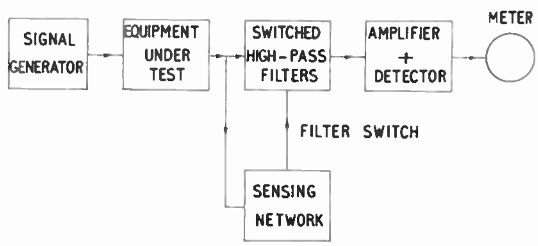


Fig. 1. Automatic total harmonic distortion measuring system.

the harmonics are in the pass-band, the meter will register only the distortion voltages. This method of fundamental rejection gives satisfactory results over about half an octave so that, in the original equipment, only fourteen filters were needed to cover the frequency range from 40 to 15,000 c/s. In use, these filters are switched sequentially by means of a stepping motor which is controlled by a series of bandstop filters which sense the frequency of the test signal.

† Marconi Instruments Ltd., St. Albans, Hertfordshire.

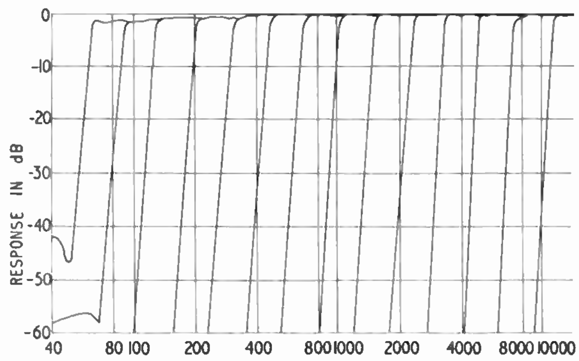


Fig. 2. The frequency response characteristics of the fourteen high-pass filters used in Fig. 1.

The amount of distortion which may be measured by this system, and hence the accuracy, is limited by the design of the high-pass filters. A further failing of the system is that there is no discrimination against noise.

2.1.2. Individual harmonic measurement method

The system used here relies on being able to gang together the main frequency control of a beat frequency oscillator and the selector switch of a 1/3 octave spectrometer.[2] The ratio of the drive is adjusted such that the spectrometer selects the desired harmonic component of the input signal and ensures that this component is always within the pass-band of the spectrometer. (See Figs. 3 and 4.) In practice the filters overlap only at the -2 or -3 dB points with the result that considerable measurement errors occur at or near the switching point.

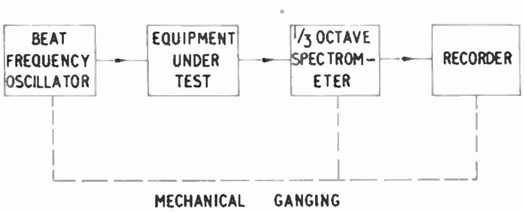


Fig. 3. Automatic individual harmonic analyser system.

2.2. *Intermodulation Distortion*

Two methods of measurement are in current use, namely C.C.I.F.† and S.M.P.T.E.‡ Automatic test equipments have been evolved for both of these and they are more elegant than the methods described in section 2.1 in that selection by means of mechanical switches has been eliminated.

2.2.1. S.M.P.T.E.

This method of intermodulation distortion measurement consists of applying two signals to the equipment under test, one having a low frequency (40–400 c/s) and the other a high frequency (> 1 kc/s) but with 1/4 the amplitude of the low frequency. The distortion products will have the spectrum shown in Fig. 10. There are two methods of measuring these.

The classical method[3, 4] is not a synchronously tuned one but it is as easy to apply because, apart from the fixed high- and low-pass filters, no tuning is necessary. (See Fig. 5.) The test signals are applied to the system under test and, if intermodulation has taken place, the high frequency signal will be amplitude modulated by the low frequency. (See Appendix.) The composite signal is fed via a high-pass filter to a detector. The modulation is selected by means of the low-pass filter and measured on the meter. The intermodulation distortion factor is given by the ratio between the modulation components and the high frequency signal amplitude. This method corresponds to the distortion factor method described in section 2.1.1 and it suffers from similar disadvantages, i.e. it does not discriminate against noise and, unless the filtered signal is observed on an oscilloscope, there is no way of knowing whether the distortion is even or odd order.

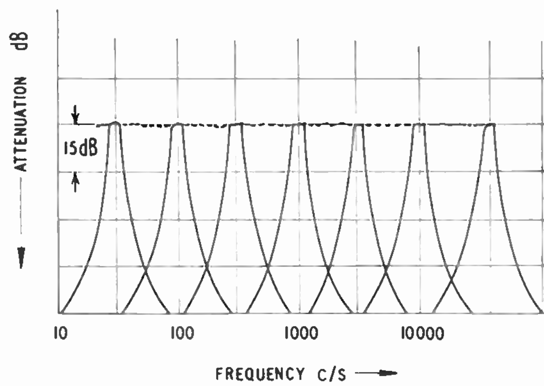


Fig. 4. $\frac{1}{3}$rd octave band-pass filter characteristics.

This type of distortion may also be measured by analysing the individual distortion components. A system[5] for doing this semi-automatically has been devised using a heterodyne wave analyser which includes a beat frequency oscillator. (See Fig. 6.)

Let f_2 = wave analyser intermediate frequency.

f_3 = wave analyser local oscillator frequency.

f_1 = wave analyser input frequency $= f_3 - f_2$.

† Comité Consultatif International de Téléphonique.
‡ Society of Motion Picture and Television Engineers.

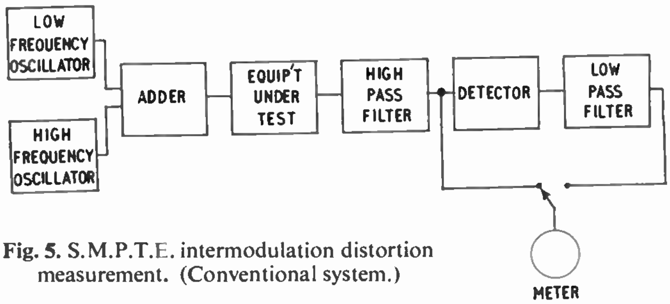


Fig. 5. S.M.P.T.E. intermodulation distortion measurement. (Conventional system.)

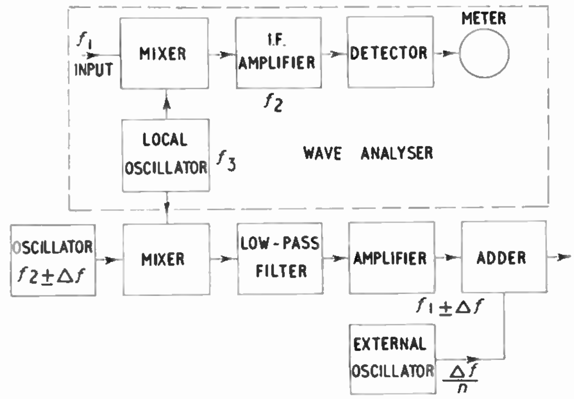


Fig. 6. Semi-automatic S.M.P.T.E. intermodulation distortion measurement. (Radiometer[5].)

The signal having a frequency of $f_1 \pm \Delta f$ is generated by mixing the local oscillator signal f_3 with a signal $f_2 \pm \Delta f$ which is generated by a secondary oscillator in the wave analyser. This provides the high-frequency test signal. The low-frequency test signal is provided by an external oscillator which is set to $\Delta f/n$. These signals are added linearly and applied to the equipment under test. The resultant distortion will be

$$(f_1 \pm \Delta f) \pm \frac{n\Delta f}{n} = f_1$$

As f_1 is the frequency to which the wave analyser is tuned, it will automatically select any preset distortion component.

2.2.2. C.C.I.F.

The method of testing is to feed two signals of equal amplitude but separated in frequency by a constant difference to the equipment under test. A band-pass filter tuned to this difference frequency will automatically select the distortion component due to square-law curvature. The two signals are generated by means of a dual beat frequency oscillator system.[6] (See Fig. 7.)

LO2 is the variable oscillator which controls the final output frequency. Its output is mixed with those of LO1 and LO3, in mixers 1 and 2 respectively giving:

$$f_1 = f_b - f_a$$
$$\text{and } f_2 = f_b - f_c$$
$$\text{Thus } f_1 - f_2 = f_c - f_a$$

As the frequencies of LO1 and LO3 are fixed, the difference frequency $f_1 - f_2$ will be constant. In the original equipment in which this method was used, provision was made for setting the difference frequency by making LO1 variable over a narrow band.

This difference frequency method is limited in that it is only useful for measuring even order distortion components. (See Appendix.)

3. Principles of New Systems

The new systems are centred around a heterodyne wave analyser, together with a series of beat frequency oscillators[7, 8] (Fig. 8.)

3.1. *Synchronously Tuned Harmonic Distortion Analysis*

The method of operation may best be understood by reference to the block diagram of Fig. 9.

Let f_1 = wave analyser input frequency
f_2 = wave analyser intermediate frequency
f_3 = wave analyser local oscillator frequency.
i.e. $f_1 = f_3 - f_2$

The local oscillator output f_3 is divided by a factor n and is applied as the switching signal to the mixer M2. The output of a crystal oscillator, having a frequency f_2, is also divided by the factor n, carefully filtered to remove any distortion, and applied to M2. The resultant difference frequency $(1/n)(f_3 - f_2)$ is then selected by a low-pass filter and amplified to provide the test signal. The level of the output signal is sensed and an automatic level control signal is fed to a voltage-controlled attenuator to ensure that the output level

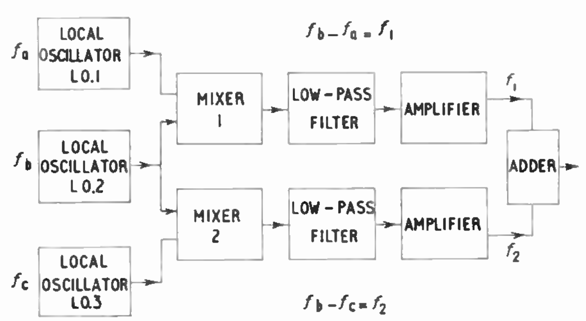


Fig. 7. Dual beat frequency oscillator system for generating two signals with a constant frequency difference (C.C.I.F. method).

remains constant with all settings of the frequency control and dividers. The output signal is applied to the equipment under test and its output, in turn, is fed to the wave analyser input terminals. If $n = 1$, $(1/n)(f_3 - f_2) = f_1$ and thus the basic frequency response of the system under test may be plotted by tuning the wave analyser over its range. However, if $n = 2$, $(1/n)(f_3 - f_2) = f_1/2$ and the wave analyser is thus tuned to the second harmonic of the test signal. Similarly higher order distortion components may be measured by choosing higher values of n. The frequency range of this system is only limited by the wave analyser used.

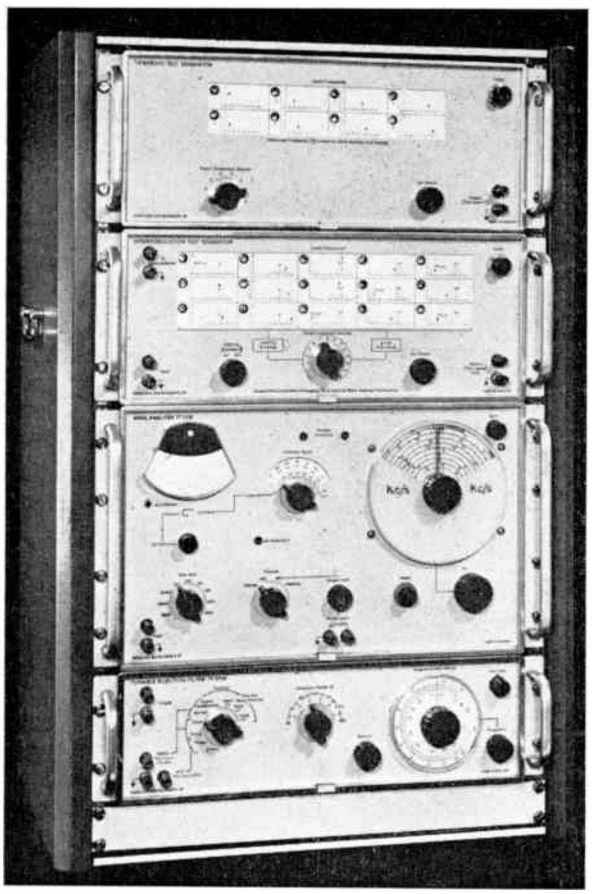


Fig. 8. General view of experimental instrument. The units are (*top to bottom*) harmonic test generator, intermodulation test generator, wave analyser and tunable rejection filter.

3.2. *Synchronously Tuned Intermodulation Distortion Analysis*

Both the accepted methods of test (C.C.I.F. and S.M.P.T.E.) can be carried out automatically using the same basic circuit blocks, but with different switching arrangements.

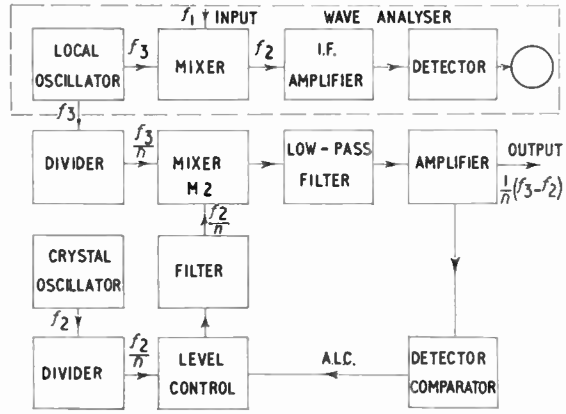


Fig. 9. Synchronously tuned harmonic distortion analyser system.

3.2.1. S.M.P.T.E.

Essentially this system requires the generation of two signals, one having a low frequency, Q, and the other high, P, but with an amplitude of $Q/4$. The distortion products to be measured will have the frequencies $(P \pm Q)$, $(P \pm 2Q)$, $(P \pm 3Q)$, etc. As the spectrum is generally symmetrical (Fig. 10), it is only necessary to measure the distortion products on one side of P, and for convenience in circuit design, the frequencies $(P+Q)$, $(P+2Q)$, etc., were chosen. The principle of operation may best be understood by reference to the block diagram of Fig. 11.

Let f_1 = wave analyser input frequency
f_2 = wave analyser intermediate frequency
f_3 = wave analyser local oscillator frequency
Δf = frequency of test signal Q.

Test signal Q is generated by applying the outputs of two crystal oscillators, XL1 and XL2, having frequencies of $(f_2 + n\Delta f)$ and $[f_2 + (n+1)\Delta f]$ respectively, where n may have values 0, 1, 2, etc., to mixer 2. The difference frequency, Δf, is selected by a low-pass filter and amplified to provide the low-frequency test signal. The output level is sensed and an automatic

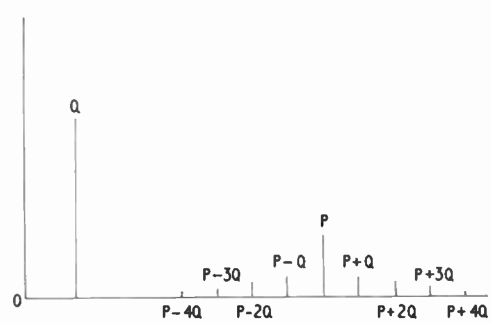


Fig. 10. S.M.P.T.E. intermodulation spectrum.

level control signal is fed back to the crystal oscillator XL2 to control its output so as to ensure that the test signal level remains constant. Test signal P is generated by applying the output of the wave analyser local oscillator and the output of XL1 to mixer 1. The difference frequency $f_3-(f_2+n\Delta f)$ is selected by a low-pass filter, amplified, attenuated by a factor of 12 dB and added linearly to Δf. Again automatic level control is applied. In order to measure the amplitude of Q, the output of the crystal oscillator having a frequency of $f_2+\Delta f$ is fed to the wave analyser mixer in place of the normal local oscillator signal. Thus $f_2+\Delta f-\Delta f = f_2$, the wave analyser intermediate frequency. In order to measure $P(n = 0)$, $P+Q$ $(n = 1)$, $P+2Q$ $(n = 2)$ etc., f_3 is fed to the wave analyser local oscillator as in normal operation, but the frequency of P is modified as shown in Table 1 (2–6) by switching different crystals into the crystal oscillator circuits. As the measuring range of a wave analyser is limited by the input voltage excursion (peak to peak), a band-stop filter tuned to the frequency of Q may be connected in series with the input terminals to the wave analyser. This increases the dynamic range by more than the expected 12 dB as virtually no interfering intermodulation can occur in the wave analyser with only the modulated high frequency applied to it.

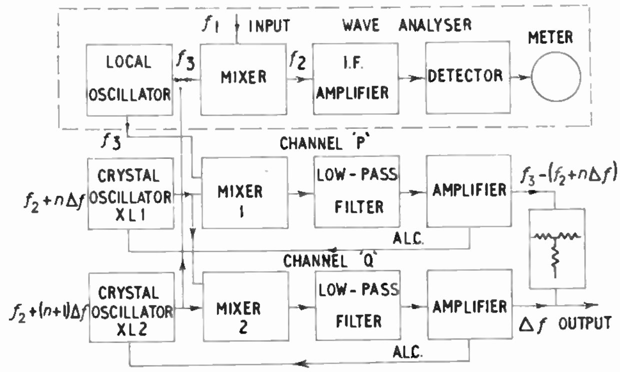


Fig. 11. Synchronously tuned intermodulation distortion measurement using the S.M.P.T.E. system.

Table 1

Signal mixing chart for S.M.P.T.E. system

Local oscillator frequency $= f_3$
Intermediate frequency $= f_2$
Crystal frequencies $= f_2+n\Delta f$
where n has values 0, 1, 2, 3, 4.

Switch position	Signal	Measure	n
1	$Q = (f_2+\Delta f) - f_2$	$Q = \Delta f$	0
2	$P = (f_3 - f_2)$	P	0
3	$P = f_3 - (f_2+\Delta f)$; $Q = (f_2+2\Delta f) - (f_2+\Delta f)$	$P+Q$	1
4	$P = f_3 - (f_2+2\Delta f)$; $Q = (f_2+3\Delta f) - (f_2+2\Delta f)$	$P+2Q$	2
5	$P = f_3 - (f_2+3\Delta f)$; $Q = (f_2+4\Delta f) - (f_2+3\Delta f)$	$P+3Q$	3
6	$P = f_3 - (f_2+4\Delta f)$; $Q = (f_2+4\Delta f) - (f_2+3\Delta f)$	$P+4Q$	4

Note. In positions 3, 4, 5, 6 the frequency of P drops by Δf for each step.

3.2.2. C.C.I.F.

Unlike the S.M.P.T.E. system, in this case the amplitudes of the test signals P and Q are equal and they are separated in frequency by a small difference Δf. The even-order distortion components will have the frequencies $(P\pm Q)$, $2(P\pm Q)$, $3(P\pm Q)$, etc., and the odd-order components will have frequencies $(2P\pm Q)$, $(P\pm 2Q)$, $(3P\pm 2Q)$, $(2P\pm 3Q)$ etc. Again the spectrum should be more or less symmetrical (Fig. 12), so it was decided to measure the difference frequencies which remain within the audio spectrum, i.e. $(P-Q)$, $2(P-Q)$, etc., for even-order distortion products and $(2P-Q)$, $(3P-2Q)$, etc., for odd-order distortion. This incidentally helps to simplify the design.

The system may best be understood by reference to the block diagram of Fig. 13.

Let f_1 = wave analyser input frequency
f_2 = wave analyser intermediate frequency
f_3 = wave analyser local oscillator frequency
Δf = difference between the test frequencies and crystal oscillator frequencies of $f_2+n\Delta f$ where n may have values of 0, 1, 2, 3, 4.

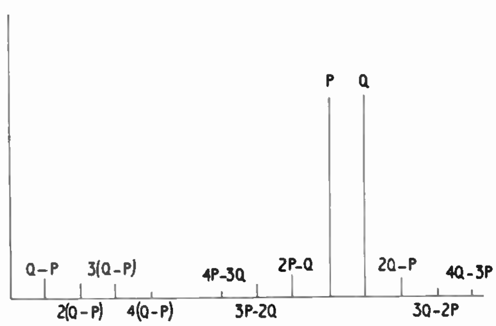


Fig. 12. C.C.I.F. intermodulation spectrum.

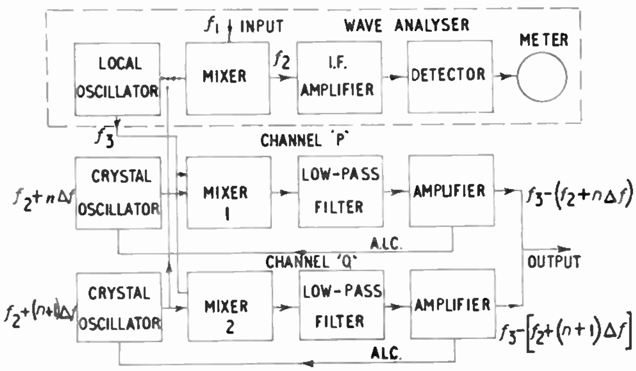


Fig. 13. Synchronously tuned intermodulation distortion measurement using the C.C.I.F. system.

The two test signals are generated by feeding f_3 and $(f_2+n\Delta f)$ to mixer 1 and f_3 and $[f_2+(n+1)\Delta f]$ to mixer 2. In each case the difference frequency is selected by a low-pass filter and then amplified so as to provide the test signals, which are then added linearly. Automatic level control again ensures that the output signal is held constant. If the wave analyser is operated normally (i.e. f_3 is applied to the mixer) the conditions shown in 1, 2, 3, 4, and 5, in Table 2 will apply. Thus the wave analyser will automatically select P, Q, $(2P-Q)$, $(3P-2Q)$, and $(4P-3Q)$, the latter being the odd-order distortion components. The apparent anomaly in position 2 of the switch is resolved by feeding f_2 to mixer 2 and $f_2+\Delta f$ to mixer 1, thus transposing the frequencies of P and Q. As the measurement in position 2 is only a circuit confidence check, this does not affect the performance of the system. In order to select the even-order distortion components, the switching signal to the mixer in the wave analyser will need to be $f_2+n\Delta f$ as shown in 6, 7, 8 and 9 of Table 2.

Table 2

Signal mixing chart for C.C.I.F. system

Local oscillator frequency $= f_3$
Intermediate frequency $= f_2$
Crystal frequencies $= f_2+n\Delta f$
where n has values 0, 1, 2, 3, 4.

Switch position	Signal	Measure	Mix with	n
1	$P = f_3 - f_2$	P	f_3	0
2	$Q = f_3 - f_2$	Q	f_3	0
3	$\begin{cases} P = f_3 - (f_2 + \Delta f) \\ Q = f_3 - (f_2 - 2\Delta f) \end{cases}$	$2P - Q = f_3 - f_2$	f_3	1
4	$\begin{cases} P = f_3 - (f_2 + 2\Delta f) \\ Q = f_3 - (f_2 + 3\Delta f) \end{cases}$	$3P - 2Q = f_3 - f_2$	f_3	2
5	$\begin{cases} P = f_3 - (f_2 + 3\Delta f) \\ Q = f_3 - (f_2 + 4\Delta f) \end{cases}$	$4P - 3Q = f_3 - f_2$	f_3	3
6		$(P - Q) = \Delta f$	$f_2 + \Delta f$	0
7	$P = f_3 - f_2$	$2(P - Q) = 2\Delta f$	$f_2 + 2\Delta f$	1
8	$Q = f_3 - (f_2 + \Delta f)$	$3(P - Q) = 3\Delta f$	$f_2 + 3\Delta f$	2
9		$4(P - Q) = 4\Delta f$	$f_3 + 4\Delta f$	3

Note. In positions 3, 4, 5, the frequencies of P and Q increase by Δf for each step.

Unfortunately, there is no simple way to filter out the test signals so that the wave analyser sees only the distortion products. This limits the dynamic measurement range to that of the wave analyser which, in a well-designed instrument, is of the order of 80 dB; this is sufficient for most purposes.

4. Practical Design Considerations

It will be obvious that these distortion measurement systems are tied very closely to the wave analyser which they are intended to complement and, hence, any design will, to a large extent, derive its pattern from its parent unit. In this case the analyser uses transistors throughout and has an intermediate frequency of 100 kc/s and a local oscillator range extending from 100 to 150 kc/s.

4.1. *Harmonic Analyser*

The two most important design aspects are, naturally enough, inherent distortion of the generated signal and stability of the frequency dividing systems.

4.1.1. Distortion

The distortion will depend upon the following factors:

(*a*) The purity of the signal f_2/n.

(*b*) The linearity of the mixer.

(*c*) The low-pass filter.

(*d*) The amplifier.

Fortunately the wave analyser will 'see' only the distortion component which is being measured and this, before mixing, will always have the same frequency equal to f_2. This means that a simple stop filter in series with the input to the mixer may make up any deficiencies in the tuned filters (see Fig. 14).

The linearity of the mixer depends on how efficiently the diodes are switched. Thus the ideal switching signal is a square wave and, provided that sufficient current is passed through the diodes, the introduced distortion is extremely low. Measurements showed distortion of the order of 0·001%. It is possible however that the distortion measurement was that of the test gear. For the rest of the stages, the distortion

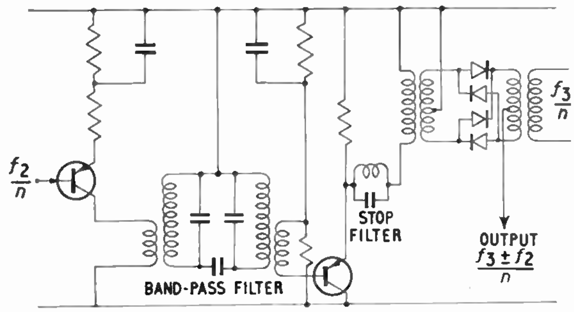


Fig. 14. Essentials of the circuit to ensure the purity of the signal f_2/n and hence of $f_3 \pm f_2/n$.

presented no real problems as the inductors in the low-pass filters were run at a very low magnetization level and sufficient feedback was applied to the output amplifier to ensure linear operation.

During the design one point was overlooked, however, and that was 'breakthrough'. It may be shown that when either f_2/n or $f_3/n = f_3 - f_2$ (i.e. when $f_3 = \frac{f_2(n+1)}{n}$ or $f_3 = \frac{f_2 n}{n-1}$), breakthrough will occur and will show up as excessive distortion. This effect may be reduced considerably by balancing the ring bridge modulator.

4.1.2. Frequency dividing system

The problem of frequency division could easily have been overcome by using a binary system, but this would have had two defects, i.e. it would have used more components and the mark/space ratio of the output would have varied with the dividing factor, thus reducing the proportion of fundamental component available for mixing. For the switching signal f_3/n it is desirable to use a signal with a 1 : 1 mark/space ratio and, in order to achieve this, the frequency

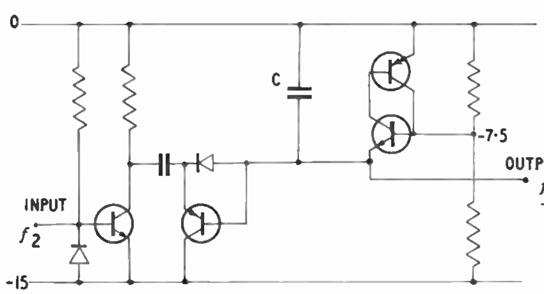


Fig. 15. Pump and trigger type frequency divider circuit.

of f_3 was first doubled by means of a full wave rectifier system. The division $2f_3/n$ was then carried out using a transistor pump staircase circuit and a pnp/npn trigger. The resultant output was then divided by 2 using a conventional Eccles-Jordan bi-stable, thus giving the requisite output waveform The division factor n is determined by the capacitor C (see Fig. 15). By choosing a trigger level of $-7 \cdot 5$ V and making the amplitude of the steps when $n = 8$ equal to 1 volt, there is no uncertainty of the dividing factor even with a large ambient temperature change.

4.1.3. Performance

Only one experimental unit has been made to date and the following inherent distortion figures were obtained with an output of 0·5 V into 600 Ω.

2nd	harmonic	< 0·1% (generally 0·01%–0·05%)
3rd	harmonic	< 0·05% (generally < 0·01%)
4th, 6th, 8th	harmonics	< 0·01% (generally < 0·003%)
5th, 7th	harmonics	< 0·02% (generally < 0·01%)

The spurious responses due to breakthrough noted in section 4.1.1 occur and their amplitude is of the order of 1–2%.

4.2. *Intermodulation Analyser*

In this unit spurious responses may come from two main sources, i.e. stray couplings between the two channels before mixing and non-linearity in the signal combining system.

4.2.1. Stray capacitance coupling

In the initial experimental circuit the first effect was not expected and the circuit was designed using normal transistor circuit impedances of the order of 1–10 kΩ. This was found to be unsatisfactory in as much as the spurious intermodulation was of the order of 1%. Investigation showed that the main cause of the trouble was the stray capacitance between the leads to the two crystal oscillators and particularly between adjacent contacts on the switch wafers. Electrostatic screening was virtually out of the question and a new approach was needed. As an improvement of the order of 100 : 1 was needed it was realized that this could be achieved if the impedance of the inputs of all the amplifier and oscillator circuits could be reduced by this factor. Thus the common base amplifier was brought into service in the input stages with the expected results; the distortion immediately fell to the order of 0·01% or better.

4.2.2. Signal combining system

It is usual to use some form of hybrid network when combining the two test signals for intermodulation measurements. This is generally cumbersome or one

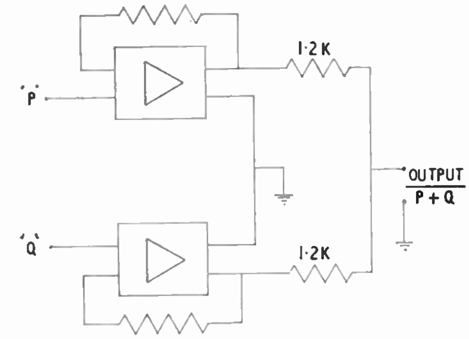


Fig. 16. System for adding the signals P and Q in a linear fashion.

of the signals needs to be balanced with respect to earth and this involves the use of a transformer with all its frequency response problems. A further disadvantage is that it is necessary to terminate the output of the hybrid network with the correct impedance which may be embarrassing in a general purpose instrument.

To overcome these difficulties a completely different approach was used, as shown in Fig. 16. The output resistance of the amplifiers in the two channels was made, by negative feedback, to be $< 0{\cdot}1\,\Omega$. This meant that the signal applied to the output of channel P due to the output of channel Q is reduced by a factor of 24 000 : 1, i.e. 0·004%. Thus the intermodulation occurring would be negligible, and on testing the amplifiers when using two separate signal generators, it was found that the spurious output was of the order of 0·0003%.

4.2.3. Performance

Again, only one unit using this system has been made. The frequency Δf was chosen to be 82 c/s for the following reasons:

(*a*) The band-pass characteristic of the wave analyser used was such that at 82 c/s off tune the response could be relied to be down by more than 80 dB.

(*b*) This frequency is well clear of most international mains frequencies, both fundamental and harmonic.

The spurious responses were worse than would be achieved in a fully developed unit, but the first even-order component in each case was less than 0·05% and the rest of the components were less than 0·01%.

5. Acknowledgments

The author would like to thank Marconi Instruments Limited for permission to publish this article.

6. References

1. H. F. Olson and D. H. Pennie, "An automatic nonlinear distortion analyser", *R.C.A. Review*, **12**, No. 1, pp. 35–44, March 1951.
2. H. Petersen, "Measurement on tape recorders", *Bruel & Kjær Tech. Rev.*, July 1957, pp. 5–11.
3. D. E. O'N. Waddington, "Intermodulation distortion measurement", *Wireless World*, **70**, p. 271, June 1964.
4. F. E. Terman and J. M. Pettit, "Electronic Measurements", p. 335, 2nd Edn. (McGraw-Hill, New York 1952).
5. Handbook on the Radiometer FRA3 Wave Analyser. Radiometer AB, Copenhagen.
6. A. P. G. Petersen, "Intermodulation distortion", *General Radio Experimenter*, **25**, No. 10, March 1951.
7. H. D. Harwood, "Apparatus for Measurement of Non-Linear Distortion as a Continuous Function of Frequency" B.B.C. Engineering Division Monograph No. 49, July 1963.
8. H. D. Harwood, British Patent No. 913,813.

7. Appendix

The transfer function of an amplifier may be written as

$$V_{out} = a_1V + a_2V^2 + a_3V^3 \ldots\ldots\ldots + a_nV^n$$

If two signals $V_1 = (P\sin pt + Q\sin qt)$ are applied, the expression may be expanded as follows:

$$V_{out} = \tfrac{1}{2}a_2P^2 + \tfrac{1}{2}a_2Q^2 + \tfrac{3}{8}a_4P^4 + \tfrac{3}{8}a_4Q^4 + \tfrac{3}{4}a_4P^2Q^2$$
$$+(a_1P + \tfrac{3}{4}a_3P^3 + \tfrac{3}{2}a_3PQ^2 + \ldots\ldots)\cos pt$$
$$+(a_1Q + \tfrac{3}{4}a_3Q^3 + \tfrac{3}{2}a_3P^2Q + \ldots\ldots)\cos qt$$
$$+(\tfrac{1}{2}a_2P^2 + \tfrac{1}{2}a_4P^4 + \tfrac{3}{4}a_4P^2Q^2 + \ldots)\cos 2pt$$
$$+(\tfrac{1}{2}a_2Q^2 + \tfrac{1}{2}a_4P^4 + \tfrac{3}{4}a_4P^2Q^2 + \ldots)\cos 2qt$$
$$+(\tfrac{1}{4}a_3P^3 + \ldots\ldots\ldots\ldots)\cos 3pt$$
$$+(\tfrac{1}{4}a_3Q^3 + \ldots\ldots\ldots\ldots)\cos 3qt$$
$$+(\tfrac{1}{8}a_4P^4 + \ldots\ldots\ldots\ldots)\cos 4pt$$
$$+(\tfrac{1}{8}a_4Q^4 + \ldots\ldots\ldots\ldots)\cos 4qt$$
$$+(a_2PQ + \tfrac{3}{2}a_4P^3Q + \tfrac{3}{2}a_4PQ^3 + \ldots)\cos(p+q)t \qquad \text{s}$$
$$+(a_2PQ + \tfrac{3}{2}a_4P^3Q + \tfrac{3}{2}a_4PQ^3 + \ldots)\cos(p-q)t \qquad \text{c s}$$
$$+(\tfrac{3}{4}a_3P^2Q + \ldots\ldots\ldots\ldots)\cos(2p+q)t \qquad \text{c}$$
$$+(\tfrac{3}{4}a_3P^2Q + \ldots\ldots\ldots\ldots)\cos(2p-q)t \qquad \text{c}$$
$$+(\tfrac{3}{4}a_3PQ^2 + \ldots\ldots\ldots\ldots)\cos(p+2q)t \qquad \text{s}$$
$$+(\tfrac{3}{4}a_3PQ^2 + \ldots\ldots\ldots\ldots)\cos(p-2q)t \qquad \text{s}$$
$$+(\tfrac{3}{8}a_4P^2Q^2 + \ldots\ldots\ldots\ldots)\cos 2(p+q)t$$
$$+(\tfrac{3}{8}a_4P^2Q^2 + \ldots\ldots\ldots\ldots)\cos 2(p-q)t \qquad \text{c}$$
$$+(\tfrac{1}{2}a_4P^3Q + \ldots\ldots\ldots\ldots)\cos(3p+q)t$$
$$+(\tfrac{1}{2}a_4P^3Q + \ldots\ldots\ldots\ldots)\cos(3p-q)t$$
$$+(\tfrac{1}{2}a_4PQ^3 + \ldots\ldots\ldots\ldots)\cos(p+3q)t \qquad \text{s}$$
$$+(\tfrac{1}{2}a_4PQ^3 + \ldots\ldots\ldots\ldots)\cos(p-3q)t \qquad \text{s}$$

s = component measured by S.M.P.T.E. system
c = component measured by C.C.I.F. system.

Manuscript first received by the Institution on 12th January 1965, and in final form on 26th April 1965. (Paper No. 999/EA 22)

DISCUSSION

Under the chairmanship of Mr. C. T. Chapman

Mr. D. A. Barlow: Mr. Waddington is to be congratulated on producing equipment which makes distortion measurements simple to carry out. Often, distortion measuring apparatus is difficult to use and requires continual checking of the zero, such that it takes perhaps half a morning to carry out one set of measurements. Mr. Waddington's equipment makes it possible to carry out such measures almost on a routine basis, and continuously throughout the whole frequency range. Presumably the equipment is stable, with no tendency to drift. Perhaps Mr. Waddington would comment on this.

The second point is whether there is any simple method of checking the equipment. The ease of operation has to be paid for in terms of the complexity of the apparatus, and with all respect to Mr. Waddington, the more complex the apparatus is, the more likely it is to break down.

Mr. D. E. O'N Waddington (*in reply*): Firstly, I would like to say that the credit for the system should go to Mr. H. D. Harwood on whose invention the design is based.

The important part of the frequency determining system is crystal controlled and, in order to prevent differential drift, the crystals are kept in a common ambient temperature. This is achieved by simply building the system in an unventilated box. The drift in the local oscillator, which is permeability tuned, is of the order of 1 c/s per deg C.

While I agree that the reliability of a system is adversely affected by increased complexity, I feel that this is a generalization which cannot always be taken at its face value. The number of components in each of the test generator units is roughly equal to that in the wave analyser, i.e. 30–40 transistors, 30–40 diodes and resistors, capacitors, inductors and quartz crystals to match. Environmental tests have shown the wave analyser to be thoroughly reliable and there is no reason to suppose that the other units will be any different. Self checking is achieved by simply connecting the output of the generator concerned directly to the input of the wave analyser. Distortion may easily be simulated by overloading.

Mr. R. L. West: Would Mr. Waddington like to comment on the technique that subtracts a suitably phased input signal from the total output as a first step in the analysis of the output signals.

Mr. Waddington (*in reply*): This technique has been applied to total harmonic distortion measurement but, due to the difficulty of adjusting the phase to effect satisfactory cancellation, it has not been widely used. This method does not permit the use of a distorted test signal due to the added difficulty in ensuring that the phasing of harmonics is also correct.

Mr. H. D. Harwood: I would like to compliment the author on the excellent results which he has achieved in translating my valve-operated instrument into a transistor version suitable for production. The very low levels of distortion which can be measured by it should satisfy any needs for some time to come.

There is one feature of the whole design, however, which I would like to stress a little more, and that is the facility of making measurements as a continuous function of frequency. When the instrument was first designed one of the principal objects was to measure the non-linearity distortion of loudspeakers. Experience has indicated the necessity of doing this as a continuous function of frequency for at least two reasons. In the first place, it has been shown[7] that even in the frequency range where a high grade monitoring loudspeaker may have an axial response uniform to $\pm$ 1 dB, harmonic distortion of the differing orders may vary irregularly with frequency by 10 to 15 dB. In addition, at some frequencies at which one harmonic is at a maximum, another order may show a minimum, so that the relative levels of these two harmonics will vary over a few cycles per second by as much as 25 dB. Measurements at only a few spot frequencies would, in these circumstances, be extremely misleading. Networks containing iron-cored components have also been shown to have harmonic distortion levels which vary irregularly with frequency and therefore the ability to measure these distortions as a continuous function of frequency is very important.

Mr. M. S. Leak: Would Mr. Waddington tell us whether the equipment he describes is suitable for making measurements in accordance with the recently issued British Standard for distortion. This Standard (B.S. 3860 : 1965) calls for harmonic distortion to be measured as the root sum square of all distortion products present and none of the distortion measuring equipments known to me measures in this way—most measure either the peak or peak to peak value of the distortion products in the waveform. The Standard also requires that the intermodulation distortion shall be expressed as the root sum square of the various modulation components appearing on the high frequency carrier in the method of testing referred to by Mr. Waddington as the S.M.P.T.E. method but which we now might call the British Standard method.

Secondly, could Mr. Waddington give any idea of an acceptable or perceivable level of intermodulation since I have been unable to find in the literature any solid figures for this?

Finally, would Mr. Waddington state whether there are difficulties in using his equipment due to the coincidence of harmonics and intermodulation products when scanning a frequency range? I refer to the points at which for instance the product $(2f_2-f_1)$ occurs at the same frequency as $2f_1$, f_1 and f_2 being the two test tones. When scanning for intermodulation this, it seems to me, will give an abnormally high reading which, unless this point is appreciated, may be taken as a point of bad intermodulation.

Mr. Waddington (*in reply*): Since the equipment measures the r.m.s. value of each distortion component individually, it is a simple matter to calculate the overall distortion.[4]

Regarding Mr. Leak's second point, several papers have been written on this subject but, as it is a purely subjective effect, no ruling can be given.†

Finally, this effect does not occur in this equipment as this product $(2f_2-f_1)$ is not measured. It is assumed that it will be equal to $(2f_1-f_2)$.

Dr. T. G. Hammerton: I have followed Mr. Waddington's paper, including his methods of analysis, with interest. Would he comment on a further method of measurement of intermodulation—namely one that has been used on wide-band line communication systems. These systems are usually operated at somewhat higher frequencies than the base-band audio channels but it is necessary to keep intermodulation levels very low and to have a means of assessing them. The method is to apply a white noise signal covering the full frequency band. A narrow 'slot' filter whose centre frequency can be varied, is inserted between the noise sources and the system input terminals. The system output under these conditions is then examined. The amount of intermodulation introduced by the system is then determined by the amount of additional noise found in the resulting output frequency slot. I am aware of course of the difficulties of producing such a slot filter for low frequencies, and for these measurements, it is necessary to limit the noise amplitude to within the maximum acceptable input levels of the system under examination.

Mr. Waddington (*in reply*): The slotted white noise test is ideal for checking amplifiers at the end of a production run. However, very little work has been done on the diagnostic capability of this sort of test. To date, the tests such as have been described yield far more useful information.

Mr. P. M. Clifford: Some caution should be exercised when using white noise to test an active network. True white noise contains components of infinite amplitude—albeit infinitely rarely—and these can cause errors by overloading amplifiers etc. No problem is introduced on a purely passive network, as due to the short duration of the large components, the heating effect is negligible.

There is probably a case for using 'random noise' whose spectral character and amplitude distribution are known, for 'white noise' testing of active networks.

The Hon. J. C. G. Dawnay: I would imagine that a pen recorder with logarithmic potentiometer is used for the read-out of the equipment. As long as pre-printed calibrated paper is available this should make the measurement very quick. Is this so?

My experience with normal non-resonant electronic circuitry (assuming that one is running wound components within their limits) is that a measurement of harmonic or intermodulation distortion taken at one frequency within the bandwidth suffices. There is normally sufficient feedback to control operating points and non-linearities quite precisely, and any fall off in performance due to reduced loop gain at the edges of the operating band can be predicted. I am not trying to say that Mr. Waddington's equipment is not highly useful as well as ingenious, but in normal development work on linear electronic circuitry a good wave analyser intelligently used provides the information which one needs. Does Mr. Waddington agree?

Mr. Waddington (*in reply*): Although the system is primarily intended for use with a recorder, only one prototype has been built and it was not convenient to include this facility.

I am afraid that I cannot agree with Mr Dawnay's suggestion. The relationship between harmonic and intermodulation distortion is very difficult to define, particularly in regions where the basic frequency response begins to fall off. Figures are quoted in a previous article[3] and also by Le Bel.‡

† F. Langford-Smith, "Radio Designer's Handbook." p. 603 et seq., 4th edition. (Iliffe, London, 1954.)

E. R. Wigan, "New distortion criterion", *Electronic Technology*, **38**, pp. 128–37, 163–74, April and May 1961.

‡ C. J. Le Bel, "An experimental study of distortion", *J. Audio Engng Soc.*, **2**, No. 4, pp. 215–8, October 1954.

COMMONWEALTH EXCHANGES IDEAS ON STANDARDS

Fourteen Commonwealth countries took part in a week of wide-ranging policy discussions in London during July, at the sixth Commonwealth Standards Conference. Meeting in London for the first time since 1951—intervening conferences have been held in New Delhi, Ottawa and Sydney—the thirty delegates were guests of the British Standards Institution. B.S.I.'s director, Mr. H. A. R. Binney, was chairman of the Conference.

An important development this year has been the participation of Commonwealth countries not previously represented—in particular, Jamaica, Nigeria, Hong Kong, Mauritius and Malta. Their presence helped to emphasize the special character of the conference—a forum not influenced by political considerations or linguistic problems, but one in which members meet to exchange ideas, to discuss mutual problems and to benefit from each other's technical experience. All the participating countries—Australia, Canada, Central Africa, India, Ireland, New Zealand, Pakistan, South Africa and this country, as well as the newcomers—had much to contribute to, and learn from, the discussions on various aspects of standardization policy that took place during the week.

At the top of the agenda came the metric system, a subject which attracted particular interest in view of the British Government's decision to adopt the system, announced in May.

B.S.I.'s associate director, Mr. Gordon Weston, outlined the programme now being planned in the U.K. to implement the Government announcement. Basic standards were being prepared; there were arrangements for increased publicity; the whole question of educational policy—particularly in regard to textbooks for schools, universities and technical colleges—was being tackled. Immediate steps were being taken in B.S.I. to consult industry on which standards should first be made available in metric units.

Indian viewpoint

The British move was in general greeted with approval, but there was some surprise that no decision had been taken to decimalize currency. Dr. Lal C. Verman, director of the Indian Standards Institution, pointed out that, in India, where the illiteracy rate was high, the metric system had been accepted surprisingly quickly because of its early introduction in commerce soon after decimalization of the currency. It was convenient 'for the common man and the engineer' if the change could be undertaken in every aspect once a country had decided to adopt the metric principle.

Dr. Verman described how the preparation of metric standards had been tackled by his own organization. A small 'metric cell' had been established, responsible for co-ordinating the approach to the problems of the different technical departments; and it reviewed all draft standards before circulation. It also helped to solve particular problems referred to it by industry and government.

The work of I.S.I.'s implementation department formed a separate, but parallel, operation. This was not designed to deal specifically with metric problems, but could do much to promote the acceptance of metric standards by—among others—industry, commerce, and the central and state governments. It organized conferences in the different Indian States and also for particular industries; operated an advisory service on the implementation of standards, and ran a number of training courses.

The I.S.I. worked closely with the Indian Government during the period of transition. It was the Government's responsibility to promote the system in commerce and education, and to see that weights and measures inspectorates were properly equipped and trained to supervise its introduction. New comprehensive legislation had been necessary to define metric units, outline the responsibility of individual States, and indicate a timetable for the completion of the changeover. The Government was able to give initial impetus through its public services—railways, postal communications, broadcasting and information services.

Although some conference delegates would not *urge* countries to make the metric change, all were agreed that such a change would greatly ease the work of reaching international standards agreements. The Commonwealth standards bodies could, it was felt, usefully undertake or encourage the initiation of detailed inquiries into the industrial needs and economic implications associated with a changeover, in line with the studies undertaken in India, South Africa and the United Kingdom.

Speeding standards

A session on "Streamlining standards" made it clear that Britain is by no means alone in her current anxiety to speed up standards work. India and South Africa, as well as the U.K., have recently made special plans to hasten standards production, and some aspects of these were discussed. There was general agreement, for instance, on the usefulness of circulating a preliminary draft before the technical committee first met. In South Africa, a project was investigated in detail before it was accepted into the standards programme, and this 'pre-acceptance investigation' was extremely useful as a preliminary to the proper planning of the work.

Informative labelling

A considerable and growing interest in the possibilities of informative labelling was evident at the conference, either as a useful supplement to the marking requirements of a standard or as an interim solution where no comprehensive standards for certain products exist. Great interest was shown in the project being developed in the U.K. by the Consumer Council with B.S.I.'s co-operation, for an informative 'Teltag' to give information on various aspects of performance, materials and construction for items such as electrical appliances and textiles.

The essential, it was agreed, was to try to ensure that the information given was accurate and based on standard methods of test (where possible internationally agreed). In some instances it might be appropriate for informative labelling to be controlled by a standards organization; in others, by an outside organization working with the standards body. But in either case it was important to see that the scheme was soundly based on the use of national standards as recommended in Draft I.S.O. Recommendation 662 *Informative labelling*. The U.K. scheme, it was noted, fulfilled these requirements.

Approvals work

Arrangements were made during the conference for delegates to visit the B.S.I. Test Centre at Hemel Hempstead—probably the fastest-growing section of the Institution. Mr. Binney, at a subsequent session, said that expansion at the Test Centre had been continuous during the first six years of its existence. It was, he said, symptomatic of the now world-wide tendency to ally standards and approvals work. In the developing countries, the provision of central testing facilities—in spite of the difficulties of finding suitable staff—was extremely important.

The South African delegation described the degree of acceptance attained by the SABS† certification mark and the importance now attached to it by both Government and consumer. Indeed, provision was now being made by the South African government to pay a preferential price for certified products, since buying these effected considerable economies for the purchaser in eliminating the need for extensive testing. The scheme was not limited to large-scale Government suppliers.

In general, certification fees added $\frac{1}{4}\%$ to the price of a product; this should be compared with the cost of inspecting and testing unmarked goods, calculated at some 4 to 6% of the purchase price.

† South African Bureau of Standards.

Although it was not thought desirable to certify partial conformity with a standard, the conference felt that a good case could be made for certifying safety features alone in some cases, identified perhaps by a special 'safety' mark. In view of the degree of common interest in approvals schemes, arrangements were made for an exchange of information on certification marking and on the legal position relating to the enforcement of marking in individual countries.

Government influence

It was clear that Government policies could have a very great influence on the prestige of a standards organization and, indeed, on the importance of its work.

In South Africa, a Central Standardizing Committee had been set up on the authority of the Prime Minister, with the object of co-ordinating the purchasing requirements of government, provincial authorities and other public bodies through insistence on purchases being made to national standards. Where no suitable national standard exists and where there is an urgent demand for an agreed specification the scheme allows for special co-ordinating specifications (known as CKS) to be issued very quickly. These are issued only for those products purchased by a number of different departments. The intention is that they be re-issued in due course as SABS standards.

Although similar schemes existed in India, New Zealand and Canada, much interest was shown by the conference in the South African project, since it showed a new and wider approach to the subject. Perhaps the most valuable part of the project was the feed-back of information to SABS on differing purchasing requirements.

Mr. Binney referred to the growing recognition in the U.K. of the need for British Standards to take proper account of export requirements, and for the requirements of large purchases (for instance, the nationalized industries) to be fitted into this framework. The National Economic Development Council was very much behind this effort. The liaison between the Government and B.S.I. in the development of Defence Specifications was described by Mr. Weston; where Government requirements had a wider application, they would normally be specified in British Standards. Mr. Weston mentioned a new series of British Standards for electronic components that would take account of the needs of Government and of other users of precision components, while less stringent requirements for more general purposes would be in the same standards.

U.D.C. 621.372.853.1

Dielectric Loaded Waveguides— A Review of Theoretical Solutions

By

S. K. CHATTERJEE, M.Sc.† AND (Mrs) R. CHATTERJEE, Ph.D., M.S.E., M.Sc.†

CONTENTS

Part I. Mathematical Methods

Part II. Propagation through Dielectric Loaded Waveguides

Part III. The Impedance Approach to Dielectric Loaded Waveguides

† Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India.

List of Principal Symbols

E	electric field intensity vector in volts per metre
H	magnetic field intensity vector in amperes per metre
B	magnetic induction vector in webers per metre
D	electric displacement vector in coulombs per metre
J	current density vector in amperes per metre
Π_e	electric Hertz vector
Π_m	magnetic Hertz vector
ω	angular frequency in radians per second
ρ	volume density of charge in coulombs per metre
ε'	real part of permittivity
ε''	imaginary part of permittivity
ε_0	permittivity of free space $= 8{\cdot}854 \times 10^{-12}$ farad per metre
μ_0	permeability of free space $= 4 \times 10^{-7}$ henry per metre
γ	propagation constant
α	attenuation constant
β	phase constant
η	surface charge density
ψ_e	electric eigenfunction
ψ_m	magnetic eigenfunction
δ	variational notation
λ_0	free space wavelength
λ_c	cut-off wavelength

Part I: Mathematical Methods

Summary: Different theoretical methods of approach, namely, field theory, network theory, variational and Rayleigh-Ritz procedure, reaction concept, and finite-difference methods used to determine the propagation characteristics of microwaves through metallic waveguides filled inhomogeneously with dielectrics are outlined. Classification of fields, boundary conditions and the orthogonality properties of modes are also discussed.

1. Introduction

The study of the radiation of electromagnetic waves started by Sir Oliver Lodge,[1]† Lodge and Howard,[2] and Thomson,[3] and the propagation of waves through metallic waveguides of different geometrical cross-sections initiated by Lord Rayleigh,[4] followed by Sommerfeld,[5] Hondros and Debye,[6] Barrow,[7] Southworth,[8] Carson, Mead and Schelkunoff,[9] Chu and Barrow[10] has yielded valuable information about the propagation characteristics of allowed modes. The attempt had been mainly to study the propagation through an isotropic, homogeneous, source-free and loss-free medium bounded by a metallic wall. The results of these investigations have been presented in an illuminating manner by Borgnis and Papas.[11] Recently, several authors[12–55] have investigated the propagation characteristics of microwaves through metallic waveguides of infinite conductivity partially or completely filled with one or two different dielectrics. Several authors[56–67] have solved the case of microwave propagation in metallic waveguides having walls of imperfect conductivity and inhomogeneously filled with dielectrics. In studying the propagation characteristics of electromagnetic waves in waveguides containing gyromagnetic media Waldron[68–70] has treated the case of waveguides containing dielectrics as a special case. In most cases of waveguide propagation we are primarily interested in the forward waves, i.e. in which the energy and wave pattern travel in the same direction, and as such the study of wave propagation by the above authors has been concerned with the forward waves. It has been found by Clarricoats and Waldron[71] that a waveguide of circular cross-section containing a concentric dielectric rod of smaller radius than the guide can support also backward waves. A detailed analysis of the existence of backward waves in such structures has been made by Clarricoats.[72, 73] The contribution by Brown[74] on backward waves in homogeneous waveguides is also significant. Waldron,[75] while considering the properties of inhomogeneous waveguides in the neighbourhood of cut-off, has also discussed the backward waves. Coaxial lines with dielectrics separated by radial planes have also been recently investigated by several authors.[76–80] An important aspect of the problem is the study of impedance characteristics[81–86] of such composite waveguides and much remains to be done in this direction.

† References will be found at the end of Part III of the paper.

2. Maxwell's Field Equations

Maxwell's electromagnetic field theory is based on the two field equations

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \qquad \text{......(1)}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{......(2)}$$

and the two scalar equations

$$\nabla.\mathbf{D} = \rho \qquad ......(3)$$

$$\nabla.\mathrm{B} = 0 \qquad ......(4)$$

The equations (1) and (2) are the generalization of Ampere's circuit and Faraday's induction laws respectively, which were originally formulated to explain the existence of currents and voltages in a loop of wire. Maxwell's field equations postulate that Ampere's and Faraday's laws are also valid in free space independent of the presence of conductor. The equations (3) and (4) define the true electric charge density and postulate the non-existence of true magnetic monopoles respectively. For sinusoidal fields in a linear, isotropic and source-free dielectric medium, the equations (1) and (2) reduce to

$$\nabla\times\mathbf{H} = \varepsilon\frac{\partial \mathbf{E}}{\partial t} \qquad ......(5)$$

$$\nabla\times\mathbf{E} = -\mu\frac{\partial \mathbf{H}}{\partial t} \qquad ......(6)$$

and the scalar equations reduce to

$$\nabla.\mathbf{E} = 0 \qquad (7)$$

$$\nabla.\mathbf{H} = 0 \qquad (8)$$

In solving boundary value problems, it is sometimes convenient to write the vector differential operator ∇ as the sum of the two operators $\nabla = \nabla_t + \nabla_n$ where, ∇_t is that portion of ∇ which represents differentiation with respect to the coordinates on the discontinuity surface S and ∇_n is the operator which differentiates with respect to the coordinates normal to S. The curl equation for the electric field then becomes

$$\nabla_t\times(\mathbf{E}_t+\mathbf{E}_n)+\nabla_n\times(\mathbf{E}_t+\mathbf{E}_n) = -j\omega\mu(\mathbf{H}_t+\mathbf{H}_n) \quad ......(9)$$

when the time variation is exp ($j\omega t$).

Maxwell's field equations therefore present a physical description of the link between the vectors **E** and **H** and their interaction with matter in space and time. The two parameters ε and μ describing the nature of the space containing the two field vectors **E** and **H** are, in general, complex and are given as follows

$$\varepsilon^* = \varepsilon' - j\varepsilon'' \qquad ......(10)$$

$$\mu^* = \mu' - j\mu'' \qquad ......(11)$$

The complex nature of ε and μ leads to the concept of energy absorption and frequency dependence of the dielectric properties of the medium through which an electromagnetic wave is propagated.

3. Solutions of Maxwell's Field Equations: Classification of Fields in Waveguides

Solutions of Maxwell's equations in orthogonal curvilinear coordinates (u_1, u_2, u_3) are obtained by assuming the time variation of the field vectors of the following form

$$\mathbf{E}(u_1,u_2,u_3) = |\mathbf{E}(u_1,u_2,u_3)| \exp(\pm j\omega t) \quad ......(12)$$

$$\mathbf{H}(u_1,u_2,u_3) = |\mathbf{H}(u_1,u_2,u_3)| \exp(\pm j\omega t) \quad ......(13)$$

The differential lengths of arcs ds_1, ds_2, ds_3 measured along the coordinate surfaces u_1 = constant, u_2 = constant, u_3 = constant, are given by the relations,[87]

$$ds_1 = h_1\,du_1, \qquad ds_2 = h_2\,du_2, \qquad ds_3 = h_3\,du_3$$

where the metric coefficients h_1, h_2, h_3 are the quantities by which increments in the variables must be multiplied to give distances measured along the coordinate axes. The intersections of the coordinate surfaces pair by pair are known as the coordinate lines. The tangents to the coordinate lines at a point are called the coordinate axes at that point. With each point P (x, y, z) in the region, there is associated a triplet u_i, i = 1, 2, 3 given by the functions $u_i = u_i$ (x_1, x_2, x_3) which are continuously differentiable functions of the three variables x_i. If such functions are chosen so as to make the coordinate axes at every point in space mutually perpendicular, the coordinates are known as orthogonal curvilinear coordinates. Most physical problems are best handled by means of orthogonal coordinates and only such will be considered. The condition that the curvilinear coordinates should form an orthogonal system is that

$$\nabla u_s . \nabla u_r = 0 \qquad \text{for} \quad r \neq s$$

In order to define a waveguide mode, the usual mathematical technique adopted is to solve Maxwell's equations in appropriate coordinate systems. The solution yields the field components in terms of certain arbitrary constants. The characteristic equation is obtained by applying proper boundary conditions and making the determinant of the coefficients of the arbitrary constants equal to zero. The eigenvalues of the system are obtained by solving the characteristic equation. The arbitrary constants are evaluated and the distinct set of field components obtained which corresponds to each eigenvalue defines a particular mode. For source-free regions (J = 0, ρ = 0), solution of Maxwell's equations can be obtained from two scalar functions, which may be chosen in different ways,[88] and each scalar function satisfies the homogeneous wave equation. A convenient way of expressing the solution is in terms of two independent solutions representing E (TM) and H (TE) waves. This classification is justified for the types of electromagnetic waves that can be propagated inside a perfectly conducting guide[74, 75] uniformly filled with a homogeneous dielectric medium. This type of waveguide is only a mathematical idealization of an actual waveguide having very small loss. If the conductivity of the boundary walls of a metallic waveguide filled with air which has very low dielectric loss is infinitely high, then this type of waveguide may be considered to be a good approximation to the homogeneous simple perfect waveguide (HSP).[81] For all practical purposes, it is sufficient to consider the field in a loss-less guide as being the same except for an attenuation constant as those of a physically realizable waveguide. Karbowiak[89] has attempted to introduce the concept of 'proper mode' and 'quasi-mode' in discussing the classification of waveguide modes. It is said that a proper mode is a mathematical concept which is usually 'concomitant with the concept of eigen-function', but the term 'quasi-mode' is applied to the field configurations which can be supported by a physically realizable guiding structure.

The field components of E wave ($H_z = 0$) and H wave ($E_z = 0$) are derived respectively from an electric Hertz vector $\mathbf{\Pi}_e$ and a magnetic Hertz vector $\mathbf{\Pi}_m$ directed along the u_3 axis. The vectors satisfy, respectively, the following two equations

$$\frac{1}{h_1 h_2}\left[\frac{\partial}{\partial u_1}\left(\frac{h_2}{h_1}\frac{\partial \Pi_e}{\partial u_1}\right)+\frac{\partial}{\partial u_2}\left(\frac{h_1}{h_2}\frac{\partial \Pi_e}{\partial u_2}\right)\right]+ + \frac{\partial^2 \mathbf{\Pi}_e}{\partial u_3^2}+\frac{\omega^2}{c^2}\mathbf{\Pi}_e = 0 \qquad \ldots\ldots(14)$$

and

$$\frac{1}{h_1 h_2}\left[\frac{\partial}{\partial u_1}\left(\frac{h_2}{h_1}\frac{\partial \Pi_m}{\partial u_1}\right)+\frac{\partial}{\partial u_2}\left(\frac{h_1}{h_2}\frac{\partial \Pi_m}{\partial u_2}\right)\right]+ + \frac{\partial^2 \mathbf{\Pi}_m}{\partial u_3^2}+\frac{\omega^2}{c^2}\mathbf{\Pi}_m = 0 \qquad \ldots\ldots(15)$$

In studying the propagation of waves through a cylindrical waveguide, the transverse electric modes may be derived from the vector $\mathbf{\Pi}_m$ having a single component directed along the axis of propagation, whereas the transverse magnetic modes may be derived from the vector $\mathbf{\Pi}_e$ with a single component along the axis of propagation. The electromagnetic field obtained by superposing the partial fields obtained from $\Pi_{e(z)}$ and $\Pi_{m(z)}$ represents the most general solution[90] of Maxwell's equations within the source-free region of the waveguide, such that it is easy to satisfy a prescribed set of boundary conditions on any cylindrical surface whose generating elements are parallel to the z-axis, i.e. on any coordinate surface (u_1 = constant) or on either a surface of the family defined by u_2 = constant or a plane z = constant. The choice of these orthogonal families is restricted to coordinate systems in which the equations are separable.

Since only the propagating modes are of interest, the following form of solution may be assumed for $\mathbf{\Pi}_e$ and $\mathbf{\Pi}_m$

$$\mathbf{\Pi}_e = \mathbf{i}_z \psi_e(u_1, u_2)\exp(\pm\gamma z) \qquad \ldots\ldots(16)$$

and

$$\mathbf{\Pi}_m = \mathbf{i}_z \psi_h(u_1, u_2)\exp(\pm\gamma z) \qquad \ldots\ldots(17)$$

where the functions ψ_e and ψ_h satisfy the two dimensional scalar Helmholtz equations as follows

$$\nabla_t^2 \psi_e + k_c^2 \psi_e = 0 \qquad \ldots\ldots(18)$$

$$\nabla_t^2 \psi_h + k_c^2 \psi_h = 0 \qquad \ldots\ldots(19)$$

in which

$$k_c^2 = k_0^2 + \gamma^2$$
$$k_0^2 = \omega^2 \mu_0 \varepsilon_0 = 4\pi^2/\lambda_0^2 \qquad \ldots\ldots(20)$$

k_c is the cut-off wavenumber, i.e. it determines the wavelength at which free propagation ceases ($\gamma = 0$). This occurs when λ_0 is so chosen that $k_c^2 = k_0^2$ for a given mode and so the cut-off wavelength is $\lambda_c = 2\pi/k_c$.

In orthogonal coordinate systems, the solution of the above equations (18) and (19) is of the form

$$\psi = U_1(u_1)U_2(u_2) \qquad \ldots\ldots(21)$$

The wave function ψ should not only satisfy the scalar Helmholtz equation but should also satisfy proper boundary conditions. The eigenvalues of k_c^2 and hence of γ^2 are determined by the boundary conditions of ψ. It is evident that as there are an infinite number of eigenvalues for k_c^2, so an infinite number of modes exist. But only a finite number of values for γ are imaginary, and so only a finite number of modes propagate freely along the guide.

Several methods of solving the partial differential equations of the type (18) and (19) are possible, but the method of separation of variables is generally the most widely used. This method consists of the following procedure:

(i) The partial differential equation is transformed into the coordinate system that fits the geometry of the problem.

(ii) This equation is then separated into three ordinary differential equations.

(iii) The solutions of these ordinary differential equations are then obtained.

(iv) The unique solution that satisfies the boundary conditions is then constructed from the particular solutions obtained in (iii).

The range of problems that can be handled by separation of variables depends to a great extent on the number of available coordinate systems. The separation equations and their solutions for the wave equation for a large number of coordinate systems have been tabulated by Moon and Spencer.[91]

By choosing an appropriate kind of space and a cylindrical analogue, Waldron[92] has solved the wave equation by the method of separation of variables for the helical coordinate system which is non-orthogonal.

The field components for E and H waves in a loss-free medium are:
E waves:

$$E_1 = \frac{1}{h_1}\frac{\partial^2 \Pi_e}{\partial u_3 \partial u_1} \qquad H_1 = \pm\frac{j\omega\varepsilon}{h_2}\frac{\partial \Pi_e}{\partial u_2}$$
$$E_2 = \frac{1}{h_2}\frac{\partial^2 \Pi_e}{\partial u_3 \partial u_2} \qquad H_2 = \mp\frac{j\omega\varepsilon}{h_1}\frac{\partial \Pi_e}{\partial u_1} \quad \ldots\ldots(22)$$
$$E_3 = \frac{\partial^2 \Pi_e}{\partial u_3^2} + \frac{\omega^2}{c^2}\Pi_e \qquad H_3 = 0$$

H waves:

$$H_1 = \frac{1}{h_1}\frac{\partial^2 \Pi_m}{\partial u_3 \partial u_1} \qquad E_1 = \mp\frac{j\omega\mu}{h_2}\frac{\partial \Pi_m}{\partial u_2}$$
$$H_2 = \frac{1}{h_2}\frac{\partial^2 \Pi_m}{\partial u_3 \partial u_2} \qquad E_2 = \pm\frac{j\omega\mu}{h_1}\frac{\partial \Pi_m}{\partial u_1} \quad \ldots\ldots(23)$$
$$H_3 = \frac{\partial^2 \Pi_m}{\partial u_3^2} + \frac{\omega^2}{c^2}\Pi_m \qquad E_3 = 0$$

The time factor exp ($\pm j\omega t$) has been omitted for convenience. The coordinate relations and the metric coefficients (h_1, h_2, h_3) in cartesian (x, y, z) and cylindrical (r, θ, z) systems are respectively

$$\begin{array}{lll} u_1 = x & & u_1 = r \quad h_1 = 1 \\ u_2 = y & h_1 = h_2 = h_3; & u_2 = \theta \quad h_2 = r \\ u_3 = z & & u_3 = z \quad h_3 = 1 \end{array} \quad \ldots\ldots(24)$$

In these two cases h_1/h_2 is a function of u_1, u_2 only and $h_3 = 1$. By using these transformations, the field components for the E and H waves in the cartesian and cylindrical coordinates can be easily written.[93] The classification of waveguide modes into H and E with respect to z (the direction of propagation) is useful as it applies also to guides of non-rectangular cross-section and provides a convenient basis to which the modes of any general system may be reduced. However, for many rectangular waveguide problems, such as dielectric slab-loaded rectangular guide, a further classification of modes is more convenient. The basic modes of propagation in such a inhomogeneous guide can also be derived from the two Hertzian vectors $\mathbf{\Pi}_e$ and $\mathbf{\Pi}_m$ having single component normal to the interface of the two media. From the vector $\mathbf{\Pi}_m$ one can obtain a mode which is characterized by the absence of any electric field component normal to the interface of the two media and possesses an electric field component which lies in the longitudinal interface plane. This mode is called a longitudinal section electric (LSE) mode. Similarly, from the electric vector $\mathbf{\Pi}_e$, a mode is obtained which does not have any magnetic field component normal to the interface. This mode is referred to as a longitudinal section magnetic (LSM) mode. The LSE and LSM modes form a complete set in which an arbitrary field distribution may be expanded.[94] For a general inhomogeneously-filled cylindrical guide, the division into LSE and LSM modes is not possible.

The field components for the LSE and LSM modes transverse to x in terms of the function ψ_h and ψ_e are respectively as follows[95]

$$E_x = 0 \qquad H_x = \frac{1}{j\omega\mu}\left(\frac{\partial^2}{\partial x^2} + k^2\right)\psi_h$$
$$E_y = -\frac{\partial \psi_h}{\partial z} \qquad H_y = \frac{1}{j\omega\mu}\frac{\partial^2 \psi_h}{\partial x\,\partial y} \quad \ldots\ldots(25)$$
$$E_z = \frac{\partial \psi_h}{\partial y} \qquad H_z = \frac{1}{j\omega\mu}\frac{\partial^2 \psi_h}{\partial x\,\partial z}$$

and

$$E_x = \frac{1}{j\omega\varepsilon}\left(\frac{\partial^2}{\partial x^2} + k^2\right)\psi_e \qquad H_x = 0$$
$$E_y = \frac{1}{j\omega\varepsilon}\frac{\partial^2 \psi_e}{\partial x\,\partial y} \qquad H_y = \frac{\partial \psi_e}{\partial z} \quad \ldots\ldots(26)$$
$$E_z = \frac{1}{j\omega\varepsilon}\frac{\partial^2 \psi_e}{\partial x\,\partial z} \qquad H_z = -\frac{\partial \psi_e}{\partial y}$$

where $k^2 = k_x^2 + k_y^2 + k_z^2$ is the separation constant. Similarly, the field components for the modes transverse to y can be written. The components will be given by equations similar to (25) and (26) with x, y, z properly interchanged. For $(\text{LSE})_{mn}$ and $(\text{LSM})_{mn}$ modes, the wave functions $(\psi_h)_{mn}$ and $(\psi_e)_{mn}$ are respectively

$$(\psi_h)_{mn} = \sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b}\exp(-\gamma z) \quad \ldots\ldots(27\text{a})$$

$$(\psi_e)_{mn} = \cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b}\exp(-\gamma z) \quad \ldots\ldots(27\text{b})$$

where $m = 0, 1, 2, 3, \ldots$ and $n = 0, 1, 2, 3, \ldots$
It is to be noted that the waves $(\text{LSM})_{0n}$ and $(\text{LSE})_{m0}$ correspond to H_{0n} and E_{m0} waves of a homogeneous guide. All the other modes given by equation (27) are linear combinations of the degenerate sets of H and E modes.

In the case of a circular cylindrical guide filled with two coaxial dielectrics, it can be shown[25] that a pure E or H wave can be supported provided the wave is circularly symmetric. For the propagation of higher order modes, it is necessary that both E and H waves co-exist in order that the boundary conditions may be satisfied. This means that both the longitudinal components E_z and H_z should be present. The resultant wave is designated respectively by $(\text{HE})_{mn}$ or $(\text{EH})_{mn}$ according as H wave predominates over E or vice versa. The field components for these modes are found by combining linearly the field components for the two modes. The components for the $(\text{HE})_{mn}$ mode in circular cylindrical coordinates are given as follows

$$E_r = -\left[j\gamma k\{AJ'_m(kr)+BY'_m(kr)\} + j\omega\mu \frac{m}{r} \{A'J_m(kr)+B'Y_m(kr)\}\right]^{\sin}_{\cos} m\theta \exp(-\gamma z)$$

$$E_\theta = -\left[\gamma \frac{m}{r} \{AJ_m(kr)+BY_m(kr)\} - j\omega\mu k\{A'J'_m(kr)+B'Y'_m(kr)\}\right]^{\sin}_{\cos} m\theta \exp(-\gamma z)$$

$$E_z = k^2\left[AJ_m(kr)+BY_m(kr)\right]^{\sin}_{\cos} m\theta \exp(-\gamma z)$$

$$H_r = -\left[\omega\varepsilon \frac{m}{r} \{AJ_m(kr)+BY_m(kr)\} + \gamma k\{A'J'_m(kr)+B'Y'_m(kr)\}\right]^{\sin}_{\cos} m\theta \exp(-\gamma z)$$

$$H_\theta = -\left[j\omega\varepsilon\{AJ'_m(kr)+BY'_m(kr)\} + \gamma \frac{m}{r} \{A'J_m(kr)+B'Y_m(kr)\}\right]^{\sin}_{\cos} m\theta \exp(-\gamma z)$$

$$H_z = \left(k^2 - \frac{2m^2}{r^2}\right)\left[A'J_m(kr)+B'Y_m(kr)\right]^{\sin}_{\cos} m\theta \exp(-\gamma z) \qquad \ldots\ldots(28)$$

where J_m, Y_m represent Bessel functions of the first and second kind and order m respectively. J'_m, Y'_m represent the derivatives of J_m and Y_m with respect to their arguments.

4. Orthogonality Properties of Fields in a Waveguide

The orthogonality properties of the field components are direct consequences of the orthogonality properties of the eigenfunctions. Any two functions f(x) and g(x) for which the inner product† (f, g) vanishes are said to be orthogonal. The norm of a function denoted by Nf is the inner product of the function f(x) with itself

$$\text{Nf} = \int f^2\,dx \qquad \ldots\ldots(29)$$

When Nf = 1, the function f(x) is said to be normalized. If any two members of a system of normalized functions $\psi_1(x)$, $\psi_2(x) \ldots \psi_n(x)$ are orthogonal, the system is said to be an orthonormal system. The orthogonality relations are expressed as follows

$$(\psi_m, \psi_n) = \delta_{mn} \qquad \ldots\ldots(30)$$

where

$$\delta_{mn} = \begin{matrix} 1 & \text{or } m = n \\ 0 & \text{for } m \neq n \end{matrix}$$

For functions of a real variable, which can have complex values, the concept of orthogonality can be applied in the following way. Two complex functions f(x) and g(x) are said to be orthogonal, if the following relations are satisfied

$$(\text{f}, \text{g}^*) = (\text{f}^*, \text{g}) = 0 \qquad \ldots\ldots(31)$$

where f* and g* indicate complex conjugates of f and g respectively. The function f(x) is said to be normalized if

$$\text{Nf} = \int |\text{f}|^2\,dx = 1 \qquad \ldots\ldots(32)$$

It may be useful to mention that the functions of an orthogonal system are always linearly independent.

Maxwell's equations within a waveguide having walls of infinite conductivity have an infinite number of possible solutions which represent various normal modes such as E and H. The E and H-modes have several useful orthogonality properties. The orthogonality properties permit a given arbitrary field represented by a general solution to be expanded as a sum of solutions for the various normal modes. In order to determine the general field which can exist in the waveguide oscillating sinusoidally with a given angular frequency ω, it is required to find the most general solution of Maxwell's equations subject to proper boundary conditions. Suppose that both E and H-modes are being propagated inside a waveguide in the positive z direction. The resultant electric **E** and magnetic H vectors are then represented by superposing vectorially all the electric and magnetic vectors, respectively, of all the modes of both the E and H type as follows

$$\mathbf{E} = \sum_p A_p(\mathbf{E}'_{tp} + \mathbf{i}_z E'_{zp}) + \sum_p B_p \mathbf{E}''_{tp} \qquad \ldots\ldots(33)$$

$$\mathbf{H} = \sum_p B_p(\mathbf{H}''_{tp} + \mathbf{i}_z H''_{zp}) + \sum_p A_p \mathbf{H}'_{tp} \qquad \ldots\ldots(34)$$

† The integral (f, g) = $\int$f(x) g(x) dx taken over the finite domain is called the inner product of the two functions f and g. The limits are omitted for convenience. It satisfies the Schwarz inequality (f, g)$^2 \leq$ (f, f) (g, g), where the equality holds if f and g are proportional.

where A_p and B_p are expansion coefficients and E′, H′ are derived from the electric Hertz vector Π_e and E″, H″ are derived from the magnetic Hertz vector Π_m. The subscript t indicates that the vectors are transverse to the direction z of propagation and $p = 0, 1, 2. \ldots$ In order to determine, A_p and B_p, it is convenient to normalize the functions E_{zn}, H_{zn}, $\mathbf{E}_{tn}$, $\mathbf{H}_{tn}$, or in other words, it is required to determine the magnitudes of these functions, so that their squares can be integrated to some convenient value. The normalized function is multiplied by an additional coefficient so that an arbitrary value of the function is obtained. The normalization procedure yields the following relations for the expansion coefficients

$$A_p = \frac{\int_s \mathbf{E} \cdot \mathbf{E}'^{*}_{tp}\, ds}{\int_s \mathbf{E}'_{tp} \cdot \mathbf{E}'^{*}_{tp}\, ds} \qquad \ldots\ldots(35)$$

$$B_p = \frac{\int_s \mathbf{E} \cdot \mathbf{E}''^{*}_{tp}\, ds}{\int_s \mathbf{E}''_{tp} \cdot \mathbf{E}''^{*}_{tp}\, ds}$$

If E and H modes are propagated in both the $+z$ and $-z$ directions, the expansion coefficients and hence the field can be determined from a knowledge of $\mathbf{E}_t$ or $\mathbf{H}_t$ over two independent cross-sections of the guide. It may be remarked that the property of orthogonality is important and useful but the property of normalization is only of formal utility.

If in a waveguide bounded by infinitely conducting walls, ψ_i and ψ_j represent the solutions for the ith and jth E or H modes, the following orthogonality relations hold.

(i) $$\iint_s \psi_i \psi_j\, da = 0 \qquad i \neq j \qquad \ldots\ldots(36)$$

which states that the axial components of the field for the two different modes (i and j) are orthogonal, i.e. the integral over the cross-section S of the guide of the product E_{zn} E_{zm} or H_{zn} H_{zm} vanishes when $n \neq m$.

(ii) For two different E or H modes, the transverse electric fields are also orthogonal, that is

$$\iint_s \nabla_t \psi_i \cdot \nabla_t \psi_j\, da = 0 \qquad \ldots\ldots(37)$$

or in other words, the integral over S of the scalar product $\mathbf{E}_{tn} \cdot \mathbf{E}_{tm}$ or $\mathbf{H}_{tn} \cdot \mathbf{H}_{tm}$ vanishes when $n \neq m$. ψ_i and ψ_j are two different eigenfunctions from which the E or H modes may be derived. The orthogonality relations hold also for the transverse magnetic fields for any two E or H modes.

(iii) The transverse electric field for one E mode and one H mode also obey the orthogonality relation

$$\iint_s (\mathbf{i}_z \times \nabla_t \psi_{hi}) \cdot \nabla_t \psi_{ej}\, da = 0 \qquad \ldots\ldots(38)$$

i.e. the integral over the cross-section S of the guide of $\mathbf{i}_z \cdot (\mathbf{E}_{tn} \times \mathbf{H}_{tm})$ vanishes when $n \neq m$. ψ_{hi} and ψ_{ej} represent the two eigenfunctions giving rise to H and E waves respectively. $\mathbf{E}_{tn}$ is the transverse electric field for the nth mode and $\mathbf{H}_{tm}$ is the transverse magnetic field for the mth mode. Similar orthogonality relation is also obtained for the transverse magnetic field of E and H waves.

In the case of a waveguide with walls of finite conductivity, the eigenfunctions ψ_e and ψ_h do not satisfy the conditions $\psi_e = 0$ and $\partial\psi_h/\partial n = 0$ respectively, on the guide boundary assumed in deriving the above orthogonality relations. The presence of finite conductivity results in a cross-coupling between the various E and H modes. However, the more general relation

$$\iint_s (\mathbf{E}_{tn} \times \mathbf{H}_{tm}) \cdot \mathbf{i}_z\, da = 0 \qquad \ldots\ldots(39)$$

holds when $n \neq m$. $\mathbf{E}_{tn}$ and $\mathbf{H}_{tm}$ represent the transverse electric field for the nth mode and transverse magnetic field for the mth mode respectively.

In the case of degeneracy ($n = m$), the above relation is not valid. It is however, possible to secure orthogonality in the following way. If ψ_1 and ψ_2 are two degenerate modes, such that

$$\iint \psi_i \psi_j\, da = P_{ij} \qquad \ldots\ldots(40)$$

where i, $j = 1$, 2 and $\psi'_1 = \psi_1$, $\psi'_2 = \psi_2 + \alpha\psi_1$ represent two subsets, where the constant α is determined so that $\iint_s \psi'_1 \psi'_2\, da = 0$, then the two modes are orthogonal, when $\alpha = -P_{12}/P_{11}$. It may be said, in general, that any subset of n degenerate modes may be converted into a new subset of n mutually orthogonal modes.

In the case of LSE and LSM modes, the orthogonality relations (36–38) are not valid. The relation (39) is valid, when $n \neq m$. When $n = m$, the relation (39) holds good if the two fields are one for LSE and one for LSM modes, since these fields are independent and non-degenerate solutions of Maxwell's field equations. In the case of a rectangular guide, inhomogeneously filled with dielectric, the orthogonality relation holds for the transverse electric or transverse magnetic fields for two different LSE modes. The orthogonality also holds for the longitudinal components. But the orthogonality does not hold for the LSM modes. The orthogonality of waveguide fields has been treated in a very comprehensive manner by Bresler and Marcuvitz[96] and also by Bresler, Joshi and Marcuvitz.[96a]

5. Boundary Conditions

In the solution of Maxwell's field equations in a region of space divided into two or more parts forming the boundaries across which the electrical properties (ε and μ) change discontinuously, it is necessary to know the relationship between the field components on adjacent sides of the boundaries. This is required in order properly to continue the solution from one region to the next such that the final solution will be a solution of Maxwell's equations. These boundary conditions are readily derived by using the integral form of the field equations.

At a smooth boundary separating a region 1 from a region 2 characterized by the constitutive parameters $\varepsilon_1, \mu_1, \sigma_1$ and $\varepsilon_2, \mu_2, \sigma_2$ respectively, the field vectors **E**, **H**, **B**, **D** behave so as to satisfy Maxwell's equations. The transition of an electromagnetic field at the interface of two media is determined, in general, by the following relations

$$\begin{aligned} \mathbf{n}\times(\mathbf{E}_2-\mathbf{E}_1) &= 0 \qquad \mathbf{n}.(\mathbf{D}_2-\mathbf{D}_1) = \eta \\ \mathbf{n}\times(\mathbf{H}_2-\mathbf{H}_1) &= \mathbf{K} \qquad \mathbf{n}.(\mathbf{B}_2-\mathbf{B}_1) = 0 \end{aligned} \qquad \ldots\ldots(41)$$

where **n** represents the unit vector normal to the boundary pointing from region 1 to region 2. η represents the surface charge density and **K** represents the surface current density and is zero unless the conductivity (σ) of the medium at the boundary is infinite. The flow of charge across or to the boundary must also satisfy the following equation of continuity in case either or both the conductivities are finite and not zero

$$\mathbf{n}.(\mathbf{J}_2-\mathbf{J}_1) = -\frac{\partial\eta}{\partial t} \qquad \ldots\ldots(42)$$

From the above conditions, it follows that the condition for the transition of the normal components of **E** and **H** and the tangential components of **D** and **B** must satisfy the following relations:

$$\begin{aligned} \mathbf{n}.\left(\mathbf{H}_2-\frac{\mu_1}{\mu_2}\mathbf{H}_1\right) &= 0 \qquad \mathbf{n}.\left(\mathbf{E}_2-\frac{\varepsilon_1}{\varepsilon_2}\mathbf{E}_1\right) = \frac{\eta}{\varepsilon_2} \\ \mathbf{n}\times\left(\mathbf{D}_2-\frac{\varepsilon_2}{\varepsilon_1}\mathbf{D}_1\right) &= 0 \qquad \mathbf{n}\times\left(\mathbf{B}_2-\frac{\mu_2}{\mu_1}\mathbf{B}_1\right) = \mu_2\mathbf{K} \end{aligned} \qquad \ldots\ldots(43)$$

The above relations are general and are to be used by inserting appropriate values of relative permittivity ε_r and relative permeability μ_r for the media concerned. The media may be classified as dielectric, metal or perfect conductors, ferrite and plasmas. The nature of ε_r and μ_r for the different media is summarized as follows:[81a]

Dielectric: μ_r is scalar and is equal to unity. ε_r is scalar and complex, $\varepsilon_r = \varepsilon' - j\varepsilon''$.

Metal: ε'' is very large, ε' is small and $\mu_r = 1$.

Perfect conductor: $\varepsilon'' = \infty$, $\varepsilon' = 0$, $\mu_r = 0$.

Ferrite: ε_r is scalar and complex. μ_r is a tensor whose elements are in general complex.

Plasmas: $\mu_r = 1$, ε_r is a tensor whose elements are in general complex.

If one of the regions (say 1) is occupied by a perfect conductor ($\sigma = \infty$, $\mu_r = 0$), then $\mathbf{E}_2 = 0$, the component of the magnetic field in medium 1 normal to the surface of the perfect conductor, vanishes. But the component of the magnetic field ($\mathbf{H}_2$) in medium 2 normal to the surface does not necessarily vanish. Hence, the boundary conditions at the interface of the above two media reduce to

$$\begin{aligned} \mathbf{n}\times\mathbf{E}_1 &= 0 \qquad \mathbf{n}\times\mathbf{H}_1 = -\mathbf{K} \\ \mathbf{n}.\mathbf{B}_1 &= 0 \qquad \mathbf{n}.\mathbf{D}_1 = -\eta \end{aligned} \qquad \ldots\ldots(44)$$

At a discontinuity surface it is sufficient to match the tangential field components only, since if the fields satisfy Maxwell's equations this automatically makes the normal components of the flux vectors satisfy the correct boundary conditions.

6. Methods of Approach

The problem of the propagation characteristics of microwaves through waveguides containing dielectrics may be studied by either field theory or network theory.

6.1. *Field Theory*

The field components for any mode for the different media, found from Maxwell's equations, involve constants which are evaluated in terms of the peak power flow through the guide. The peak power flow is found with the help of the Poynting vector involving the tangential field components. By applying proper boundary conditions at the interface between any two media inside the guide, the characteristic equation is formed. The equation which is generally transcendental is solved graphically or by trial and error methods with digital computers to determine the propagation constant. This approach has been adopted by the author in solving problems on composite cylindrical guide.[24–26, 29, 30, 33, 82]

6.2. *Network Theory: Transmission Line*

The composite structure is described in terms of a transmission region characterized by a single dominant mode and a discontinuity region characterized by an infinite set of higher order rapidly evanescent modes. The field problem is then reformulated in terms of conventional network concepts. If the waveguide cross-section possesses an appropriate geometrical symmetry, the characteristics of the propagating mode are determined by regarding not the longitudinal guide axis but rather one of the transverse directions perpendicular thereto as a transmission direction and setting up a transverse equivalent network representation of the desired mode. This method has been followed by Marcuvitz[97] and Montgomery.[98] An alternative approach is to reformulate the problem in terms of a scattering matrix. The reformulation of field problems either as network problems or as scattering problems provides a fully equivalent and equally rigorous description of the far field in a microwave structure. It may however be remarked that this method gives merely the propagation constant. But a knowledge of the behaviour of fields requires that a suitable solution of Maxwell's equations be subjected to the proper boundary conditions at the walls of the guide and at the interface separating any two media.

6.3. *Approximate Methods*

An exact solution of the equations encountered in problems on inhomogeneous guides may be obtained for only a limited number of cases. For instance, in the case of the scalar Helmholtz equation ($\nabla^2\psi + k^2\psi = 0$) the method of separation of variables can be used for some coordinate systems. But if the surface on which the boundary conditions are to be satisfied is not one of these coordinate surfaces or if the boundary conditions are not the simple Drichlet ($\psi = 0$) or Neumann $\left(\frac{\partial\psi}{\partial n} = 0\right)$ types but of the mixed type $\frac{\partial\psi}{\partial n} + f(s)\psi = 0$, where f is in general, a function of the surface coordinates and f varies over the boundary, the separation of variable method fails. This has led to the theoretical techniques such as the integral equation method, the equivalent static and transform methods devised by Schwinger,[99] to determine the circuit parameters and obtain approximation to the eigenvalues. In the case of integral equation formulation, the transform technique may be used to obtain a solution only if the kernel of the integral equation is of a particular form. For instance, the Fourier transform may be used if the kernel is a function of the difference of two variables and if the range of integration is from $-\infty$ to $+\infty$, or from 0 to ∞. In such cases, it is more convenient to use approximate methods for the solution. The different methods that have been discussed in the paper for obtaining approximate solutions are (i) the perturbation method, (ii) the variational method, (iii) the Rayleigh-Ritz procedure, (iv) the reaction concept, and (v) the finite-difference equation method.

Perturbation methods are useful, whenever the problem under consideration closely resembles one which can be solved exactly. It is the theory of the changes which occur in the eigenvalues and eigenfunctions in an eigenvalue problem when only small changes are made in the conditions of the problem such as surface or volume perturbation. The problems may involve the determination of the resonant frequencies of microwave cavities of general shape in which small dielectric or magnetic samples are introduced, or the determination of the scattered and diffracted field when an obstacle is placed in the path of electromagnetic waves. When the deviations from the exactly soluble problem become large, the perturbation method becomes inconvenient. In this case it is more fruitful to use the variational method. The variational procedure gives an approximation to the desired quantity itself, rather than to changes in the quantity. In the variational procedure, it becomes necessary to find a quantity involving the unknown function which is to be stationary upon variation of the function. If the desired quantity is real, the variational relation may be an upper or lower bound to the desired quantity. In order to determine the accuracy of a variational calculation, it is necessary to have a systematic procedure for improving on the original trial function. The method of inserting additional nonlinear parameters will increase the accuracy, but it cannot be said that it will ultimately lead to the correct answer. One way to avoid this difficulty is to insert a linear combination of a complete set of functions, the coefficients forming a set of linear variational parameters. For instance, if an assumed field is expressed as a series of functions with undetermined coefficients, then the coefficients can be adjusted by the Rayleigh–Ritz procedure. If a complete set of functions is used for the assumed field, it is sometimes possible to obtain an exact solution. The Rayleigh–Ritz procedure can be utilized to obtain approximation for the first N-eigenvalues and N-eigenfunctions in an eigenvalue problem. Stationary formulas for electromagnetic problems can also be obtained by using the concept of 'reaction'. The propagation characteristics of guided electromagnetic waves through isotropic media are obtained from the solution of one or more two-dimensional eigenvalue wave equations

$$(\nabla_t^2 + k^2)\psi = 0$$

subject to certain prescribed boundary conditions. The total electromagnetic field is then defined by ψ, the eigenvalue k, the spatial derivatives of ψ and the applied frequency. When the boundary conditions are not simple, or in case of coordinate systems where the analytical separation of variables technique fails, the approximate variational and perturbation techniques are applied but it tends only to give the eigenvalues of the problem and little information on the values of ψ throughout the medium. It has been shown[43] that finite-difference techniques in conjunction with matrix calculations on digital computers can be used more conveniently in such cases.

6.3.1. The perturbation method

Many physical problems reduce to finding the eigenvalues k_n and eigenfunctions $\psi_n = \psi_n(x)$ of an equation of the following form

$$L\psi_n + (k_n - p)\psi_n = 0 \qquad \text{......(45)}$$

where L is a differential operator and $p = p(x)$, where x represents a set of coordinates. It is often necessary to find solutions under circumstances slightly different from conditions for which the solutions are known. In some problems, the boundary of the region is deformed slightly and the boundary conditions are assumed to be unaltered, whereas, in some cases, the boundary remains unchanged, but the boundary conditions for ψ are slightly altered from the case when the solutions are known. The problem can be approached in several ways.[100] The problem can be solved by introducing additional terms to the differential operator and then employing the usual perturbation technique. The problem can also be approached by transforming the differential equation to an integral equation by using Green's function.[101] The integral equation which includes the boundary conditions is then reduced to a secular equation, the solution of which is expressed in a form suitable for obtaining its value to any approximation.[102] The problem may also be solved by introducing perturbation terms in ψ and k and then solving the differential equation. A brief discussion of this method when k has discrete values and when the boundary is slightly deformed but the boundary conditions are assumed to be unaltered is given below.

In electromagnetic problems of interest, discussed in this paper, the differential operator is a Laplacian and the differential equation assumes the form

$$\nabla^2\psi_n + (k_n - p)\psi_n = 0 \qquad \text{......(46)}$$

which is assumed to be valid over a certain region R. Let the function $p(x)$ be modified to

$$P(x) = p(x) + \varepsilon t(x)$$

due to slight deformation of the boundary, and as a consequence, change ψ_n and k_n respectively as follows:

$$\begin{aligned} \Psi_n(x) &= \psi_n(x) + \varepsilon\psi_n^{(1)}(x) + \varepsilon^2\psi_n^{(2)}(x) + \dots \\ K_n(x) &= k_n + \varepsilon k_n^{(1)} + \varepsilon^2 k_n^{(2)} + \dots \end{aligned} \qquad \text{......(47)}$$

where the perturbed wave functions Ψ_n and eigenvalues K_n are assumed to be analytic for all possible values of the parameter ε which takes only small positive values including zero. Equation (46) is then changed to

$$\nabla^2\Psi_n(x) + [K_n - p(x) - \varepsilon t(x)]\Psi_n(x) = 0 \qquad \text{......(48)}$$

Substituting equation (47) in the perturbed equation (48) and equating the coefficients of ε, the following equations are obtained:

$$\begin{aligned} &\nabla^2\psi_n(x) + [k_n - p(x)]\psi_n(x) = 0 \\ &\nabla^2\psi_n^{(1)}(x) + [k_n - p(x)]\psi_n^{(1)} + (k_n^{(1)} - t(x))\psi_n(x) = 0 \\ &\nabla^2\psi_n^{(2)}(x) + [k_n - p(x)]\psi_n^{(2)} + \\ &\quad + (k_n^{(1)} - t(x))\psi_n^{(1)}(x) + k_n^{(2)}\psi_n(x) = 0 \end{aligned} \qquad \text{......(49)}$$

Using the last two equations and the expansions

$$\psi_n^{(1)}(x) = \sum_{q=0}^{\infty} \gamma_{n,q}\psi_q(x)$$

$$\psi_n^{(2)}(x) = \sum_{q=0}^{\infty} \delta_{n,q}\psi_q(x)$$

in terms of the unperturbed wave functions $\psi_q(x)$, the following equations are obtained:

$$\sum_{q=0}^{\infty} \gamma_{n,q}(k_n - k_q)\psi_q(x) + [k_n^{(1)} - t(x)]\psi_n(x) = 0$$

$$\begin{aligned} &\sum_{q=0}^{\infty} \delta_{n,q}(k_n - k_q)\psi_q(x) + \\ &+ [k_n^{(1)} - t(x)] \sum_{q=0}^{\infty} \gamma_{n,q}\psi_q(x) + k_n^{(2)}\psi_n(x) = 0 \end{aligned} \qquad \text{......(50)}$$

which lead to the following relations:

$$\gamma_{n,n} = 0$$

$$\gamma_{n,m} = \frac{\alpha_{mn}}{k_n - k_m} \qquad n \neq m$$

$$\delta_{n,n} = -\tfrac{1}{2} \sum_{q \neq n} \frac{\gamma_{n,q}^2}{(k_n - k_q)^2}$$

$$\delta_{n,m} = \frac{1}{k_n - k_m} \sum_{q \neq n} \frac{\alpha_{m,q}\alpha_{n,q}}{k_n - k_q} - \frac{\alpha_{n,n}\alpha_{m,n}}{(k_n - k_m)^2}$$

$$k_n^{(1)} = \alpha_{n,n}$$

$$k_n^{(2)} = \sum_{q \neq n} \frac{\alpha_{n,q}^2}{k_n - k_q}$$

where

$$\alpha_{n,m} = \int_R t(x)\psi_m(x)\psi_n(x)$$

Since $\delta_{n,q} = \delta_{n,n} + \delta_{n,m}$

and $\gamma_{n,q} = \gamma_{n,n} + \gamma_{n,m} = \gamma_{n,m}$,

$\psi_n^{(1)}$ and $\psi_n^{(2)}$ can be determined and hence $\Psi_n(x)$ corresponding to K_n can be found. This treatment holds good for a non-degenerate system. The perturbation theory both for the discrete as well as continuous spectrum has been treated in an illuminating manner by Titchmarsh.[103] The method has been applied successfully to solve some problems in microwave cavity resonators.[95, 104–107] The theory has also been used by Hayashi[108] in dealing with the electromagnetic fields in anisotropic inhomogeneous media.

6.3.2. Variational methods

Many of the problems which arise in electromagnetic wave propagation may be formulated as problems in variational calculus.[109–111] In differential calculus one is interested to find points at which functions of one or more variables possess maximum or minimum values, whereas, in the calculus of variations, we try to find the functional forms for which a given integral assumes maximum or minimum values. Or, in other words, the calculus of variations deals with the problem of finding the paths of integration for which a given integral assumes maximum or minimum values. A variation indicated by the symbol δ signifies an infinitesimal change by analogy with the differentiation process, but this infinitesimal change is imposed externally on a set of variables and is not due to the actual change of an independent variable. Consider a function $y = \mathrm{f}(x)$. Both δy and $\mathrm{d}y$ represent infinitesimal changes in y. But $\mathrm{d}y$ represents an infinitesimal change of the function $\mathrm{f}(x)$ caused by the infinitesimal change $\mathrm{d}x$ of the independent variable, whereas δy refers to an infinitesimal change of y which produces a new function $y+\delta y$. The independent variable x does not take part in the process of variation. The problem of finding the position of a point at which a function has a relative maximum or minimum requires that the function has a stationary value at the point. If the rate of change of the function in every possible direction from that point vanishes, then the function is said to have a stationary value. The variational method is applicable to single and to multiple integrals, and to cases where the integrand is a function of one or more independent variables and their first or possibly higher order derivatives, and to cases where the limits of integration are variable.

If L is a function of the independent variables $x_i (i = 1, 2, 3......n)$ and the functional $u = u(x_i)$, $v = v(x_i)$ and their first derivatives

$$\dot{u}_i\left(=\frac{\partial u_i}{\partial x_i}\right), \qquad \dot{v}_i\left(=\frac{\partial v_i}{\partial x_i}\right)$$

and if it is assumed that L, u, v are continuous and $\dot{u}_i$, $\dot{v}_i$ are also continuous or piecewise continuous, then the essence of the variational process is to determine u and v, such that the definite integral in the n-dimensional space

$$I = \int\int\int \dots \int_N L(x_i, u, v, \dot{u}_i, \dot{v}_i)\mathrm{d}x_i \qquad ......(51)$$

becomes stationary ($\delta I = 0$), when there are first-order variations δu, δv in u and v. The functions u and v satisfy the prescribed boundary conditions, so that the variations in u and v vanishes at the two extreme limits of integration. The necessary and sufficient condition that a function L of N variables shall be stationary at a certain point is that the N partial derivatives of L with respect to all the N variables shall vanish at the point in question. In order that the integral I may be stationary, the function L must satisfy independently the Euler-Lagrange partial differential equations as follows:

$$\begin{aligned} \frac{\partial L}{\partial u} - \sum_{i=1}^{N} \frac{\partial}{\partial x_i}\frac{\partial L}{\partial \dot{u}_i} &= 0 \\ \frac{\partial L}{\partial v} - \sum_{i=1}^{N} \frac{\partial}{\partial x_i}\frac{\partial L}{\partial \dot{v}_i} &= 0 \end{aligned} \qquad ......(52)$$

The variational operator δ obeys the following laws:

(i) $$\delta(L_1, L_2) = L_1\,\delta L_2 + L_2\,\delta L_1$$

$$\delta\left(\frac{L_1}{L_2}\right) = \frac{L_2\delta L_1 - L_1\delta L_2}{L_2^2}, \text{ etc.}$$

which are analogous to the corresponding laws of differentiation.

(ii) The variational operator δ and the differential $\mathrm{d}/\mathrm{d}x$ are commutative, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}x}(\delta u) = \delta\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)$$

(iii) The variation and integration processes are commutative, i.e.

$$\delta I = \delta \int_{x_1}^{x_2} L\,\mathrm{d}x = \int_{x_1}^{x_2} \delta L\,\mathrm{d}x$$

i.e. the variation of a definite integral is equal to the definite integral of the variation.

In some physical problems which need the determination of several functions by a variational method, the variations cannot all be arbitrarily assigned but are restricted by some auxiliary conditions. If the stationary condition is of the form

$$\delta \int_{x_1}^{x_2} L(x, u, v, \dot{u}, \dot{v})\,\mathrm{d}x = 0$$

where $u = u(x)$, $v = v(x)$ and if u and v are not independent but are related by the restrictions

$$\phi(u, v) = 0$$

then in the variational integral, the variation in u and v are governed such that $\phi = 0$ always. The integral of ϕ over the limits $x = x_1$ and $x = x_2$, also vanishes. Then any multiple $\lambda\phi$ or ϕ may be added to the function L under the integral sign without changing the value of the integral over the prescribed limits, where

λ is called the Lagrange's undetermined multiplier. This method reduces a variational problem with constraint to a variational problem without any constraint. In this method, the function L is modified by adding the left-hand side of the constraint equations after multiplying each by an undetermined multiplier λ. The modified problem is then treated as a variational problem without any auxiliary conditions. The resulting conditions, together with the prescribed constraints, determine the unknown and the undetermined multiplier λ.

The following equations:

$$\frac{d}{dx}\left(\frac{\partial L}{\partial \dot{u}}\right) - \frac{\partial L}{\partial u} - \lambda\frac{\partial \phi}{\partial u} = 0$$

$$\frac{d}{dx}\left(\frac{\partial L}{\partial \dot{v}}\right) - \frac{\partial L}{\partial v} - \lambda\frac{\partial \phi}{\partial v} = 0$$

with the auxiliary equation $\phi(u, v) = 0$, determine u, v and λ. For N-dimensional space, the conditional equations to find the functions u, v and λ are

$$\left.\begin{aligned} &\frac{\partial}{\partial u}(L+\lambda\Phi) - \sum_{i=1}^{N}\frac{\partial}{\partial x_i}\frac{\partial}{\partial \dot{u}_i}(L+\lambda\Phi) = 0 \\ &\frac{\partial}{\partial v}(L+\lambda\Phi) - \sum_{i=1}^{N}\frac{\partial}{\partial x_i}\frac{\partial}{\partial \dot{v}_i}(L+\lambda\Phi) = 0 \end{aligned}\right\} \quad \ldots\ldots(53)$$

with the auxiliary conditions

$$\Phi(x_i, u, v, \dot{u}_i, \dot{v}_i) = 0$$

In some cases, the auxiliary condition may be of the form

$$\int_{x_1}^{x_2} G(x, u, \dot{u})dx = k \text{ (a constant)}$$

If the variational integral is of the form

$$\delta I = \delta \int_{x_1}^{x_2} L(x, u, \dot{u})dx$$

and is stationary ($\delta I = 0$) and u and $u+\delta u$ satisfy the auxiliary condition and u satisfies the prescribed values at $x = x_1$ and $x = x_2$, then δu is not arbitrary. If

$$\delta u = \varepsilon f(x) + \alpha g(x)$$

involving the arbitrarily fixed function $g(x)$, where $f(x)$ and $g(x)$ are continuously differentiable functions, which vanish at the end points, and ε and α are constants, then the necessary condition that $\delta I = 0$ is

$$\left[\frac{\partial L}{\partial u} - \frac{d}{dx}\left(\frac{\partial L}{\partial \dot{u}}\right)\right] + \lambda\left[\frac{\partial G}{\partial u} - \frac{d}{dx}\left(\frac{\partial G}{\partial \dot{u}}\right)\right] = 0 \quad \ldots\ldots(54)$$

where the undetermined multiplier is given by the relation

$$\lambda = \frac{\int_{x_1}^{x_2}\left(\frac{\partial L}{\partial u} - \frac{d}{dx}\frac{\partial L}{\partial \dot{u}}\right)g\,dx}{\int_{x_1}^{x_2}\left(\frac{\partial G}{\partial u} - \frac{d}{dx}\frac{\partial G}{\partial \dot{u}}\right)g\,dx} \quad \ldots\ldots(55)$$

The condition (54) can be generalized in the case of two or more independent variables. As an example of the application of the variational calculus, let us determine the function u for which the integral

$$I = \int_0^1 L(x, u, \dot{u})dx$$

is stationary if $u = 0$ when $x = 0$ and $u = 2$ when $x = 1$. The function L is given as follows

$$L(x, u, \dot{u}) = \{3xu^2(\dot{u}^2-1)+u^3\dot{u}^3\}$$

The procedure is to find $\partial L/\partial \dot{u}$ and $\partial L/\partial u$ and substitute them in the Euler-Lagrange equation

$$\frac{d}{dx}\frac{\partial L}{\partial \dot{u}} - \frac{\partial L}{\partial u} = 0$$

which yields

$$u^2 + x^2 = 2\alpha x + \beta$$

where α and β are constants of integration. Applying the boundary conditions we obtain $\alpha = 5/2$, $\beta = 0$. Therefore the function u is given by the relation

$$u^2 + x^2 - 5x = 0$$

for which $\delta I = 0$. Variational methods have been extensively used to solve eigenvalue problems and the interested reader is referred to the books by Lanczos,[109] Gould,[110] Fox,[111] and Hildebrand.[112]

6.3.3. Sturm-Liouville problems

A certain type of boundary value problem involves a differential equation of the form

$$\frac{d}{dx}\left(p\frac{dy}{dx}\right) + (\lambda q - r)y = 0 \quad \ldots\ldots(56)$$

with some prescribed boundary conditions at the limits of the interval (a, b). In this equation, p, q, r are continuous functions of x in the interval (a, b) and λ is a parameter independent of x. By a suitable choice of p, q, r, most of the important differential equations governing electromagnetic wave propagation can be obtained from equation (56). For example the equation reduces to a simple equation of periodic motion when $p = q = 1$, $r = 0$. But when $p = x$, $q = x$, $r = -n^2/x$, (56) leads to the Bessel equation and when $p = 1-x^2$, $q = 1$, $r = 0$, it reduces to the Legendre equation. Mathieu's equation, Gauss's hypergeometric equation, Chebyshev's equation, and Hermite and Laguerre polynomials are also special cases of the above equation.

In order to have a non-trivial solution of equation (56) which will satisfy the boundary conditions at the end of the interval (a, b), λ must assume one of the eigenvalues λ_k. To each eigenvalue there corresponds an eigenfunction ϕ_k. The solution may be expressed in terms of the eigenfunction as $y = c\phi_k(x)$, where c is an arbitrary constant. The boundary conditions of the Sturm-Liouville system include the following

(i) At $x = a$ or $x = b$, either $y = 0$ or $\dot{y} = 0$ or a linear combination $\alpha y + \beta \dot{y} = 0$.

(ii) If $p(x) = 0$ at $x = a$ or $x = b$, we require y and $\dot{y}$ to remain finite at that point and impose one of the conditions (i) at the other point.

(iii) If $p(b) = p(a)$, then only $y(b) = y(a)$ and $\dot{y}(a) = \dot{y}(b)$ are required.

It can be shown [113] that any continuous function can be developed in terms of Sturm-Liouville functions which are the solutions of equation (56). The system is also characterized by the following properties:

(i) An infinite number of eigenfunctions ϕ_k's corresponding to definite eigenvalues of λ_k exists.

(ii) The eigenfunctions possess orthogonality with respect to the weighting function r over the prescribed interval (a, b), i.e.

$$(\lambda_j - \lambda_i) \int_a^b r\phi_i\phi_j \,\mathrm{d}x = 0 \qquad \lambda_i \neq \lambda_j$$

(iii) The eigenfunctions ϕ_k's form a complete set in terms of which a well behaved function $f(x)$ can be represented by a convergent series.

(iv) The solution to the problem may be formulated as the problem of finding the function which will make the variational integral stationary.

6.3.4. Rayleigh-Ritz method

The method consists of a procedure for determining approximately the eigenfunctions and eigenvalues to problems expressed in variational form. These are obtained by means of a variational integral, whose stationary values correspond to the true eigenvalues, when the true eigenfunctions are used in the integrand. The object of this method is to replace the problem of finding the maxima and minima of integrals by that of finding the maxima and minima of functions of several variables. This then reduces the problem to that of the calculus of functions.

Suppose we are required to determine the functional form of y so that the integral

$$I = \int_a^b F(x, y, \dot{y}, \ddot{y} \ldots) \,\mathrm{d}x$$

becomes stationary. The procedure is to assume that the desired extremal of the problem can be expressed approximately as

$$y \simeq \sum_{i=1}^{n} c_i \mathrm{f}_i(x)$$

where $\mathrm{f}(x)$ are arbitrarily chosen, such that the expression satisfies the specified boundary condition for any choice of C_i and C_i are undetermined. Substituting for y and evaluating the integral, we obtain an expression in terms of C_i. Using the method of differential calculus, I is stationary if C_i are such that $\partial I/\partial C_i = 0$. These n equations can then be solved for finding the n parameters C_1, C_2, C_n and hence y. In most of the cases, an approximate result is obtained. The accuracy of the result depends on the choice of the function $\mathrm{f}_i(x)$. A closer approximation can be obtained by increasing the value of n. If a large number of $\mathrm{f}_i(x)$ are used to obtain close approximation, it is desirable to choose a sequence of functions which are complete. A function $\mathrm{f}_i(x)$ is said to be complete if for any piecewise continuous function $F(x)$, a set of coefficients C_i can be chosen such that the following relation is satisfied

$$\lim_{n\to\infty} \int_a^b \{F(x) - \sum_{i=1}^{n} C_i \mathrm{f}_i(x)\}^2 \mathrm{d}x = 0$$

Frequently, it is convenient to choose as the nth approximation a polynomial of degree i satisfying the prescribed boundary conditions. In certain cases, it is of advantage in numerical calculation to choose sine, cosine harmonics, Legendre polynomial and so on. Use of the Rayleigh-Ritz procedure avoids evaluation of many complicated expressions, especially for cases when a waveguide cross-section is divided into more than two regions by dielectric media of different dielectric constants. As an example, let us determine the eigenvalue λ in the boundary value problem as follows:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[(1+x)\dot{y}\right] + \lambda y = 0$$

with the boundary conditions $y(0) = 0$ and $y(1) = 0$. Let us assume

$$y(x) \simeq C_0 x(1-x)$$

If $\delta y = \varepsilon\alpha(x)$ represents a small variation vanishing at both end points, the variational problem is

$$\int_0^1 \left[\frac{\mathrm{d}}{\mathrm{d}x}\{(1+x)\dot{y}\} + \lambda y\right] \delta y \,\mathrm{d}x = 0$$

Substitituting the value of y and evaluating the integral, we obtain $\lambda = 15$. If $y = (C_0 + C_1 x)\, x(1-x)$, then substituting this value of y in the integral, and equating the coefficients of δC_0 and δC_1 after evaluating the integral, the following equations are obtained:

$$\left(-\frac{1}{2}+\frac{\lambda}{30}\right)C_0+\left(-\frac{17}{60}+\frac{\lambda}{60}\right)C_1 = 0$$

$$\left(-\frac{17}{60}+\frac{\lambda}{60}\right)C_0+\left(-\frac{7}{30}+\frac{\lambda}{105}\right)C_1 = 0$$

which lead to the equation

$$\lambda^2-78\lambda+917 = 0$$

which yields the two lowest eigenvalues as $\lambda_1 \simeq 14{\cdot}42$ and $\lambda_2 \simeq 63{\cdot}58$.

6.3.5. Reaction concept

Stationary formulas can also be established by using the concept of 'reaction' introduced by Rumsey.[114] The use of this concept considerably simplifies the formulation of boundary value problems in electromagnetic theory. If the volume distribution of electric $d\mathbf{J}(a)$ and magnetic $d\mathbf{M}(a)$ currents represents the source (simply called source) of a monochromatic electromagnetic field and the electric and magnetic fields generated by these source distributions are represented respectively by $\mathbf{E}(a)$ and $\mathbf{H}(a)$ and similarly, $\mathbf{E}(b)$ and $\mathbf{H}(b)$ represent the electric and magnetic fields due to another set of sources b of the same frequency and existing in the same linear medium, the 'reaction' (a physical observable) or a coupling between the sources a and b is defined by the relation

$$\langle a,b\rangle = \int\limits_{\substack{\text{volume}\\ \text{containing}\\ \text{the source}}} [\mathbf{E}(b).d\mathbf{J}(a)-\mathbf{H}(b).d\mathbf{M}(a)] \quad \ldots\ldots(57)$$

If there are any three sources a, b, and c radiating at the same frequency in the same region, then

$$\langle a,(b+c)\rangle = \langle a,b\rangle+\langle a,c\rangle \quad \ldots\ldots(58)$$

If the strength of the source a is multiplied by a scalar A to give Aa, then the following relation holds

$$\langle Aa,b\rangle = A\langle a,b\rangle \quad \ldots\ldots(59)$$

The best approximate relation $\langle a, b\rangle$ to a desired relation $\langle c_a, c_b\rangle$ is obtained if

$$\langle a,b\rangle = \langle c_a,b\rangle = \langle a,c_b\rangle \quad \ldots\ldots(60)$$

which means that all trial sources look the same to themselves as to the correct sources (c). If the small variation in a and b about c_a and c_b are represented by

$$\begin{aligned} a &= c_a+p_a e_a \\ b &= c_b+p_b e_b \end{aligned} \quad \ldots\ldots(61)$$

then $\langle a, b\rangle$ is stationary if

$$\left.\frac{\partial\langle a,b\rangle}{\partial p_a}\right|_{p_a=p_b=0} = \left.\frac{\partial\langle a,b\rangle}{\partial p_b}\right|_{p_a=p_b=0} = 0 \quad \ldots\ldots(62)$$

In the case of a resonant cavity, the true field at resonance is a source free field, so the reaction of any field with the correct source is zero. Hence, if $a = b$ represents a trial field and associated source then eqn. (60) reduces to

$$\langle a,a\rangle = 0 \quad \ldots\ldots(63)$$

6.3.6. Finite-difference method

The calculus of finite differences deals with the changes that occur in the values of a function, the dependent variable due to changes in the independent variable. Finite differences form the basis of numerical differentiation, integration and solution of differential equations. In this method, it is assumed that a given function $f(x)$ has a specified numerical value $f(x_r)$ at each of a sequence of equally spaced values $x = x_r$ of the independent variable x. If the common interval between any two values x_0, x_1, $x_2 \ldots$ is denoted by h, the first difference $\Delta f(x)$ of the function $f(x)$ is defined as

$$\Delta f(x) = f(x)+h)-f(x) \quad \ldots\ldots(64)$$

i.e. $\Delta f(x)$ gives the difference in values of the function for two neighbouring values of x, h units apart. We can regard the difference operator Δ as acting upon $f(x)$ in the same way as we regard the differentiation process $d/dx\ f(x)$ as being the operator d/dx acting on $f(x)$. The operator Δ like the operator d/dx is linear, i.e.

$$\Delta[a\,f(x)+b\,g(x)] = a\Delta f(x)+b\Delta g(x) \quad \ldots\ldots(65)$$

where a and b are constants. The Δ operation on the product of two functions $f(x)$ and $g(x)$ yields

$$\Delta[f(x)\,g(x)] = f(x+h)\Delta g(x)+g(x)\Delta f(x)$$

or,

$$\Delta[f(x)\,g(x)] = f(x)\Delta g(x)+g(x+h)\Delta f(x) \quad \ldots\ldots(66)$$

The Δ operation on the ratio of two functions gives

$$\Delta\frac{f(x)}{g(x)} = \frac{g(x)\Delta f(x)-f(x)\Delta g(x)}{g(x)\,g(x+h)} \quad \ldots\ldots(67)$$

The second difference $\Delta^2 f(x)$ is defined as the difference of the first difference of $f(x)$ for two neighbouring values of x, h units apart, i.e.

$$\Delta^2 f(x) = \Delta[\Delta f(x)] = \Delta f(x+h)-\Delta f(x) \quad \ldots\ldots(68)$$

Higher differences are defined similarly, and it is evident that any forward difference should be expressible in terms of the function values at various abscissae x_r. In general,

$$\Delta^n f(x) = \Delta[\Delta^{n-1} f(x)], \qquad n = 1,2,\ldots \quad \ldots\ldots(69)$$

When using higher differences, it is useful to set up a table as follows.

A special form of a diagonal difference table is one which is called a central difference table and is identical to the diagonal difference table, except that the values

Table 1

Diagonal difference table

x	$y(x)$	$\Delta y(x)$	$\Delta^2 y(x)$	$\Delta^3 y(x)$	$\Delta^4 y(x)$
0	$y(0)$				
		$\Delta y(0)$			
1	$y(1)$		$\Delta^2 y(0)$		
		$\Delta y(1)$		$\Delta^3 y(0)$	
2	$y(2)$		$\Delta^2 y(1)$		$\Delta^4 y(0)$
		$\Delta y(2)$		$\Delta^3 y(1)$	
3	$y(3)$		$\Delta^2 y(2)$		
		$\Delta y(3)$			
4	$y(4)$				

Table 2

Central difference table

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
-4	$y(-4)$				
		$\Delta y(-4)$			
-3	$y(-3)$		$\Delta^2 y(-4)$		
		$\Delta y(-3)$		$\Delta^3 y(-4)$	
-2	$y(-2)$		$\Delta^2 y(-3)$		$\Delta^4 y(-4)$
		$\Delta y(-2)$		$\Delta^3 y(-3)$	
-1	$y(-1)$		$\Delta^2 y(-2)$		$\Delta^4 y(-3)$
		$\Delta y(-1)$		$\Delta^3 y(-2)$	
0	$y(0)$		$\Delta^2 y(-1)$		$\Delta^4 y(-2)$
		$\Delta y(0)$		$\Delta^3 y(-1)$	
1	$y(1)$		$\Delta^2 y(0)$		$\Delta^4 y(-1)$
		$\Delta y(1)$		$\Delta^3 y(0)$	
2	$y(2)$		$\Delta^2 y(1)$		$\Delta^4 y(0)$
		$\Delta y(2)$		$\Delta^3 y(1)$	
3	$y(3)$		$\Delta^2 y(2)$		
		$\Delta y(3)$			
4	$y(4)$				

of the function are arranged around a central value $y(0)$. The values of the independent variable are numbered as ... $x(-3)$, $x(-2)$, $x(-1)$, $x(0)$, $x(1)$, $x(2)$, $x(3)$ ... and the corresponding values of the function are denoted as ... $y(-3)$, $y(-2)$, $y(-1)$, $y(0)$, $y(1)$, $y(2)$, $y(3)$ ....

Problems on inhomogeneous guides need the solution of partial differential equations. In the finite difference method, the partial differential equation is first converted to a difference equation which is then solved for finding the eigenvalues. A difference equation is one or more of the differences of the dependent variable. The difference quotient is defined as

$$\frac{f(x+h)-f(x)}{h}$$

which in the limit, if $h \to 0$, is the derivative of a simple function. So, if all the derivatives in a differential equation are replaced by the corresponding difference quotients, we obtain a difference equation. The numerical solution of difference equation is an extensive subject. We shall concern ourselves with partial difference quotients of the second and higher orders. Let $u(x, y)$ be a function of two variables. Let the x, y plane be divided into a network by two families of parallel lines (Fig. 1)

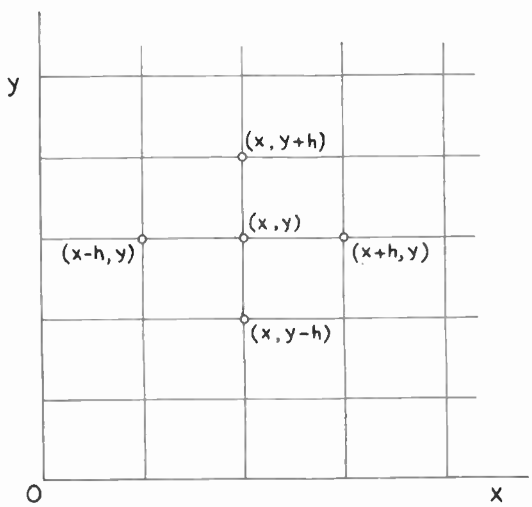


Fig. 1

$$x = \mu h \qquad \mu, \nu = 0, 1, 2, \ldots$$
$$y = \nu h$$

The points of intersection of these lines are called lattice points or nodes. For each of the variable of the function $u(x, y)$, we have a forward and a backward difference quotient. Thus with respect to x and y, the forward (u_x, u_y) and backward ($u_{\bar{x}}$, $u_{\bar{y}}$) quotients are respectively

$$u_x = \frac{u(x+h, y)-u(x, y)}{h}$$
$$u_{\bar{x}} = \frac{u(x, y)-u(x-h, y)}{h} \qquad \ldots\ldots(70)$$
$$u_y = \frac{u(x, y+h)-u(x, y)}{h}$$
$$u_{\bar{y}} = \frac{u(x, y)-u(x, y-h)}{h}$$

The second difference quotients of $u(x, y)$ with respect to x and y can be defined as the difference quotient of the first difference quotients, i.e.

$$u_{\bar{x}, x} = \frac{u_x - u_{\bar{x}}}{h} = \frac{u(x+h, y)-2u(x, y)+u(x-h, y)}{h^2}$$
$$\ldots\ldots(71)$$
$$u_{\bar{y}, y} = \frac{u_y - u_{\bar{y}}}{h} = \frac{u(x, y+h)-2u(x, y)+u(x, y-h)}{h^2}$$

These definitions are used to replace the partial derivatives in a partial differential equation and then the resulting difference equation is obtained. The functions occurring in the difference equation are defined at the nodes only. So to get a better description of the function, it is necessary to increase the number of nodal points. As an illustration, let us consider how Laplace's equation can be transformed to an equivalent difference equation.[115] This difference equation is then rearranged to obtain an expression for a particular value of the function in terms of the values of the function at adjacent points and other constants or known functional values.

Laplace's equation in two dimensions is

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0$$

$$u_{\bar{x}x} \simeq \frac{\partial^2 v}{\partial x^2}, \qquad u_{\bar{y}y} \simeq \frac{\partial^2 v}{\partial y^2}$$

Laplace's equation is then transformed with the help of the relations (71) to the following difference equation:

$$h^{-2}[u(x+h, y) - 2u(x, y) + u(x-h, y)] + \\ + h^{-2}[u(x, y+h) - 2u(x, y) + u(x, y-h)] = 0$$

which yields

$$u(x, y) = \tfrac{1}{4}[u(x+h, y) + u(x, y+h) + \\ + u(x-h, y) + u(x, y-h)]$$

i.e. the value of $u(x, y)$ at any interior nodal point is the arithmetic mean of the values of $u(x, y)$ at the four lattice points surrounding the point in question.

The numerical solution of partial differential equations is an important subject and the literature on this subject is growing rapidly. Almost every problem arising out of the physical sciences requires original thinking and modifications of the existing methods. With minor exceptions, the numerical solution of difference equations corresponding to partial differential equations is so laborious that it is carried out only with the help of automatic digital computers.[115–117]

[*To be continued*]

(Paper No. 1000)

Editorial Note:

Part II of this paper will be published in the October and November issues, and Part III, with the References, will be published in the December issue.

U.D.C. 519.27 : 621.039 : 681.84

A Correlator for Investigating Random Fluctuations in Nuclear Power Reactors

By

F. D. BOARDMAN, M.A.,*

E. L. E. HARRINGTON, B.Sc.
(*Associate Member*) †

AND

D. J. A. CARSWELL, M.A.‡

Summary: In nominally steady operation of a nuclear power reactor small fluctuations are observed in various parameters, e.g. reactor power, coolant flow and temperature. The cross-correlation between simultaneous recordings of these fluctuations can give information on the kinetics of the reactor system without disturbing its normal operation. An apparatus has been built to explore the practical possibilities of the technique and this paper describes its design principles, practical implementation and the results of tests. Some typical results obtained on the Dounreay Fast Reactor are also given.

The equipment records the fluctuations in four channels on magnetic tape, the correlation functions being obtained subsequently by analogue computing techniques in an accelerated time scale. A novel feature is the use of a continuously-rotating magnetic drum with multiple heads to delay one signal relative to another. Delays equivalent to 25 seconds in recording time can be obtained, the smallest delay being 0·05 second. Tests with sinusoidal signals show that frequencies can be measured to better than 2% and the r.m.s. error of all points on the correlation curve is less than 2% of full scale. These errors are shown to be considerably less than the statistical errors due to the finite recording time.

List of Symbols

$a(t)$ impulse response of a system
$A(\omega)$ frequency response of a system
$b(t)$ reactivity response to δ-function change in inlet temperature
$B(s)$ transfer function between reactivity and inlet temperature
$c_i(t)$ population of ith group of delayed neutron precursors
C_i mean value of $c_i(t)$
$f(t)$ impulse response of thermocouple
$F(s)$ transfer function of thermocouple
$g(t)$ response of simple model of reactor to δ-function reactivity change
$G_0(s)$ zero energy reactor transfer function
$k(t)$ effective multiplication constant
$K(s)$ transfer function of reactor power feed-back
prompt neutron lifetime
$n(t)$ extraneous noise source
$n'(t)$ neutron population
N mean neutron population
P reactor mean power
$P_{uv}(\omega)$ Fourier transform of $X_{uv}(\tau)$
T duration of noise sample
T_1 time constant of a lag
T_θ thermal time constant of inlet thermocouple
$u(t)$ input to system
$u'(t) = u(-t)$
$v(t)$ output of system
$X_{uv}(\tau)$ cross-correlation function of $u(t)$ and $v(t)$
α prompt reactivity feed-back per unit power
β_i fraction of neutrons in ith group
γ_i fractional change in population of ith group of percursors
$\theta(t)$ input temperature noise
λ_i decay constant of ith group
v fractional change in neutron population
ρ reactivity
τ correlation parameter
$\phi(t)$ thermocouple output

* Dounreay Experimental Reactor Establishment, Caithness.
† U.K. Atomic Energy Authority, Engineering Group Headquarters, Risley, Lancs.; formerly at Dounreay Experimental Reactor Establishment.
‡ Bruce Peebles Ltd., Edinburgh.

1. Introduction

In common with most physical processes, fluctuations are observed in various parameters (power, temperature, etc.) of a nuclear power reactor operating at a nominally constant power.[1-3] The fluctuations (noise) can be attributed to a variety of causes. Fission itself is fundamentally a random process and, in low-power experimental reactors, is the primary cause of power noise. However, in a power reactor the fission rate is high and the statistical variations are relatively small, several orders of magnitude less than is observed in practice. Other causes must be sought.

Power reactors are sensitive to changes in temperature of fuel and coolant in the core since fuel temperature affects the fission cross-section and the temperature of the coolant and structural materials in the core affect the probability of neutron capture. In most power reactors the flow of coolant entering the core is often turbulent, and there will be random fluctuations in temperature and flow. These then are likely causes of power fluctuations. Another possible source is small random movements of the control rods.

Since the fluctuations at one point in the reactor system can be regarded as input signals, the possibility exists of measuring the dynamic characteristics of the system by considering the fluctuations at another point. For example, if a significant proportion of the fluctuations in reactor power can be correlated with fluctuations in the inlet coolant temperature, it is possible in theory to obtain information on the kinetics of the core by regarding the inlet fluctuations as the input to a linear system.

The fundamental parameter of the reactor which causes changes in power is the multiplication factor associated with neutron population. In steady operation it is, of course, unity. For our purpose we define reactivity as the excess multiplication factor over unity. In considering the stability of the reactor it is important to determine the transfer function of reactor power with respect to changes in reactivity. This is often measured by means of an oscillator, a rod coated on one side with an absorber, which rotates in a position of high neutron flux gradient and which varies the reactivity at a known frequency. However, the oscillator takes up valuable space in the reactor, is not always available, and seriously perturbs the operating condition of the reactor.

Since random fluctuations in temperature of the inlet coolant also affect reactivity, it is possible in theory to use these to measure the transfer function without causing any more disturbance to the reactor than is normally present. Work along these lines is reported by Rajagopal.[4] It seems unlikely that random noise analysis can offer as accurate an estimate of the transfer function as some form of driven perturbation, but the possibility of keeping a continuous check on the kinetic characteristics of the reactor without disturbing its operation could be a valuable safety asset.

The theoretical principles of noise analysis are well established. Our purpose here is to outline the design principles of an equipment to investigate noise in the Dounreay reactors. A brief description is given of the kinetics of a power reactor and the main parameters of the noise analysis equipment are then established. The design of the apparatus is treated as an exercise in systems engineering, since most of the component units are available commercially and only the more important design considerations will be discussed here. Some preliminary results on the Dounreay Fast Reactor are given.

2. Theory

Two alternative methods can be used to describe the characteristics of noise signals in a system in such a way as to derive information on its dynamics. They are analogous to the two ways of describing the response of a linear system to an input signal, namely, by its frequency response or by its impulse response. The former is often more familiar to an electronics engineer because of the simple relation between input and output spectra, but has the disadvantage that it requires a complex representation or alternatively both amplitude-frequency and phase-frequency characteristics. Operation in the time domain is strictly real and requires therefore only one relation to define it. In the analysis of noise the equivalent quantities are, in terms of frequency, the power spectrum of a noise signal and the cross-power spectrum of two signals and, in the time domain, the auto-correlation function of a signal and the cross-correlation function of two signals. The time and frequency concepts are completely equivalent, one can be obtained from the other by Fourier transformation and no more information is available from one than from the other.[5] The choice between them is purely one of convenience and ease of interpretation. Our equipment is based on the time concept and the emphasis will be accordingly on this aspect, although it will sometimes be convenient to consider the equivalent frequency representation.

Consider a linear system specified by its impulse (delta-function) response $a(t)$. The frequency response $A(\omega)$ is the Fourier transform of $a(t)$. The output $v(t)$ of the system due to any input $u(t)$ is given by the well-known convolution integral

$$v(t) = \int_{-\infty}^{\infty} a(x)u(t-x)\,dx \qquad \ldots\ldots(1)$$

which we shall write for convenience

$$v = a \star u \qquad \ldots\ldots(2)$$

The corresponding expression in the frequency domain is

$$V(\omega) = A(\omega)U(\omega) \qquad \ldots\ldots(3)$$

Suppose now that the input is a random fluctuation. In many practical cases it will consist of a small fluctuation about a much larger mean and, for our purpose, we shall assume that the mean has been suppressed and consider only the fluctuation $u(t)$. The auto-correlation function is defined by

$$X_{uu}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{1}{2}T}^{\frac{1}{2}T} u(t)u(t+\tau)\,dt \quad \ldots\ldots(4)$$

which is seen to measure the mean correlation between the signal $u(t)$ and itself delayed by a time interval τ. In practice it is not, of course, possible to proceed to the limit and we have to consider a long but finite sample of $u(t)$ so that the auto-correlation function we compute

$$X_{uu}(\tau) = \frac{1}{T} \int_{-\frac{1}{2}T}^{\frac{1}{2}T} u(t)u(t+\tau)\,dt \quad \ldots\ldots(5)$$

must be regarded as a sample of an ensemble. Its statistical properties are discussed later with a view to determining an acceptable sampling period (T).

By a simple change of variable, equation (4) can be written

$$X_{uu}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{1}{2}T}^{\frac{1}{2}T} u(-x)u(\tau-x)\,dx \quad \ldots\ldots(6)$$

The integral can be recognized as the convolution integral (equation (1)), and by analogy with equation (2), we can write

$$X_{uu} = \lim\,(u'\star u) \quad \ldots\ldots(7)$$

where the limit sign represents the process of deriving the mean as well as proceeding to the limit as T tends to infinity and where

$$u'(t) = u(-t) \quad \ldots\ldots(8)$$

Consider now the auto-correlation of the output $v(t)$ due to the fluctuations $u(t)$ at the input. Appendix 1 shows that, in our notation,

$$X_{vv} = a'\star a\star X_{uu} \quad \ldots\ldots(9)$$

In some systems the input fluctuations can be regarded as 'white' noise, that is, as having a uniform power spectrum. The equivalent auto-correlation function is a delta-function at $\tau = 0$, i.e.

$$X_{uu} = K\delta(\tau) \quad \ldots\ldots(10)$$

Substituting into equation (9) shows that

$$X_{vv} = Ka'\star a \quad \ldots\ldots(11)$$

or, in terms of frequency, the power spectrum $P_{vv}(\omega)$ of the output is, from equation (47), writing $P_{uu}(\omega) = K$.

$$P_{vv}(\omega) = K|A(\omega)|^2 \quad \ldots\ldots(12)$$

In these circumstances the auto-correlation function or the power spectrum of the output is a measure of the frequency response of the system. If the input noise is not 'white', however, at least over the relevant frequency band, its auto-correlation function or its power spectrum must also be obtained.

A further difficulty in using the auto-correlation function of the output to derive information on the system is apparent when we consider the effects of a second noise source. Suppose that, in addition to those due to the input, the output contains other fluctuations, as will often be the case in practice. We can then write

$$v = a\star u+n \quad \ldots\ldots(13)$$

where $n(t)$ represents fluctuations uncorrelated with $u(t)$. Applying equation (7)

$$\begin{aligned} X_{vv} &= \lim\,(v'\star v) \\ &= \lim\,(a'\star u'+n')\star(a\star u+n) \\ &= a'\star a\star X_{uu}+X_{nn} \quad \ldots\ldots(14) \end{aligned}$$

where X_{nn} is the auto-correlation function of the extraneous noise $n(t)$. The cross terms in the product vanish in the limit, since $n(t)$ is assumed uncorrelated with $u(t)$. Thus a measurement of the auto-correlation function of the output will contain that of the uncorrelated noise, which can make the derivation of information on the response of the system difficult if not impossible.

A more fruitful approach is to consider the cross-correlation function between input $u(t)$ and output $v(t)$ defined by

$$X_{uv}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{1}{2}T}^{\frac{1}{2}T} u(t)v(t+\tau)\,dt \quad \ldots\ldots(15)$$

By analogy with equation (7) this can be written

$$X_{uv} = \lim\,(u'\star v) \quad \ldots\ldots(16)$$

Substituting from equation (2) gives

$$X_{uv} = \lim\,(u'\star a\star u)$$

By an argument similar to that of Appendix 1 it can be shown that

$$X_{uv} = a\star X_{uu} \quad \ldots\ldots(17)$$

and the response of the system can be obtained. In frequency terms this is

$$P_{uv}(\omega) = A(\omega)P_{uu}(\omega) \quad \ldots\ldots(18)$$

where $P_{uv}(\omega)$ is the Fourier transform of the cross-correlation function X_{uv}.

Consider now the effect of extraneous noise. Substituting equation (13) into equation (16) and rearranging gives

$$X_{uv} = a\star X_{uu}+\lim\,(u'\star n)$$

in which the second term vanishes in the limit, since $n(t)$ is uncorrelated with $u(t)$. Hence, again

$$X_{uv} = a\star X_{uu}$$

even in the presence of uncorrelated noise. In practice the ability to discriminate against extraneous noise is limited by the finite sampling time which should be long compared with the correlation times of both wanted and unwanted noise signals.

It is relevant at this stage to consider the effect of the finite sampling. A measurement of a correlation function must necessarily be performed on a finite sample of the noise in the relevant parameters and so represents one estimate of the 'true' correlation function which, in a stationary system, will be the mean of all such samples. We require to know the standard deviation of these samples. It will, in general, depend not only upon the sample length (T), but also on the correlation function itself, so that we must make some assumptions about the characteristics of the noise we wish to measure. In many physical systems the auto-correlation function of the noise in a parameter will approximate to an exponential or a sum of exponentials. For example, the impulse response of a lag having a time constant T_1 is

$$a(t) = \frac{1}{T_1}\exp(-t/T_1) \quad [t > 0] \qquad \ldots\ldots(19)$$

and if the input to the lag is 'white' noise, equation (11) shows that the auto-correlation function of the output is

$$X_{vv} = \frac{K}{2T}\exp(-|\tau|/T_1) \qquad \ldots\ldots(20)$$

We shall therefore consider the standard deviation in the estimates of the auto-correlation function of a noise signal whose 'true' auto-correlation function is $\exp(-|\tau|/T_1)$, the estimates being obtained from samples of length T very much greater than the time constant (T_1). The details are given in Appendix 2. There it is shown that, in any one sample, there is high correlation between the errors at two points on the curve separated by less than one time constant, so that the scatter on the curve obtained from one sample will be small. This is observed in practice. Also the ratio r.m.s. error/mean will be $\sqrt{(2T_1/T)}$ for values of τ small compared with the time constant T_1, and approximately twice this if $\tau = T_1$.

3. Nuclear Reactor Kinetics

Before discussing the design of a correlator to measure the noise properties of a nuclear reactor we shall first consider a model of the reactor system. We shall regard the reactor as a point system for which the kinetic equations for the neutron population (n') are (see, for example, Sastre[6])

$$\frac{dn'}{dt} = [(1-\beta)(k-1)-\beta]n'/l + \sum_{i=1}^{6}\lambda_i c_i \quad \ldots(21)$$

and

$$\frac{dc_i}{dt} = \beta_i k n'/l - \lambda_i c_i \qquad = 1 \ldots 6 \qquad \ldots\ldots(22)$$

In these equations the assumption is made that most of the neutrons produce fission after a short time (l), the prompt neutron lifetime, but that a small fraction (β) produce precursors which later decay to produce neutrons. There six such groups of delayed neutrons, the fraction of precursors produced in the ith group being denoted by β_i and the time constant of the decay by $1/\lambda_i$. In the steady state the effective multiplication constant (k) is unity. However, in practice the multiplication constant can vary with time due to movement of control rods or, as we have discussed, due to temperature changes in the reactor core. It is then apparent that the equations are non-linear and do not permit of an analytic solution for a general time-varying multiplication factor.

However, when considering small changes about a steady state we may linearize the equations by writing

$$n'(t) = N(1 + \nu(t))\ ;\ c_i(t) = C_i(1 + \gamma_i(t))\ ;$$
$$k = 1 + \rho(t) \qquad \ldots\ldots(23)$$

where N and C_i are the mean values of neutron and precursor population, $\nu(t)$ and $\gamma_i(t)$ the fractional changes, assumed small, and $\rho(t)$ is the reactivity. Neglecting second order quantities the kinetic equations then become

$$\frac{d\nu}{dt} = (1-\beta)\rho/l - \beta\nu/l + \sum_i \lambda_i C_i \gamma_i/N \quad \ldots(24)$$

$$(C_i/N)\frac{d\gamma_i}{dt} = \beta_i(\rho+\nu)/l - \lambda_i C_i \gamma_i/N \qquad \ldots\ldots(25)$$

We can now solve these for ν in terms of ρ by taking Laplace transforms and eliminating γ_i. The Laplace transfer function of the system can then be written (denoting Laplace transforms by a superscript bar)

$$\bar{\nu}/\bar{\rho} = \left[1 - \sum_{i=1}^{6} s\beta_i\Big/(s+\lambda_i)\right]\Big/ s\left[l + \sum_{i=1}^{6}\beta_i\Big/(s+\lambda_i)\right] \qquad \ldots\ldots(26)$$

It is clear that the denominator will have seven factors so that the response is complicated and difficult to handle analytically. To simplify what follows and to obtain an intelligent insight into the performance of the system we consider one delayed neutron group so that, with an obvious notation, the transfer function becomes

$$\bar{\nu}/\bar{\rho} = [1 - s\beta(s+\lambda)]/s[l + \beta/(s+\lambda)]$$

which simplifies to

$$\bar{\nu}/\bar{\rho} = [s(1-\beta)+\lambda]/s[ls+\beta+l\lambda] \qquad \ldots\ldots(27)$$

In practice the delayed neutron fraction (β) is about 7×10^{-3} and hence small compared with unity so that we can write

$$\bar{\nu}/\bar{\rho} = (s+\lambda)/ls(s+\lambda^*) = G_0(s) \qquad \ldots\ldots(28)$$

where $\lambda^* = \lambda + \beta/l$.

The above remarks refer to a reactor system in which changes in temperature due to changes in power can be neglected. In practice in a power reactor a small change in power produces a proportional small change in temperature of the fuel, coolant and structure and these in turn produce proportional small changes in reactivity due to the induced changes in absorption, fission and escape of the neutrons. We thus have a feed-back system as shown in Fig. 1 where $K(s)$ represents the feed-back per unit change in power. The feed-back will not be instantaneous since there will in general be lags due to heat transfer effects. Since $v(t)$ is defined as a fractional change in neutron population or power (neutron population and power are proportional) the actual feed-back transfer function is $PK(s)$ and from feed-back theory the modified response of the system is

$$\bar{v}/\bar{\rho} = G_0(s)/[1-PK(s)G_0(s)] \quad ......(29)$$

where $G_0(s)$ is given by equation (28).

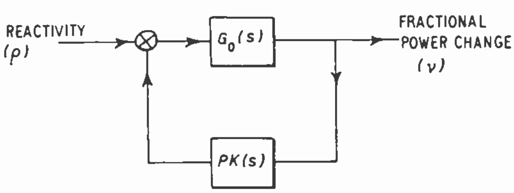


Fig. 1. Model of reactor feed-back.

Again to simplify the argument let us consider the effect of a rapid negative feed-back $K(s) = -\alpha$ per unit power. Substituting for $G_0(s)$, the overall response of the reactor is given by

$$\bar{v}/\bar{\rho} = (s+\lambda)/l[s^2+s(\lambda^*+P\alpha/l)+\lambda\alpha P/l] \quad ...(30)$$

The expression has two poles s_1 s_2 given by

$$s_1, s_2 = \tfrac{1}{2}(\lambda^*+\alpha P/l)\pm\tfrac{1}{2}[(\lambda^*+\alpha P/l)^2-4\lambda\alpha P/l]^{\frac{1}{2}} \quad (31)$$

To simplify this further we consider typical values of λ, β and l. The delayed neutron fraction (β) is typically about 7×10^{-3}. The prompt neutron lifetime (l) is of the order 10^{-4} seconds in a thermal reactor and as little as a few microseconds in a fast reactor. The value $\lambda = 0{\cdot}08$ makes the approximate form of $G_0(s)$ in equation (28) a good approximation to the more accurate form given by equation (26). Under these circumstances it is readily shown that good approximations to s_1, s_2 are given by

$$s_1 = (\beta+\alpha P)/l \quad ; \quad s_2 = \lambda\alpha P/(\beta+\alpha P) \quad ...(32)$$

In a thermal reactor s_1 is of the order $10\,s^{-1}$ and in a fast reactor of the order $10^3\,s^{-1}$ corresponding to time constants of 0·1 s and 1 ms respectively. The second root will be considerably less than $1\,s^{-1}$.

Returning now to equation (30) we have

$$G(s) = \bar{v}/\bar{\rho} = (s+\lambda)/l(s+s_1)(s+s_2) \quad ......(33)$$

and the impulse response of our simple system will be of the form

$$g(t) = A_1\exp(-s_1 t)+A_2\exp(-s_2 t) \qquad t>0 \;...(34)$$

If the reactor power noise can be assumed due to changes in k having the characteristics of white noise at least within the frequency band from d.c. to $\omega = s_1$ the auto-correlation function of reactor power noise will be

$$X_{vv}(\tau) = Kg\star g'$$

When $g(t)$ is of the form of equation (34) it is readily shown that $X_{vv}(\tau)$ is of the form

$$X_{vv}(\tau) = B_1\exp(-s_1|\tau|)+B_2\exp(-s_2|\tau|) \quad ...(35)$$

With the assumption of white noise disturbance a measurement of the auto-correlation function of reactor power noise (X_{vv}) will give estimates for s_1 and s_2 and hence of the feed-back.

It is however by no means certain that the random changes in reactivity in a power reactor have the characteristics of white noise. For example some of the noise may be due to random changes in coolant flow which produce changes in temperature and hence reactivity and will inevitably have relatively long correlation time-constants because of the inertia of pumps etc.

Rapid changes in reactivity cannot be measured directly and one is forced to postulate mechanisms which can cause them and carry out measurements on the parameters of the initiating mechanism. One such source could be fluctuations in temperature of the coolant entering the reactor due perhaps to the random mixing of several coolant streams at slightly different mean temperatures. The model then becomes that shown in Fig. 2, in which

$\theta(t)$ = random temperature noise

$\phi(t)$ = measured temperature fluctuation

$B(s)$ = transfer function between inlet temperature and reactivity

$F(s)$ = transfer function of the temperature measuring instrument (usually a thermocouple)

and the other parts of the system have already been defined. If we denote

$b(t)$ = impulse response of reactivity to changes in temperature

$f(t)$ = impulse response of thermocouple

we have

$$v = b\star g\star\theta \quad \text{and} \quad \phi = f\star\theta$$

Hence the power auto-correlation function is given by

$$X_{vv} = b'\star b\star g'\star g\star X_{\theta\theta} \qquad ......(36)$$

and the cross-correlation function of $\phi(t)$ and $\nu(t)$ by

$$X_{\phi\nu} = f' \star b \star g \star X_{\theta\theta} \qquad \ldots\ldots(37)$$

Thus if we have an estimate of $f(t)$ and $b(t)$ from theoretical considerations, we can obtain some information on the feed-back reactor transfer function by measuring the cross-correlation function between temperature fluctuations and fractional power changes and the auto-correlation function of the temperature fluctuations.

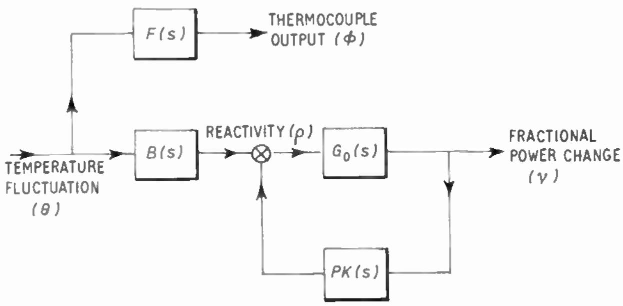


Fig. 2. Model of reactor response to inlet temperature fluctuation.

There remains the practical problem of obtaining quantitative information from the correlation functions. We may assume the impulse response of the thermocouple to be

$$f(t) = \frac{1}{T_\theta}\exp(-t/T_\theta) \qquad t > 0 \qquad \ldots\ldots(38)$$

where T_θ is its thermal time constant. Now the auto-correlation function of the measured temperature is given by

$$X_{\phi\phi} = f' \star f \star X_{\theta\theta} \qquad \ldots\ldots(39)$$

and in practice the correlation time of the true temperature noise which is a property of the turbulent mixing, is very much less than the thermocouple time-constant, so that $X_{\theta\theta}$ may be regarded as a δ-function and

$$X_{\theta\theta} = \bar{\theta}^2\,\delta(\tau) \qquad \ldots\ldots(40)$$

where $\bar{\theta}^2$ is the mean square temperature noise. Thus, from equation (20)

$$X_{\phi\phi} = \frac{\bar{\theta}^2}{2T_\theta}\exp(-|\tau|/T_\theta) \qquad \ldots\ldots(41)$$

and a measurement of the auto-correlation function $X_{\phi\phi}$ gives a measure of both $\bar{\theta}^2$ and T_θ. Also

$$f' {}_{\star} X_{\theta\theta} = (\bar{\theta}^2/T_\theta)\exp(t/T_\theta) \qquad t < 0 \quad \ldots\ldots(42)$$

Substituting in equation (37) we have to solve the equation

$$X_{\phi\nu} = c \star X$$

where $c = b \star g$, $X = f' \star X_{\theta\theta}$, which is known, and $X_{\phi\nu}$ has also been measured.

Three methods of solution suggest themselves. We can build up a model and derive expressions for b and g and hence c with perhaps some unknown parameters we wish to measure and then attempt to fit the measured correlation function. Secondly, we can formally solve the equation by using Fourier transforms:

$$F(X_{\phi\nu}) = F(c).F(X)$$

Hence at frequencies at which $F(X)$ is not negligible we can obtain $F(c)$ and perform an inverse transformation to find $c(t)$. It should be appreciated that in general $F(X_{\theta\nu})$ and $F(c)$ are complex quantities. Thirdly, following a method suggested by Mr. E. P. Hicks of the U.K.A.E.A., we can use the algebraic identity

$$X_{\phi\nu}(\tau_i) = \sum_j c(\tau_j).X(\tau_{i-j})$$

in which we assume $X_{\phi\nu}$ and X to be given at a series of discrete times of equal intervals. These equations form a set of linear equations for the values $c(\tau_i)$ which can be solved on a computer. Having derived $c = b \star g$ we can use similar techniques to obtain $g(t)$ if some formulation of $b(t)$ can be obtained from a physical model.

One further practical problem remains. The correlation functions have inevitably to be terminated at finite values of τ and it is often the case that the value of the correlation function at the terminal points is by no means zero. This 'window' problem is well-known and is discussed in detail by Blackman and Tukey.[7] Their method of solution is to weight the correlation function by a suitable weighting function which reduces the value of the correlation function at the terminal points to zero and to use the Fourier transform technique. The effect of the weighting function is similar in physical terms to attempting to measure a frequency spectrum with a filter of finite bandwidth. We have used a different technique. We have assumed that in our application the long-term correlation falls off exponentially, as appears to be the case, and have fitted an exponential decay to the 'tail' of the correlation function. This enables us to extrapolate the correlation function as far beyond the terminal points as we wish.

Most of these techniques of analysis are currently being used in work on the Dounreay Fast Reactor.

4. Equipment

4.1. *General Design Considerations*

In order to investigate the practical problems of employing noise analysis to determine the kinetic characteristics of the power reactors at Dounreay a correlator has been built. The design was based on known techniques using as far as possible commercially available equipment. Considerable attention was given to the system engineering aspects, i.e. the selection of a chain of units which when linked together would produce an equipment which satisfied a specification determined by the theoretical aspects of noise analysis and the parameters of the reactor systems. The units fall into one of two classes, proprietary items with perhaps minor modifications and some special units of relatively straightforward design. By this philosophy development time and effort were kept to a minimum. There were however some interesting and novel features in the system engineering which will be highlighted in the following sections.

The basic requirement of the analysis system is defined by equation (15) omitting the limiting process. To compute the value of the correlation function at a selected value of τ, one signal is delayed relative to the other by the time interval τ, the two signals are multiplied together and the product integrated over the chosen sample period (T). A selection of values of delay (τ) should be available using the same samples of the two noise signals. First estimates of the reactor feed-back suggested that a maximum delay of 25 seconds should be provided, while the estimated thermal time-constants in the system defined a minimum delay of 0·05 second. A wider range of delays would, of course, increase the flexibility of the equipment but could lead to increased complexity. A selection of values of τ was required to enable time constants from about 0·2 second to 20 seconds to be measured.

To carry out the computation both analogue and digital techniques were considered. Digital techniques offer considerable flexibility and speed of operation although the sampling, digitizing and punching of several channels to, say, 10% accuracy at 0·05 second intervals poses equipment problems. In addition the storage capacity required is large, e.g. sampling four channels at 0·05 second intervals for one hour requires about 3×10^5 words and this presents problems in the organization of the program if a general-purpose computer is used. The greatest objection to digital techniques for our application, however, was that access was not readily available to a general-purpose computer. Since it was felt desirable to be able to analyse a recording as soon as possible to provide the possibility of controlled experiments on the reactor, the system would have to include its own special-purpose or general-purpose computer. The cost of this encouraged us to consider the possibility of analogue computing.

The main design considerations in an analogue system are the accuracy and the method of storage to provide the delay. It can readily be shown from Appendix 2 that although samples of several hours duration are required to measure time constants of the order of 20 seconds to a statistical accuracy better than 10%, nevertheless considerably greater accuracy can be achieved in the measurement of more rapid phenomena for the same sample duration. To exploit this it was felt that an overall instrumental accuracy of $\pm 2\%$ should be aimed at and this was considered to be within the capabilities of a high quality analogue system. A more precise statement of equipment performance is given later.

Various techniques were considered to provide the delay, including recording optically on film and using an electronic film reading system, or magnetic tape recording with either an auxiliary tape loop with movable reading heads or a rotating magnetic drum. The drum appeared to be the most straightforward and required the least development, since drums of the required performance were available commercially. In the practical application of the equipment it was desirable to record the signals and carry out the computation as separate operations. The most convenient recording system compatible with a drum delay unit is frequency-modulated recording on magnetic tape, since the signals can then be transferred directly from the tape reading heads to the drum writing heads via an amplifier.

We have already seen that recordings of several hours duration are required and it is apparent that, to restrict within acceptable limits the time to analyse a recording, the instrument must either work in parallel, i.e. individual delay, multiplication and integration circuits must be provided for each value of delay required in the correlation function, or the recording must be played back at increased speed and the computations carried out in series by one set of equipment. On grounds of economy the latter is preferable and has the additional advantage that it enables reasonable drum speeds to be used. For example, a speed ratio of 128 between replay and recording, which is readily obtainable from instrumentation recording decks, reduces the maximum real-time delay of 25 seconds to about 0·2 second in the replay system. This time corresponds to one drum revolution at 300 rev/min.

There remained the problem of providing the wide range of delays, a minimum increment of 0·05 second and a maximum of 25 seconds. This could not be achieved only by the spacing of heads round the drum. It was in practice obtained by a suitable combination of head spacing, drum speeds, play-back speeds and the use of two tracks in series on the drum. Preliminary measurements of the amplitude and spectrum of the noise introduced by writing a signal on the drum, reading it off and demodulating it showed that its amplitude in the relevant signal band was negligible at all relevant drum speeds.

The preliminary assessment showed that an analogue system based on frequency modulated recording on magnetic tape, together with a rotating drum to provide the delay, would provide sufficient accuracy and flexibility, could be assembled from commercially available units with little adaptation, and would be considerably cheaper than a digital system.

4.2. *System Design Details*

We shall now consider some of the more interesting aspects of the detailed design of the system. Block diagrams of the recording and analysis systems are shown in Figs. 3 and 4. Four channels are provided for recording and the auto-correlation of any one or the cross-correlation function of any pair can be computed in the analysis phase, the selection being carried out on a patch board.

The input signals in general consist of a small fluctuation imposed on a much larger mean level, for example the fluctuations in a thermocouple output might have an r.m.s. value of 10 microvolts superimposed on a mean level of about 20 millivolts. The theory in the previous section is based on the assumption that the recorded signals have zero mean. It is readily shown that, if the mean is not zero, an additional term whose value is independent of τ is introduced into the correlation function. This complicates its interpretation and should be kept as small as possible. In practice the error in the mean should be considerably less than the r.m.s. value of the noise which is being measured. This is achieved by providing in the recording equipment a backing-off voltage which must be adjustable and stable to a few microvolts over a period of several hours. A Mallory cell is used which has been aged sufficiently so as to be operating on the plateau of its voltage-time curve.

To assist in setting-up the backing-off voltage before recording, an integrating amplifier, having a time-constant of several seconds and an approximately logarithmic response, can be switched into any channel. The direction of deflection of a meter indicates the sign of the error and the drift rate its magnitude.

Since the backing-off can be adjusted, in the recording phase, only on signals prior to the actual recording, it is still possible that some adjustment of the mean will be required in the analysis phase and means for this are provided in the analysis equipment. A relatively small d.c. signal is added to the recorded noise signal, after demodulation, which can be adjusted until the mean signal from the whole recording, as measured by the integrator, is zero.

At the input terminals of each of the four recording channels arrangements are made to accommodate signals from thermocouples as well as from ion chambers etc. Dewar flasks containing an ice/water mixture are used to provide the cold junction. The system is capable of measuring fluctuations in temperature of the order 0·1 deg C and to avoid the introduction of unwanted fluctuations of this magnitude by random contact of ice with the cold junction, the latter is protected by a wire mesh.

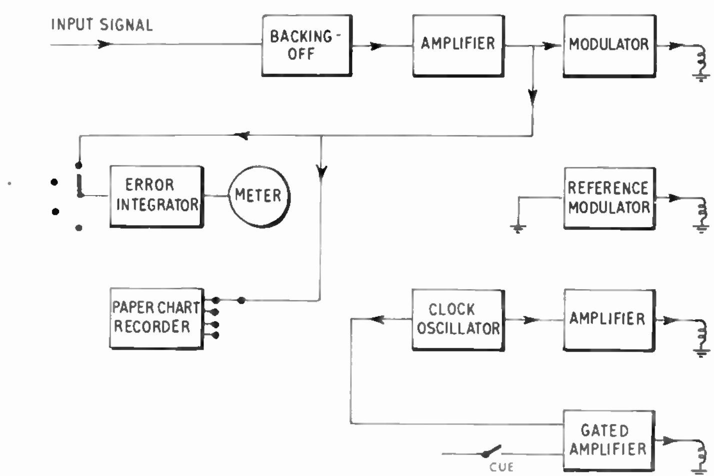


Fig. 3. Block schematic of recording system.

The signal amplifiers were selected for low noise, low drift, high gain and good common mode rejection to reduce to an acceptable level the high level of 'hum' picked up on the long leads in a reactor environment. Commercially available chopper-stabilized transistor d.c. amplifiers were used having fixed voltage gains of

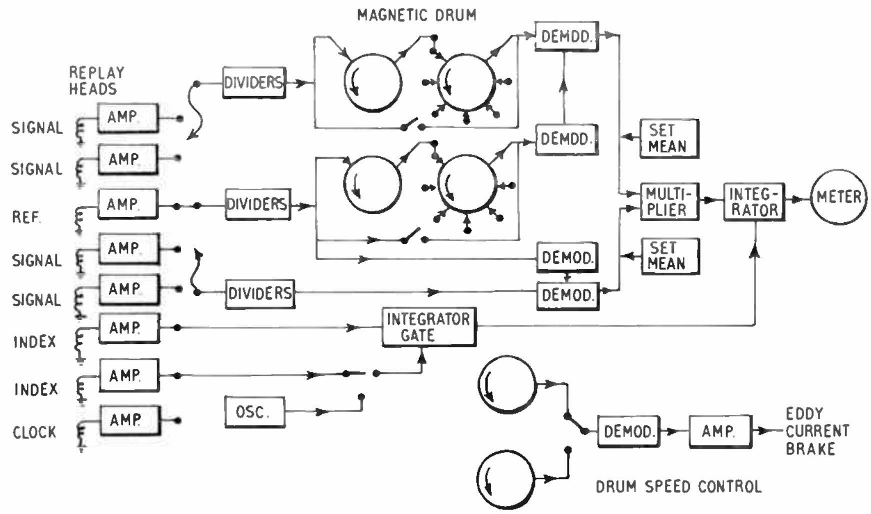


Fig. 4. Block schematic of analysis system.

up to 10^4. Further amplification is provided by standard computing amplifiers to raise the peak signal to ± 1 volt for insertion into the magnetic tape recording system. The signal to be recorded is displayed on a four-pen paper chart recorder to enable the gain in each channel to be adjusted to make full use of the dynamic range of the recording equipment.

The signals are recorded on 1-in wide magnetic tape running at 15/16 in/s by a commercially available instrumentation equipment. The system uses frequency modulation of a 937·5 c/s carrier, maximum deviation frequency being 312·5 c/s. A reference carrier is also recorded so that first-order correction can be made in the demodulators for slow variations in tape speed of both recording and replay decks. Simple subtractive correction is used and although this cannot compensate for differential drift in the centre frequencies between one channel and another nor for differential drift in the slopes of the signal and reference channel modulators and demodulators, the net error can be kept well below $\pm 2\%$ of full scale in terms of signal amplitude by routine adjustment.

In addition to the signals and the reference, a 1000-c/s crystal-controlled clock signal is also recorded, and a further two tracks are available for indexing signals. The accurate clock signal not only allows accurate calibration of tape speeds and delay times to be made, but also serves as a signal of known (20% full scale) amplitude for overall calibration of the gains of the analysis equipment.

The design of the analysis equipment is largely dictated by the requirements of the magnetic drum. This has 17 equally spaced heads, one for reading and 16 for writing (thus avoiding the necessity to switch the relatively small read signals when changing the delay parameter (τ)). Its maximum speed of rotation is about 2400 rev/min and it can be controlled by drive voltage and eddy current braking over a speed range of about 4 : 1. In addition, measurements showed that two tracks could be used in series without increasing the extraneous noise to an unacceptable level. We thus have a ratio of 128 : 1 between the smallest increment of delay and the maximum delay. If in addition we can replay the recording at different tape speeds on the replay deck over a range of 4 : 1 the maximum/minimum delay ratio can be increased to 512 and is adequate for our purpose.

The replay magnetic tape deck, identical in design to the recording deck, has replay speeds of 30, 60 or 120 in/s, representing speed-up factors of 32, 64 or 128 respectively. If the smallest increment of delay in recording time is 0·05 second, this will be obtained at a replay speed of 30 in/s and the required drum speed (w) is given by

$$60/17w = 0{\cdot}05/32$$

where w is in rev/min, whence the maximum required drum speed is 2250 rev/min. By using replay speeds of 30, 60 or 120 in/s, drum speeds of 562·5 or 2250 rev/min and two tracks in series on the drum, the delays given in Table 1 (measured in recording time) are available.

Table 1

Range of delay	(seconds)	0–1·6	0–3·2	0–6·4	0–12·8	0–25·6
Increments	(seconds)	0·05	0·10	0·2	0·4	0·8

This meets our requirement since rapid phenomena are measured at intervals of 0·05 second, the longest delay is in excess of 25 seconds and it is usually permissible to measure longer term phenomena at increasingly coarse intervals.

Two further design criteria in the use of the drum are of importance. Because of the high rotation speed and the possibility of wear, the heads are out of contact with the drum, the clearance being about 0·001 in. Hence the resolution of the reading and writing heads is of the order of 0·002 in and the wavelength of the carrier should be considerably greater than this. Now with a tape replay speed of 120 in/s the carrier is 120 kc/s which corresponds to a wavelength on the drum of about 0·002 in. It is therefore necessary to reduce the carrier frequency on the drum. Moreover for simplicity of operation it is desirable that the demodulators operate at the same carrier frequency irrespective of replay speeds. Before writing on the drum, therefore, the frequency of the replayed f.m. signal is divided by binary dividers in the ratio of 2, 4 or 8, depending on replay speed, to give a carrier of 15 kc/s which corresponds to a wavelength on the drum of 0·016 in. The division has no effect on the signal information since both the percentage deviation and its rate of change are unaffected. The division is made possible by the unusually high ratio (approximately 100) of recording carrier frequency to maximum signal frequency, which even after division remains sufficiently large (10 in the worst case).

A further design consideration is the small tolerance on delay required by the system. If, for example, the correlation function is of the form $A\exp(-\tau/T_1)$ and there is an error $\delta\tau$ in the delay at some value τ, the corresponding fractional error in the computation of the correlation function at that point is $\delta\tau/T_1$. Since one is usually interested in values of τ up to T_1, this implies that the acceptable fractional error in the delay is less than the magnitude of the required instrumental accuracy, namely $\pm 2\%$. Errors in delay can be obtained by inaccurate positioning of heads on the drum and by deviations in drum speed from its nominal value. The former is reduced by accurate adjustment of the heads. The drum speed is held constant by servo control to better than 0·5% using pre-recorded clock tracks of known wavelength to measure true speed and thus control an eddy current brake system.

The design of the remainder of the analysis system is now straighforward. After demodulation the delayed and undelayed signals are multiplied together in a crossed-fields multiplier having an accuracy of $\pm 1\%$ full scale and a phase error of less than 1 deg in the band 0–1000 c/s. This bandwidth is not strictly adequate for the original 10 c/s recorded bandwidth when multiplied by a speed-up factor of 128, but, in general, when operating with the highest speed-up factor, one is interested in components of the signal having long correlation times corresponding to low frequencies, when the bandwidth limitation can be ignored.

The multiplier output is integrated by a conventional electronic integrator provided with a choice of low leakage capacitors in the feed-back path to preset the integrator constant. The duration of the integration is controlled electronically by signals on the index channels of the tape to ensure that, for each value of delay, the same section of record is analysed. Since, to obtain an acceptable statistical accuracy, the duration of the recording is at least 100 times the maximum delay, end effects can be ignored.

4.3. *System Tests*

To examine the overall performance of the system sinusoidal signals of known amplitude (200 μV r.m.s.) at frequencies of 0·1, 1·0 and 6·25 c/s were recorded on all four channels and the equipment used to compute their auto and cross correlation functions. The test signals were sufficiently large to allow the effects of noise in the recording amplifiers to be neglected. It is readily shown that the mean auto-correlation function of a sinusoidal signal of frequency f and r.m.s. amplitude A is

$$A^2\cos(2\pi f\tau)$$

The measured results were compared with this at the values of τ available in the equipment. It was found that the major error could be described as a phase shift proportional to delay (τ) and was equivalent to an error of less than 2% in the measurement of the frequency of the test signal. After correction had been made for this, the r.m.s. error averaged over all points was less than 2% of full scale. The consistency with which the value of the correlation function at any value of delay could be obtained was better than ½% of full scale.

A measure of the overall noise in the equipment was obtained by shorting the inputs to all four channels and recording the noise in the recording system at full gain. The mean square noise was computed by the analysis equipment and the results showed that the equivalent r.m.s. signal at the input was less than 3 microvolts. The presence of the drum delay system in the circuit appeared to make no measurable difference to the measured mean square noise.

5. Typical Results

The apparatus is being used to analyse fluctuations in temperature and power in the Dounreay Fast Reactor. It is not proposed to discuss the results in detail here since they are as yet incomplete, but to illustrate the use of the equipment.

Power fluctuations were measured by an ion chamber and are expressed as a percentage of the mean power (60 MW). Fluctuations in temperature of the inlet coolant flow entering the centre of the core were measured by a thermocouple. The duration of the recording selected for analysis was about 30 min.

The auto-correlation function of the inlet temperature is shown in Fig. 5 in which the circled points are the measured values obtained by the equipment.

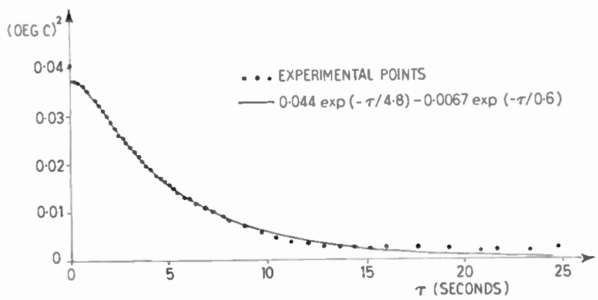


Fig. 5. Auto-correlation function of inlet thermocouple noise.

It is seen to be well fitted by the full line which is the analytic expression.

$$X_{\phi\phi} = 0{\cdot}044\exp(-\tau/4{\cdot}8) - 0{\cdot}0067\exp(-\tau/0{\cdot}6)$$

The auto-correlation function shows two time-constants 4·8 and 0·6 seconds. The former can be attributed to the thermocouple which is mounted in a stainless steel sheath. The cause of the latter time-constant is not known. It can readily be shown that the auto-correlation function corresponds to white temperature noise of mean square amplitude ($\overline{\theta^2}$) 0·24 (deg C)^2 passed through a lag of 0·6 second and measured by a thermocouple of time constant (T_θ) 4·8 seconds.

A typical auto-correlation function of reactor power noise is shown in Fig. 6. The value at $\tau = 0$ gives the mean square fluctuation about the mean (equation (4)) and shows that the r.m.s. fluctuation was 0·46%. Evidence of a decaying oscillation can be seen, its frequency being 6 c/s and its mean square amplitude about one tenth of the total mean square noise. A detailed model[3] of the response of reactivity to inlet temperature changes, denoted by $b(t)$ in our previous discussion, shows it to contain a damped resonance whose frequency is proportional to flow. In our model of the reactor power noise, from equation (36),

$$X_{vv} = b' \star b \star g' \star g \star X_{\theta\theta} + X_{nn}$$

where X_{nn} is the component of reactor power noise uncorrelated with inlet temperature fluctuations. It is apparent that if the inlet temperature noise is 'white', the power auto-correlation function will contain a component corresponding to the damped resonance in $b(t)$. This is observed and moreover measurements at different flow rates show that the frequency is proportional to flow thus helping to confirm the model. It will also be noted that the auto-correlation function contains a term of very short correlation period on which it is not possible to obtain any accurate information since the minimum delay provided by the correlator is 0·05 second. The 'tail' of the auto-correlation function can be fitted by an exponential of time-constant about 50 seconds. Referring to equation (35) we can ascribe this to the root s_2 so that

$$1/s_2 = (\beta + \alpha P)/\lambda\alpha P = 50$$

Writing $\beta = 7\times10^{-3}$ and $\lambda = 0{\cdot}08$ we obtain $\alpha P = 1{\cdot}8\times10^{-3}$. Since the mean power P was 60 MW, the prompt negative feed-back is about 3×10^{-5} per MW which should be compared with a measurement by other means of $3{\cdot}5\times10^{-5}$ per MW.

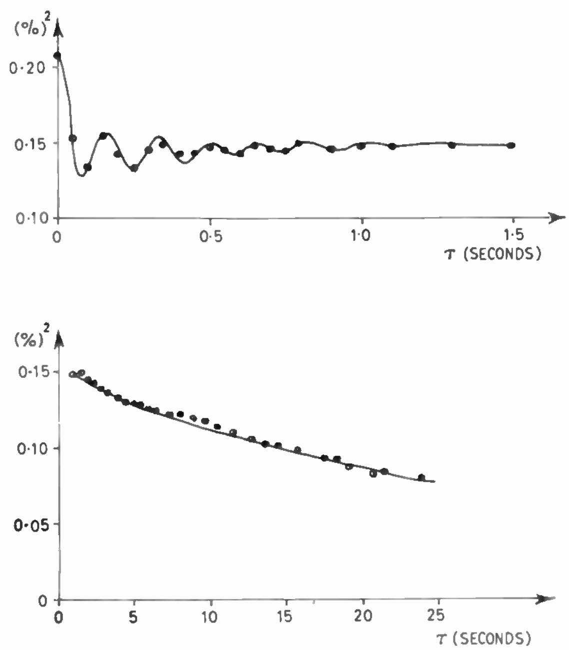


Fig. 6. Auto-correlation function of reactor power noise.

An attempt has been made to measure the reactor transfer function more directly by cross-correlating the inlet temperature fluctuations with power fluctuations. This presents practical problems since in the reactor there is only one thermocouple measuring inlet temperature fluctuations and it is unlikely that the temperature fluctuations due to turbulent mixing of the inlet coolant from many channels will be well correlated over the whole of the reactor core. The net reactivity fluctuation will be the sum of a number of noise sources of which we are measuring one with our thermocouple. It is to be expected that the amount of reactor noise correlated with the noise in the indications of this thermocouple will be small compared with the total reactor noise. A convenient method of estimating this is the 'coherence factor' defined by $X_{\phi v}(0)/[X_{\phi\phi}(0).X_{vv}(0)]^{\frac{1}{2}}$. Our measurements show that only 4% of the reactor power noise is correlated with the noise in this thermocouple. We are thus attempting to measure the correlation between an input noise

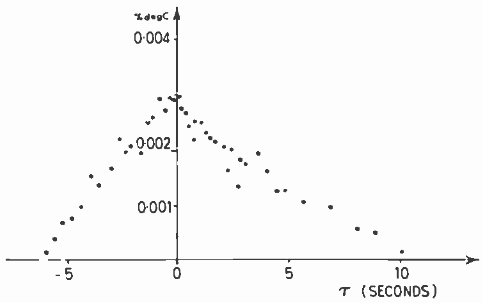


Fig. 7. Cross-correlation function of reactor power and inlet thermocouple noise.

source and the output noise in the presence of a considerable amount of uncorrelated noise. We have seen that this should be possible in theory, but in practice there will be errors due to the finite recording time and to the fact that to keep the total noise signal within the dynamic range of the equipment the required noise signal will be little greater than the instrumental errors. A typical result is shown in Fig. 7 where it will be seen that the scatter is relatively large and it is difficult to obtain any quantitative information. One result is significant. It will be noticed that the correlation is positive, that is, an increase in temperature of the inlet coolant gives rise to an increase in reactor power. It is known from steady state measurements on the reactor that an increase in inlet coolant temperature over the whole of the reactor core produces a decrease in power. The thermocouple measures the temperature of the coolant entering the centre of the core so that there appears to be a positive temperature coefficient at the core centre but an overall negative temperature coefficient. A quantitative assessment suggests that the positive coefficient is an order of magnitude smaller than the negative coefficient. It would be useful to carry out similar measurements with thermocouples placed at the coolant inlet to other parts of the core but this has not yet proved possible.

6. Discussion

The determination of the kinetics of a reactor system is important to its safety. A model of the power feed-back can usually be postulated but there will inevitably be uncertainty in the knowledge of some of the parameters and it is important to check the model experimentally. Some power reactors are equipped with apparatus to produce known disturbances, usually sinusoids, to measure the overall reactor transfer function but this inevitably disturbs the steady operation of the reactor and it is not generally acceptable that it can be used to keep a constant check on the reactor kinetics. There are, however, always random fluctuations in the reactor which could be monitored for this purpose, and the present apparatus was built as a research tool to investigate what parameters can be monitored to give useful information on the kinetics. It is already apparent that some information can be gained from the power auto-correlation function. More valuable information could be obtained in theory from the use of cross-correlation functions but at present the instrumentation in the reactor is insufficient to provide useful results. Even when sufficient instrumentation is available it is apparent that the interpretation of the measurements is a complex procedure. However, the prize is a rewarding one since a continuous monitor of the reactor kinetics which does not disturb its operation would be a valuable addition to its safety instrumentation.

The range of correlation times required is relatively large. Thermal effects in the reactor have time-constants as low as 1/10th second whereas an important time-constant in the reactor transfer function is of the order of 50 seconds. Our equipment covers the faster time-constant adequately but is not really suitable for the longer time-constant. Its range could be extended readily to about 50 seconds by using the slowest recording speed on the tape recording equipment. Alternatively, it is considered feasible to extend the available range by using more than two tracks on the drum. In addition, a re-recording technique can be used. The original signals are recorded at 15/16 in/s and played back at 15 in/s. The signal is demodulated, re-recorded at 15/16 in/s and then analysed in the usual way. This increases the maximum correlation time to 400 seconds but also increases the smallest increment of delay to 0·8 second. The technique has been employed with limited success to investigate the longer time-constant components in the correlation functions.

It will be appreciated that the model of the reactor system which we have discussed has been simplified in at least two important aspects. We have used one delayed neutron group instead of the six usually assumed and we have assumed the temperature feed-back to be rapid. In practice, apart from the 'fast' feed-back due to changes in temperature of the fuel, there will be 'slow' feed-back terms due to heating of the whole reactor structure, as well as the slow feed-back through the heat exchangers to the inlet temperature. The mean time of transit of coolant through the heat exchangers is about 45 sec and an attempt is being made to study the heat exchanger feed-back by the re-recording technique. Work is also proceeding on refining the feed-back model of the reactor but it is not appropriate to discuss it here.

So far, relatively few initial measurements have been carried out but a number of factors emerge. The expected high correlation between adjacent points on a correlation curve has been realized. It has proved possible to detect and measure small resonances in the presence of considerably greater noise and this could be useful in detecting an incipient instability in the reactor system. It also appears possible to obtain some information from the cross-correlation function of two parameters in the presence of twenty times as much uncorrelated noise. Although a great deal more experience is required, our experience suggests that an on-line monitor of the reactor kinetics using noise analysis techniques is feasible.

7. Acknowledgments

The authors wish to acknowledge the assistance of Messrs. Farrer and Nicholls and of Dr. Roberts (all formerly of Bruce Peebles) in the design and development of the equipment. Acknowledgment is also made to Mr. R. V. Moore, Managing Director, Reactor Group, U.K.A.E.A. for permission to publish.

8. References

1. J. A. de Shong, "Flux, reactivity, steam void and steam-water interlayer noise spectrums in EBWR", *J. Nuclear Engng, Pt. A, Reactor Science*, **10**, pp. 147–56, 1959.
2. C. W. Griffin and R. L. Randall, "At-power, low-frequency, reactor power spectrum measurements", *Nuclear Science and Engineering*, **15**, pp. 131–8, 1963.
3. P. G. Bentley and D. C. Menzies, Private communication.
4. V. Rajagopal, "Determination of reactor transfer functions by statistical correlation methods", *Nuclear Science and Engineering*, **12**, pp. 218–24, 1962.
5. A. Z. Akcasu, "Measurement of noise spectra by Fourier analysis", *J. Appl. Phys.*, **32**, No. 4, pp. 565–8, 1961.
6. C. Sastre, "Reactor transfer functions". "Advances in Nuclear Science and Technology", Vol. 2 (Academic Press, New York, 1964).
7. R. B. Blackman and J. W. Tukey, "Measurement of power spectra", *Bell Syst. Tech. J.*, **37**, No. 2, pp. 485–569, 1958.
8. S. O. Rice, "Mathematical analysis of random noise" in "Noise and Stochastic Processes", edited by Nelson Wax (Dover Publications, New York, 1954).
9. M. S. Bartlett, "Stochastic Processes" (Cambridge University Press, 1956).
10. M. G. Kendall and A. Stuart, "The Advanced Theory of Statistics", Vol. I (Griffin, London, 1958).

9. Appendix 1: Fundamental Relations

Let $U(\omega)$ be the Fourier transform of $u(t)$ over the interval $t = -\frac{1}{2}T$ to $t = +\frac{1}{2}T$, defined by

$$U(\omega) = \int_{-\frac{1}{2}T}^{\frac{1}{2}T} u(t)\exp(-j\omega t)\,dt \qquad \ldots\ldots(43)$$

Let $A(\omega)$ be the Fourier transform of the impulse response $a(t)$, defined by

$$A(\omega) = \int_{-\infty}^{\infty} a(t)\exp(-j\omega t)\,dt \qquad \ldots\ldots(44)$$

Then, if the output is denoted by $v(t)$,

$$V(\omega) = A(\omega)U(\omega) \qquad \ldots\ldots(45)$$

Define the power spectrum $P_{uu}(\omega)$ of $u(t)$ as

$$P_{uu}(\omega) = \lim_{T\to\infty} [U(\omega)U^*(\omega)/T] \qquad \ldots\ldots(46)$$

where the superscript star denotes the conjugate complex.

The power spectrum $P_{vv}(\omega)$ of $v(t)$ is then

$$\begin{aligned} P_{vv}(\omega) &= \lim_{T\to\infty} [A(\omega)A^*(\omega)U(\omega)U^*(\omega)/T] \\ &= A(\omega)A^*(\omega) \lim_{T\to\infty} [U(\omega)U^*(\omega)/T] \\ &= A(\omega)A^*(\omega)P_{uu}(\omega) \\ &= |A(\omega)|^2 P_{uu}(\omega) \qquad \ldots\ldots(47) \end{aligned}$$

Now Rice[8] shows that the auto-correlation function and the power spectrum are Fourier pairs, i.e.

$$X_{uu}(\tau) = \int_{-\infty}^{\infty} P_{uu}(\omega)\exp(j\omega\tau)\,d\tau \qquad \ldots\ldots(48)$$

Remembering also that $[A(\omega), a(t)]$ and $[A^*(\omega), a(-t)]$ are also Fourier pairs and applying the convolution theorem to equation (47) we have

$$X_{vv} = a\star a'\star X_{uu} \qquad \ldots\ldots(49)$$

where $a'(t) = a(-t)$, and the star denotes the convolution process.

10. Appendix 2: Effect of Finite Samples

Consider a stationary system in which the output signal has a random fluctuation $v(t)$ about the mean. Consider also a large number of samples of equal length T from which we obtain estimates $X_i(\tau)$ of the auto-correlation function $\overline{X}(\tau)$ corresponding to an infinitely long sample. Bartlett[9] shows that the mean of all such estimates $X_i(\tau)$ at any value of τ is $\overline{X}(\tau)$. Also from Bartlett it can be deduced that the mean square deviation in the values of $X_i(\tau)$ is

$$[X_i(\tau)-\overline{X}(\tau)]^2 = \frac{1}{T}\int_{-\infty}^{\infty} [(\overline{X}(x))^2 + \overline{X}(x-\tau).\overline{X}(x+\tau)]\,dx \quad (50)$$

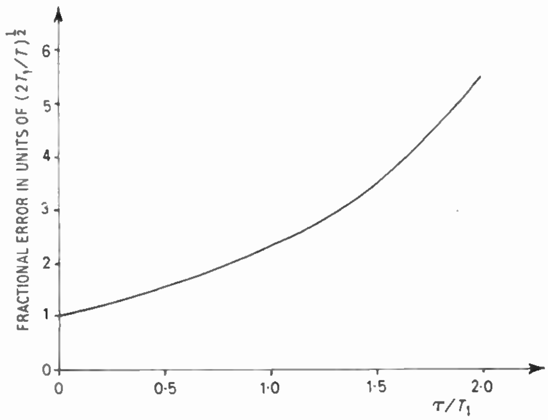


Fig. 8. Statistical error in measurement of an exponential auto-correlation function due to finite noise sample.

which is seen to depend not only on the sample length T but also on the auto-correlation function $(\bar{X}\tau)$.

Consider now the auto-correlation function

$$\bar{X}(\tau) = \sigma^2 \exp(-|\tau|/T_1) \qquad \ldots\ldots(51)$$

a form often encountered in noise analysis. Substitution into equation (50) gives the mean square error in the estimates of $X(\tau)$ as

$$[T_1 + (T_1 + 2\tau)\exp(-2\tau/T_1)]\sigma^2/T \qquad \ldots\ldots(52)$$

Hence the fractional r.m.s. error is given by

$$\text{r.m.s. error/mean} = [T_1 + (T_1+2\tau)\exp(-2\tau/T_1)]^{\frac{1}{2}} \exp(\tau/T_1)/\sqrt{(T)} \qquad \ldots\ldots(53)$$

This is least when $\tau = 0$ when its value is $\sqrt{(2T_1/T)}$ and is approximately twice this when $\tau = 0{\cdot}8\ T_1$ (see Fig. 8).

It must not be assumed, however, that this represents the scatter to be expected on the points of the measured auto-correlation function since there is, in fact, good correlation between the errors in the estimates of $X(\tau)$ at different values of τ from one sample. Let x_1 and y_1 denote the errors in the estimates of $X(\tau)$ and $X(\tau+s)$ in the ith sample, i.e.

$$x_i = X_i(\tau) - \bar{X}(\tau)$$
$$y_i = X_i(\tau+s) - \bar{X}(\tau+s) \qquad \ldots\ldots(54)$$

The means of x_i and y_i are zero and Kendall[10] shows that the mean correlation (p_{xy}) between the errors x_i and y_i is given by

$$p_{xy} = \text{cov}(x,y)/[\text{var}(x).\text{var}(y)]^{\frac{1}{2}} \qquad \ldots\ldots(55)$$

where cov(x,y) is the co-variance or first product moment of x and y, and var(x) is the variance of x. Now Bartlett[9] shows that

$$\text{cov}(x,y) = \frac{1}{T}\int_{-\infty}^{\infty} [\bar{X}(x).\bar{X}(x+s) + \bar{X}(x-\tau).\bar{X}(x+\tau+s)]\,dx \qquad (56)$$

Applying these formulae to the auto-correlation function of equation (51), assuming both s and τ to be positive, gives

$$p_{xy} = \frac{[T_1 + s + (T_1 + 2\tau + s)\exp(-2\tau/T_1)]\exp(-s/T_1)}{[T_1 + (T_1+2\tau)\exp(-2\tau/T_1)]^{\frac{1}{2}} \times [T_1 + (T_1 + 2\tau + 2s)\exp(-2(\tau+s)/T_1)]^{\frac{1}{2}}} \qquad \ldots\ldots(57)$$

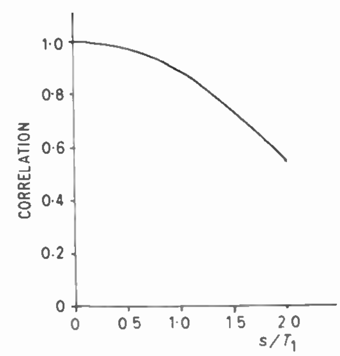


Fig. 9. Correlation of statistical errors between points on a measured auto-correlation function.

If we take the first of the two points on the auto-correlation function to be near the origin, i.e. $\tau \ll T_1$, this reduces to:

$$p_{xy} = \frac{\sqrt{2}(1 + s/T_1)\exp(-s/T_1)}{[1 + (1 + 2s/T_1)\exp(-2s/T_1)]^{\frac{1}{2}}} \qquad \ldots(58)$$

Figure 9 shows that the correlation is high, better than 0·9, if s is less than time-constant T_1, and is still better than 0·7 if s is less than $1{\cdot}5\ T_1$.

Manuscript first received by the Institution on 16th September 1964 and in final form on 16th February 1965. (Paper No. *1001.*)

U.D.C. 621.317.7:621.397.6

The Performance Requirements of a Television Monitor Receiver (Nyquist demodulator) and Methods of Measurement

By

F. G. JOHANNESEN†

Summary: A high-quality monitor receiver is normally used at television transmitter stations for measuring and monitoring purposes. The performance of this receiver is discussed on the basis of the ideal characteristics, as specified by C.C.I.R., and the characteristics which lie within practical tolerance limits. The problem of testing the receiver is considered, and selected measuring procedures, which have been used in an extensive test service for the Norwegian Telecommunication Administration, are described.

1. Introduction

The signals transmitted by television stations are customarily monitored on a high-quality control receiver which is connected to a directional coupler in the transmitter antenna feeder. In case of reflections from the antenna, only the direct wave is delivered to the receiver. A correct assessment of the radiated signal is therefore always possible—provided the characteristics of the control receiver are close enough to the ideal characteristics specified for the actual television system. However, the ideal characteristics (with tolerance limits) are not completely specified internationally—this specially applies to the phase versus frequency characteristics. The C.C.I.R. specifications on phase which do exist[1] have been formulated relatively recently and are in general terms. Two alternatives are given: (1) phase linear transmitter, (2) precompensation in the transmitter not exceeding '*one half* of that necessary to compensate a receiver using normal minimum phase shift networks and with an amplitude characteristic corresponding to the television standard concerned'; this precorrection 'applies only to frequencies between zero and up to approximately half the video bandwidth'. *Full* precompensation in the transmitter of the phase distortion in a model receiver was formerly specified by at least one national telecommunication administration.[2-4] The influence of the phase characteristic on the transient response, and hence picture quality is decisive and to assure a correct assessment of the transmitted signal, the phase characteristic of the control receiver must be accurately known and be in accordance with the intended characteristic. The display unit, which is normally a high-grade video picture monitor, presents no particular problem as its bandwidth is in excess of the nominal bandwidth of the system by a safe margin. Its amplitude response is allowed to fall off gradually which leads to phase linearity and good transient response. This display unit is therefore not considered further here. The sound section is also left out of consideration, because the performance and ways of testing are well established. The part of the receiver which is then left to be treated comprises the radio frequency circuits and the video detector/amplifier. This part usually forms a constructional unit which is termed 'Nyquist demodulator' or 'vestigial sideband demodulator'.

† Electronics Laboratory, The Engineering Research Foundation of the Technical University of Norway, Trondheim.

2. General Considerations

2.1. *Nyquist Demodulator Design*

In the conventional design the Nyquist demodulator consists basically of an r.f. input stage, a mixer and a crystal controlled local oscillator, an i.f. amplifier, a video detector and a video amplifier. The construction and characteristics are similar to an ordinary television receiver; however, the requirements to be met with regard to performance and stability are greater in the Nyquist demodulator. It is not necessary for the Nyquist demodulator to provide gain because a large signal is available from the directional coupler in the transmitter antenna feeder. At least one type of a completely passive Nyquist demodulator is manufactured which performs the shaping of the amplitude response at r.f. directly (Bands I and III). Conversion to i.f. is then dispensed with. This gives a simplified and rigid design.

2.2. *Testing and Monitoring of the Transmitter*

The Nyquist demodulator is used as part of the system transmitter + receiver for measurement of the characteristics of the transmitter and for monitoring of transmitted signal.

The transmitter parameters normally measured through the Nyquist demodulator are:

(*a*) Video amplitude versus frequency characteristic.

(*b*) Video phase versus frequency characteristic. (Alternatively group delay or phase delay.)

(*c*) Transient response.

(*d*) Linearity.

(*e*) Differential gain and differential phase.

(*f*) Modulation depth.

(*g*) Noise.

These will be considered in turn.

(*a*) The video amplitude/frequency response is measured from the video modulation input of the transmitter to the video output of the demodulator. Either point-by-point or sweep measurements are used. It is sometimes necessary to add line pulses to the test signal to ascertain proper operation of the transmitter clamp circuitry. However, most transmitters have an 'a.c.-test' operation mode in which the frequency characteristics are measured with a continuous sine-wave signal—without sync and blanking pulses.

For a rapid check of the response, a 'burst' signal with frequency bursts of, e.g., 1, 2, 3, 4 and 5 Mc/s, might be used. This signal usually has line sync and blanking pulses added.[5-7]

Before the video response is measured, it should be ascertained that the sideband amplitude/frequency characteristic is within the intended tolerance limits. A 'sideband-response analyser' is then most conveniently used.[8]

(*b*) The video phase versus frequency characteristic is most accurately measured with a precision video frequency phase meter (accuracy approx. $\pm 1°$). High accuracy of measurements is necessary when equalizers should be matched to a transmitter.

A less tedious procedure is to use a sweep group delay meter which gives a display of the group delay/frequency characteristic on a cathode-ray tube. However, the accuracy is seldom better than 10 nanoseconds, and it is usually not possible to measure below approximately 0·5 Mc/s video frequency.

(*c*) To measure the transient response a square wave and/or impulse-type test signal is fed to the video input of the transmitter. A wide-band oscilloscope, connected to the demodulator output, displays the transient response. The versatile 'pulse-and-bar' test signal is used to advantage.[9-19]

Amplitude, phase and transient response should be measured at a small modulation index to exclude distortion from the quadrature component inherent in all vestigial sideband systems. Other types of non-linear distortion must also be negligible if these measurements are to have any significance.

(*d*) The linearity is measured as a function of the video level within the specified range of picture signal levels. Usually a saw-tooth or stair-step signal with added line sync and blanking is used. In the more refined procedure a sine-wave of constant frequency and amplitude is superimposed on the saw-tooth or stair-step and filtered off at the output. Non-linearity in the tested system will produce 'level-dependent gain' or 'differential gain' for the superimposed sine-wave. Variations in the amplitude of the filtered-off sine-wave are then a measure of the non-linearity in the system, Poor linearity affects the picture gradation, i.e. the correct reproduction of tonal values.[6, 7, 20, 21]

(*e*) Differential gain is measured in the same way as linearity. The superimposed sine-wave then has the frequency of the chrominance subcarrier of a compatible colour television system.[6, 7, 20-22]

Differential phase is a measure of shift in chrominance subcarrier phase caused by level variations of the luminance signal. Special measuring instruments have been devised.[6, 7, 20-22]

The differential gain and differential phase of a Nyquist demodulator of conventional design are considerable. It might therefore be preferable to use a linear diode detector for these measurements.[8]

Low values of differential gain and differential phase are important for faithful reproduction of colours in compatible colour television systems.

(*f*) For television systems with negative modulation a definite minimum carrier level is assigned to 'white'. This 'white level', usually 10%, should be maintained to ascertain proper operation of inter-carrier sound receivers. For adjustment of maximum modulation depth and the 'white clipper' of the transmitter, the Nyquist demodulator could be used provided its linearity is sufficient. Alternatively a linear diode detector (operated at high signal level) might be used. Recently it has become customary to use a h.f. oscilloscope to display the modulation envelope at the transmitter output. The linearity and modulation depth can then be assessed without the implication of non-linearity in a demodulating device.

(*g*) The signal/noise ratio of the transmitted signal is also measured at the output of the Nyquist demodulator. The noise level of the demodulator should therefore be as small as possible.

The Nyquist demodulator is furthermore required to deliver the demodulated video signal to a picture monitor for continuous monitoring of the outgoing programme. In case of compatible colour transmissions, both black-and-white and colour picture monitors can be used.

2.3. *The Ideal Transmitter Receiver Characteristics*

Measurement of transmitter characteristics through the Nyquist demodulator, as previously described, obviously requires the demodulator characteristics to be known. These characteristics should preferably be such as not to make necessary undue corrections of the measuring results on a complicated theoretical basis.

For accurate results the Nyquist demodulator characteristics should be determined using a vestigial sideband transmitter which is in accordance with standard specifications. However, such a transmitter is a complex device, which, in turn, relies upon the Nyquist demodulator for correct alignment. The Nyquist demodulator should therefore be tested by simpler means.

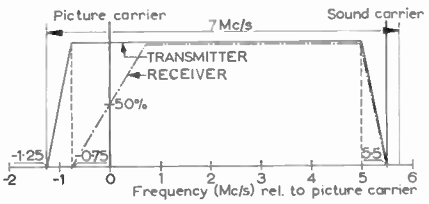


Fig. 1. Ideal amplitude vs. frequency characteristics of transmitter and receiver as specified by C.C.I.R. (European 625-line system). The equivalent low-pass system has constant amplitude and linear phase within the nominal video frequency band.

The questions of preferable characteristics and test methods of the Nyquist demodulator can partly be answered by considering the ideal system characteristics as specified by the C.C.I.R. Figure 1 shows the ideal amplitude versus frequency characteristics for the 625-line system with 5 Mc/s nominal video bandwidth. The overall video phase versus frequency characteristic is linear.

From Fig. 1 it may be concluded that part of the lower sideband, below $-1{\cdot}25$ Mc/s relative to picture carrier, is not transmitted. The lower sideband maintains its full amplitude only down to $-0{\cdot}75$ Mc/s from picture carrier. If a double sideband demodulator is used for demodulation of the transmitted signal, the correct balance between the amplitudes of the modulating frequency components will not be maintained in the demodulated output signal, unless a special amplitude correction is provided. This amplitude correction would have to increase the gain 6 dB for video frequencies higher than 1·25 Mc/s, relative to the response below 0·75 Mc/s, in order to compensate for the attenuation of the lower sideband. A television receiver which applies the principle of double sideband demodulation and subsequent amplitude correction, has been proposed by Ruston.[23]

The correct amplitude balance is also obtained if the receiver r.f. characteristic is shaped as shown in Fig. 1. The picture carrier is positioned halfway up on the sloping part (the 'Nyquist flank') of the response. It can be deduced from the figure that the sum of both sidebands is constant within the nominal video band, 0–5 Mc/s. Hence, the overall video amplitude/frequency characteristic of the system of transmitter and receiver will also be constant within this frequency band.

Because the Nyquist flank gradually suppresses the lower sideband, amplitude- and phase distortion of this sideband will have a relatively small influence on the overall response. In this respect a receiver with a Nyquist flank is preferable because deficiencies of the present technique of transmitter phase correction lead to large phase distortion of the lower sideband (Sect. 2.4). For this and other reasons the Nyquist principle is generally applied in domestic receiver design and should therefore also be used for transmitter control demodulators. However, if the phase properties of the transmitter are improved, the less critical Ruston principle might be attractive to receiver designers.[24]

All vestigial sideband demodulators at present commercially available apply a Nyquist flank and are generally referred to as 'Nyquist demodulators'. Only such demodulators will be assumed hereafter.

The overall system is identical to 'the ideal low-pass system' having constant amplitude, linear phase and sharp cut-off. The response of the ideal low-pass system to a sine-squared impulse of half-amplitude duration (h.a.d.) $T = 1/2f_c$ (f_c = cut-off frequency of system), is shown in Fig. 2. This ideal response should be kept in mind as the performance target of the system.

It can be concluded from Fig. 1 that the performance of the system will be the same if the lower sideband of the transmitter is *not* attenuated (provided the phase characteristics are the same), because the

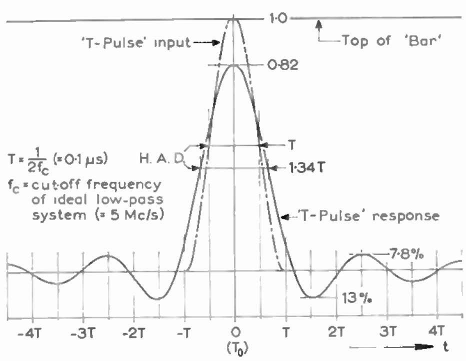


Fig. 2. The response of the ideal low-pass system to a sine-squared impulse of half-amplitude duration $T = \frac{1}{2f_c}$ (f_c = cut-off frequency of system). This ideal response can be taken as the performance target of the practical system.

receiver r.f. characteristic excludes the lower sideband below −0·75 Mc/s, referred to picture carrier.

If the receiver (or Nyquist demodulator) sufficiently limits the lower sideband, it is therefore correct to use a double sideband signal generator for measurement of its characteristics.

It now remains to consider the practical characteristics of both transmitter and receiver to include possible corrections of the measured results.

2.4. *Practical Characteristics of Transmitter and Receiver*

We shall confine ourselves to treat the European 625-line system because phase- and quadrature-distortion is inherently more predominant in this system due to the large bandwidth of the main sideband relative to the width of the vestigial sideband.[25] However, the conclusions to be drawn would largely be the same for other systems.

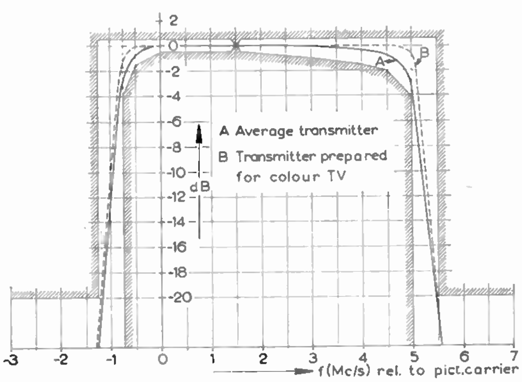


Fig. 3. Practical amplitude vs. frequency characteristics of an average transmitter (A) and a transmitter prepared for compatible colour television transmissions (B). The tolerance limits shown are widely used, although laid down as a national standard.[26]

Figure 3, A, shows the sideband characteristic of an average transmitter. The lower sideband falls off to about −3 dB at −0·75 Mc/s relative to picture carrier, as compared with the ideal response (Fig. 1) which is flat in this region. This loss in amplitude is not so harmful, because the receiver response (Nyquist flank) gradually excludes the lower sideband (Fig. 1). The upper sideband also rolls off and is down by more than 3 dB at the nominal video cut-off frequency 5 Mc/s.

The reduction of picture resolution, caused by the loss of high video frequency components, will probably be outweighed by an improved transient response.

For compatible colour television, however, the upper part of the video band is used for transmission of additional colour information (saturation and hue). The chrominance signal is transmitted on a subcarrier at 4·43 Mc/s video frequency, and its upper sideband frequencies extend to 5 Mc/s. If the upper sideband frequencies are unduly attenuated by the transmitter, cross-talk between colour information components might result which, in turn, leads to colour fringed transitions. Modern transmitters designed to meet the requirements of future compatible colour television therefore have a more square-shaped sideband characteristic—similar to Fig. 3, B. The tolerance limits shown in Fig. 3 are widely used in European countries although specified as a national standard.[26] It has been proposed to tighten the lower limit in case of colour transmissions.[27]

Figure 4, A, shows the group delay versus frequency characteristic of the transmitter with the amplitude characteristic of Fig. 3, A. The group delay is seen to decrease gradually from picture carrier to mid-band by approximately 150 nanoseconds. The effects on the transient response, and hence the picture, are the typical 'preshoot' and 'smear'—which are rather serious degradations. The increase in group delay towards the upper channel edge causes, on the transient response, unsymmetry and enlarged amplitudes of the 'Gibbs ringing' inherently associated with sharp high-frequency cut-off. These deteriorating effects might be fully compensated by use of phase correcting networks.

For high-level modulated transmitters, it is considered impracticable to apply phase correction in the r.f. path. The phase equalizer is therefore inserted ahead of the video modulation amplifier. Because the inherent group delay characteristic of the transmitter

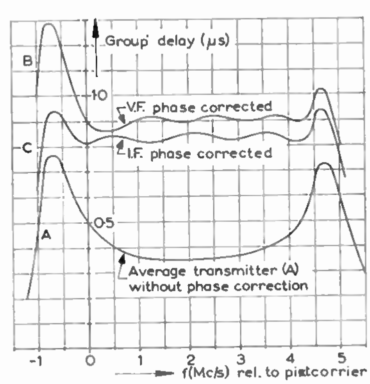


Fig. 4. Group delay vs. frequency characteristic of the transmitter with the amplitude characteristic of Fig. 3, A, with no phase correction applied (A). V.f. phase correction, generally used today, leads to increased distortion of the lower sideband (B). For i.f. modulated transmitters, i.f. phase correction could be applied. Both sidebands might then be equalized (C).

is skew-symmetrical with respect to the picture carrier (Fig. 4, A), this skew-symmetry will be *increased* by video frequency phase correction—as is seen in Fig. 4, B. Figure 4, B, shows the best result which could be obtained using a three-section phase equalizer in the video path. The upper sideband is fully equalized. (Residual delay variations are probably within the tolerances of the control and measuring equipment.) The lower sideband, however, is seen to be even more phase-distorted. The effect of this is mainly to produce a notch in the overall video amplitude characteristic at about 0·75 Mc/s (Fig. 7, a).

A new and interesting concept in transmitter design is modulation on an intermediate frequency with subsequent frequency conversion and amplification in the desired r.f. band.[28] The vestigial sideband filter is built for the i.f. band, which permits lumped elements to be used. The phase distortion might then be ideally corrected after the vestigial sideband filter. Figure 4, C, shows the result which could be obtained with i.f. phase correction using a three-section equalizer. Full correction of both sidebands is possible by increasing the number of equalizer sections.

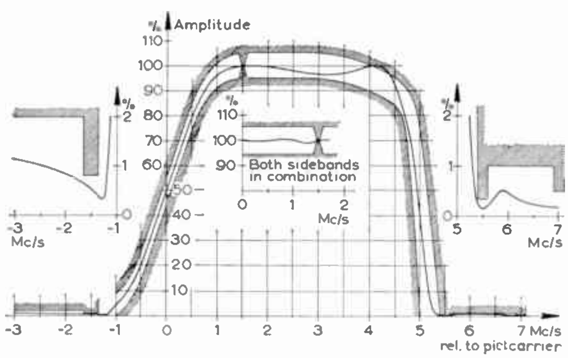


Fig. 5. Measured amplitude vs. frequency characteristics of a Nyquist demodulator which meets essentially the tolerance limits specified by A.R.D.,[31] also shown.†

The characteristics of the *receiver* should preferably be such as to need few poles and zeros for realization, which means few tuned circuits—and, consequently, lower costs and better stability. It was originally thought that only 'minimum phase-shift networks' would be practicable in the domestic receiver. According to the minimum phase-shift law,[29] the phase characteristics would then be similar for receivers having the same standard amplitude response. On that ground, it was proposed to precorrect at the transmitter the phase distortion of the 'average receiver'.[30] Full precompensation was formerly specified by at least one national administration.[4] However, according to the C.C.I.R.,[1] a maximum of half the precorrection might now be used—or alternatively, no precorrection at all. Figure 5 shows measured amplitude characteristics of a Nyquist demodulator, which lie essentially within the tolerance limits specified by A.R.D.[31] Only minimum phase-shift networks are used in this demodulator; the measured phase characteristic Fig. 6, A is then uniquely determined by the amplitude response (Fig. 5). The group delay, derived from the measured phase characteristic by numerical derivation ($\tau_g = \frac{d\phi}{d\omega}$),

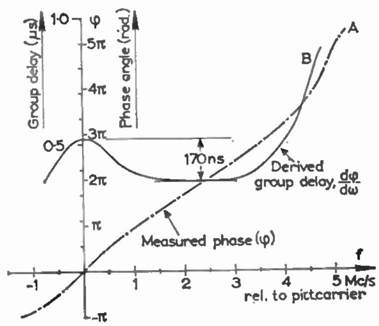


Fig. 6. Measured phase vs. frequency characteristic (A) and derived group delay (B) of the Nyquist demodulator with the amplitude response of Fig. 5. These characteristics are uniquely determined by the amplitude response, because only 'minimum phase-shift networks' are used in this demodulator.

is shown in Fig. 6, B. The delay characteristic of the demodulator is similar to the delay characteristic of the uncorrected transmitter, Fig. 4, A. If the delay is not equalized, serious preshoot and smear will affect the transient response and the picture reproduction (Sect. 3.7).

In modern receiver design 'non-minimum phase-shift networks' have also come into use, known as 'infinite attenuation networks'.[32] The phase characteristic is therefore not entirely decided by the amplitude response; this is more so when stray couplings are present, as is often the case in domestic receivers. It is therefore inevitably necessary to specify a definite phase or group delay characteristic for the receiver. This ideal characteristic might either be phase linear or similar to Fig. 6, B, but with a maximum of half the delay variations (e.g. 90 ns instead of 170 ns in Fig. 6, B).[33] In the present state of the art it is necessary to shape the Nyquist flank of the receiver closer to the 'Gaussian curve'[34] to keep within the C.C.I.R. phase recommendations. The Nyquist flank will then not be strictly symmetrical, and the upper part will fall outside the tolerance limits of Fig. 5. The receiver

† A.R.D. = Arbeitsgemeinschaft der Rundfunkanstalten der Bundesrepublik Deutschland.

designer is then left to find the best compromise between amplitude, phase and transient response. However, what is gained by specifying a definite ideal phase characteristic for the receiver, is probably more uniform phase properties and transient response between receivers of different manufacture. In addition, future design is not tied to past compromises.

The Nyquist demodulator need not be a compromise solution—because the additional costs of adhering closely to ideal characteristics only amount to a very small percentage of the total transmitter installation costs. Furthermore, the quality of transmission is seen to be largely dependent on the Nyquist demodulator performance.

Nyquist demodulators designed to meet C.C.I.R. phase specifications all use phase equalizing networks—which is considered unacceptable in domestic receiver design. Some demodulators are designed with a sharp 'sound notch', leaving considerable residual response for upper adjacent channel (Fig. 13). These trends break an early concept that the Nyquist demodulator should simulate the characteristics of a domestic 'model receiver'.[3, 4, 30] The Nyquist demodulator should now be looked upon as a precision control instrument to be used for measurement and assessment of the specific parameters of the transmitter.

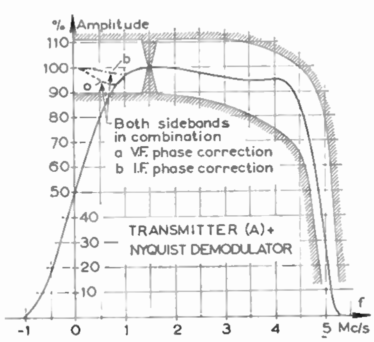


Fig. 7. Combined amplitude vs. frequency characteristics of transmitter (Fig. 3, A) and Nyquist demodulator (Fig. 5). The influence of the alternative transmitter delay characteristics, Fig. 4, B and C, on the overall video response is shown by curves a and b respectively.

The combined characteristics of transmitter and Nyquist demodulator remain to be considered. Figure 7 shows the combined amplitude response of the average transmitter (Fig. 3, A) and Nyquist demodulator (Fig. 5). The Nyquist flank is seen to be practically not influenced by the transmitter lower sideband attenuation. The performance of the Nyquist demodulator would therefore be the same for a double sideband laboratory transmitter as for the standard vestigial sideband transmitter—provided the phase relationships are the same. The true double sideband transmitter possesses equal delays for both sidebands. Equal sideband delays might also be achieved for i.f. phase corrected transmitters, as previously mentioned. Figure 7, a and b, shows the overall video amplitude response presuming the transmitter delay characteristics of Fig. 4, B and C respectively. Video frequency phase correction is seen to produce a dip in the video response at approximately 0·75 Mc/s, due to the non-symmetry of the sideband vectors relative to the carrier vector. This dip might be amplitude corrected in the transmitter; its influence on the transient response seems to be negligible. However, for amplitude measurements on the transmitter, the reason for the dip should be recognized. For both types of phase correction the amplitude response is seen to be well within the tolerance limits for the system.[26]

2.5. *Input Parameters; Influence of Quadrature Distortion*

The Nyquist demodulator is normally to be fed via a coaxial cable of specified characteristic impedance, usually 50, 60 or 75 ohm. Conventional Nyquist demodulators are designed to have almost constant, resistive, input impedance within the frequency range of a channel. The reflection coefficient at the input should preferably be lower than 1·03 to avoid errors caused by standing waves.

The passive Nyquist demodulator usually has an input impedance which varies considerably, both in magnitude and phase angle, with frequency. A nominal source impedance is specified, and if source and cable are properly matched, no faults due to standing waves will occur. However, the source e.m.f. must not be influenced by the varying load.

The negative excursions of the modulated picture carrier normally fed to the Nyquist demodulator, is shown in Fig. 8 (A) (C.C.I.R. 625 line-system). Maximum carrier level occurs during synchronizing pulses (100%), and blanking level is fixed to 75%. The picture modulation levels are assigned to the range 70% (black) to 10% (white).

A special sort of non-linear distortion, inherent in the single sideband operation, needs to be considered. The well-known vectorial representation of single sideband modulation is shown in Fig. 8 (B). The resultant carrier vector v varies both in amplitude and phase. This vector may be resolved into an in-phase and a quadrature component. Due to the quadrature component, the envelope of modulated carrier is no longer a replica of the modulating signal. Normal amplitude demodulators respond to the envelope of the carrier ('envelope detector'), and the quadrature distortion will therefore also be present in the demodulated signal. The distortion in case of sine-wave modulation with modulation index $m = 1$ is depicted in Fig. 8 (C).

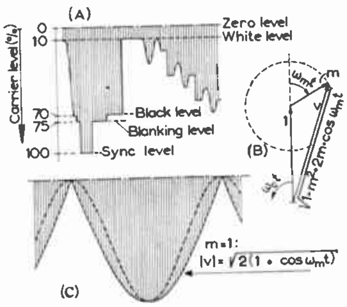


Fig. 8.

(A) Modulation envelope of the r.f. signal (C.C.I.R. 625-line system).

(B) Single-sideband modulation gives rise to a quadrature component which distorts the modulation envelope.

(C) Envelope distortion for modulation index $m = 1$.

It is evident from Fig. 8, B that the quadrature distortion increases with the degree of modulation. The modulation index for a black-to-white transition (Fig. 8, A) is 0·75, but decreases rapidly for decreasing modulation depth, being 0·17 for a transition 50/70%. The modulation factors for the sine-waves near white, at grey and near black, shown in Fig. 8 (A), is 0·5, 0·25 and 0·17 respectively; operation near black therefore gives little quadrature distortion. For accurate measurements of amplitude characteristics, a modulation index lower than 0·1 must be ascertained.

It is possible to reduce quadrature distortion by the principle of increasing the carrier amplitude before demodulation.[24] Although this principle has proved very satisfactory, it is considered too expensive for domestic receivers.

For Nyquist demodulators, the extra costs should not be prohibitive. In fact, 'exalted carrier diode detection' has recently been applied in a Nyquist demodulator with great success. It was demonstrated that frequency characteristics and transient response could be correctly assessed, regardless of modulation depth, and also that linearity near 'white' was improved.[35]

3. Measurement of Nyquist Demodulator Performance

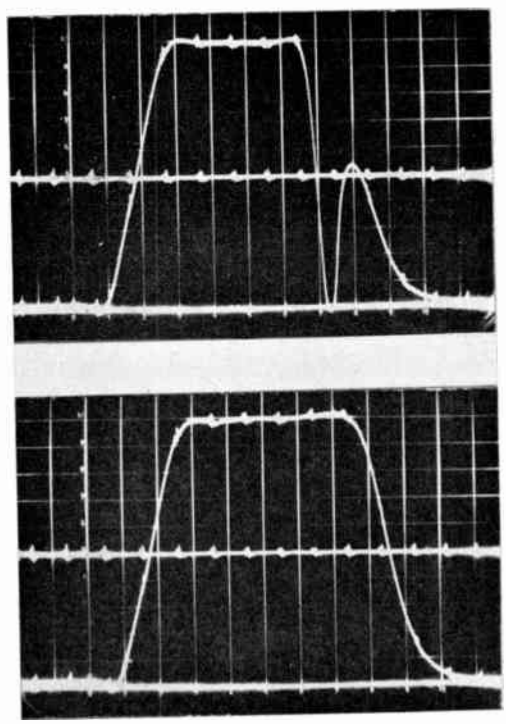
The methods of testing the Nyquist demodulator will now be described. Only high-quality laboratory instruments should be used, and it is essential that reflections from mismatch be kept as low as possible. The standing wave ratio at the input of the demodulator should preferably be lower than 1·02 within the frequency range of the channel. This could normally be obtained by proper use of precision attenuator pads at reflecting connections.

Sufficient warm-up time of equipment should be allowed to avoid drift during measurement. Besides, it might be necessary when measuring frequency characteristics point-by-point, to repeat the measurement and correct for any drift and instability which might occur during the time of the measurement.

The tolerances required in the Nyquist demodulator characteristics are seen to be rather tight, in fact, they almost call for the accuracy of good laboratory instruments. It is therefore advantageous to use different test methods as a control, e.g. the transient response which, when correctly assessed, provides a good control on measured amplitude and phase characteristics.

3.1. *R.F. Amplitude/Frequency Response*

For a rapid check of the r.f. amplitude response a sweep generator is conveniently used. Figure 9 shows



(a)

(b)

Fig. 9. The r.f. amplitude response of a Nyquist demodulator measured with a sweep generator. Horizontal line indicates − 6 dB. Frequency markers 1 Mc/s apart.

(a) 'Sound notch' switched in.

(b) 'Sound notch' switched out.

the result displayed on a cathode-ray tube. The horizontal line indicates −6 dB relative to maximum response. Crystal-controlled marker 'pips', 1 Mc/s apart, are also shown. The Nyquist demodulator tested had a switchable 'sound notch' for attenuation of the signal from the sound transmitter. On Fig. 9 (a), this sound notch is switched in, on Fig. 9 (b), it is switched out.

The non-linearity of the video detector causes a compression of the lower part of the response display which might obscure residual response lobes in the adjacent channel regions. To reduce this non-linearity as much as possible, the sweep signal should be strong enough to make the video detector work over its entire normal dynamic range.

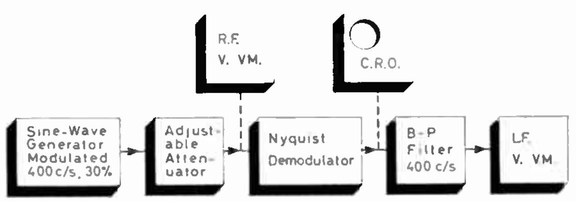


Fig. 10. For accurate measurement of the r.f. response, a point-by-point procedure must be used. The output voltage is kept constant to exclude the non-linear influence of the video detector.

For accurate measurements, especially of the Nyquist flank, a more tedious procedure is used (Fig. 10). A sine-wave generator is modulated with a low frequency, e.g. 400 c/s. The degree of modulation is kept constant, at about 30%. This signal is fed, through an adjustable attenuator, to the input of the Nyquist demodulator. The demodulated output is indicated on an oscilloscope or a valve voltmeter (L.F. V.V.M.). To measure the response, the modulated carrier is shifted within the frequency range of the appropriate channel and the attenuator adjusted for constant output signal from the demodulator. The reading of the attenuator is plotted against frequency to give the attenuation characteristic (which might be reversed to give the amplitude response). Because the level at the video detector is kept constant, its non-linearity will not influence the measurements. The response might therefore be measured accurately within a range of 60 dB or more. For measurements in the adjacent channel regions, where high attenuation is encountered, it might be necessary to use a 400-c/s bandpass filter at the demodulator output to exclude noise (Fig. 10). To avoid overload and 'clipping', it is advisable to control the demodulator output signal on an oscilloscope, as indicated in Fig. 10.

If the adjustable attenuator is not accurately calibrated, a r.f. valve voltmeter (R.F. V.V.M.) might be used to measure the input voltage, provided the input resistance of the demodulator is sufficiently constant (Fig. 10). In the case of frequency-dependent input impedance, normally encountered with passive demodulators, it would not be correct to measure the input voltage. The e.m.f. of the signal source should then be measured. It must be ascertained that the e.m.f. is not influenced by load variations.

3.2. *R.F. Group Delay/Frequency Characteristic*

The r.f. group delay characteristic is measured according to the principle proposed by Nyquist and Brand.[36] A carrier is modulated with a constant frequency, called 'slit' frequency, e.g. 100 kc/s. The modulated carrier is passed through the circuit under test and demodulated. After demodulation, the signal is compared in phase with a fixed reference phase in a phase meter, as shown in Fig. 11.

If the phase characteristic of the circuit being tested can be taken as linear within a region ±100 kc/s relative to the carrier, each sideband of the 'slit' frequency will be shifted in phase by equal but opposite amounts with respect to the carrier. Consequently, the *modulation envelope* will be subject to the same amount of phase shift. The measured phase shift can then be converted to equivalent 'envelope delay'—also called 'group delay'. In the case of 100 kc/s slit frequency, 36 deg of phase shift means a group delay of 1 μs. It follows that the group delay

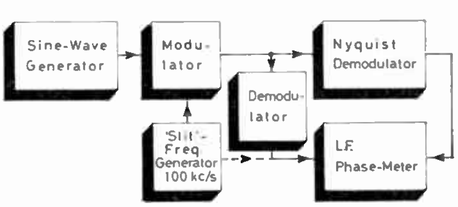


Fig. 11. The r.f. group delay characteristic is measured using the Nyquist and Brand principle. Sweep measurement is possible, although for high accuracy a point-by-point procedure must be applied.

is the slope of the phase characteristic, $\tau_g = d\phi/d\omega$. To measure the group delay characteristic, the carrier frequency is shifted within the appropriate range and the phase shift plotted in terms of equivalent group delay against frequency. It is also possible to display the delay curve on an oscilloscope. A sweep generator is then used to supply the carrier voltage.

It is important to keep the voltage level at the video detector constant during measurement. This means that the input signal level must be adjusted with carrier frequency to compensate for gain variations. This adjustment could easily be done in the point-by-point procedure if a valve voltmeter is connected to the output of the demodulator for level control. In the sweep procedure, automatic gain control must be used. The gain in the modulator (Fig. 11) is then regulated.[37] If the level at the video detector is allowed to vary, the detector will present a varying load to the last i.f. stage, and that will cause level-dependent amplitude and phase variations.

In Fig. 11 it is indicated that the slit frequency reference phase might be taken alternatively from the slit frequency source itself, or from the input of the test object via a (phase linear) demodulator. The last alternative is normally preferred, because possible delay variations in the modulator will then not be included in the measuring results.

The accuracy is usually not better than 10 ns when using sweep measurements. To increase accuracy, the slit frequency must be lowered, but this, in turn, requires a more sensitive phase meter. For an accurate point-by-point measurement, the slit frequency should be as low as 20 kc/s. To measure delay variations in the order of 1 ns, the phase meter would then have to resolve angles as small as 0·01 deg. This accuracy can be reached if the phase meter is given a narrow frequency bandwidth to maintain sufficient signal/noise ratio.

3.3. *Overall Video Amplitude/Frequency Response; Each Sideband Separately*

The measuring procedures which have been described do not include the characteristics of the video detector and the video amplifier, because one fixed modulation frequency was used. To include the influence of the video circuitry (e.g. peaking coils), the *modulation frequency* must be varied within the entire video frequency range. The carrier frequency is then to be held fixed at the nominal picture carrier frequency of the appropriate channel. The carrier frequency tolerance should be better than ±5 kc/s, which means that crystal control is required.

As explained in section 2.3, the Nyquist receiver reacts only to one sideband for higher modulation frequencies. The overall video amplitude/frequency characteristic in the frequency range from 1·25 Mc/s to cut-off, 5·5 Mc/s, or higher, is then measured with a single sideband signal source, as depicted in Fig. 12. The carrier and the single sideband are generated separately and combined in a resistive T-network.

Taking into account the extra 6 dB receiver gain for the sideband relative to the carrier, a modulation index of 0·1 at the video detector corresponds to a carrier/sideband voltage ratio of 26 dB at the receiver input. If possible, a voltage ratio of 30 dB should be used in order to bring the quadrature distortion to a negligible minimum (Sect. 2.5).

The response is normally measured with a carrier voltage corresponding to a black picture. For acceptance tests, the response corresponding to a white picture should also be measured, because the level at the detector might have some influence on the response.

The nominal output voltage from a Nyquist demodulator for a standard r.f. signal (Fig. 8, A) modulated to white, is usually 1 V peak-to-peak. The relation between input carrier level and nominal video output level must be established. One method is to engage the 'zero-line chopper' in the Nyquist demodulator which periodically short-circuits the output and

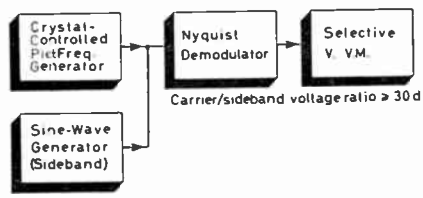


Fig. 12. Overall video amplitude response is most accurately measured using the single sideband method. In this case, possible imperfections of a modulator would not be included in the measured result.

gives a zero reference line on an oscilloscope connected to the output. If only the carrier is applied, a rectangular wave is displayed. The input carrier level corresponding to 1·1 V p–p output signal is noted and assigned as 100% carrier level (cf. Fig. 8, A).

To measure the response, the sideband frequency signal is shifted within the appropriate range and the demodulated output voltage measured with a selective valve voltmeter (Fig. 12). A broad-band voltmeter might be used for in-band measurements, but near cut-off—and especially for measurement of the response corresponding to white picture—a selective voltmeter must be used because of the low signal levels encountered.

The response for video frequencies below 1·25 Mc/s is measured for each sideband separately by shifting the sideband frequency within the range ±1·25 Mc/s relative to the carrier frequency. The response in this range is almost entirely decided by the r.f. response alone, i.e. the Nyquist flank. It is possible to measure the Nyquist flank very accurately by this method. Also the response in the lower adjacent channel range might be measured.

In Fig. 13 is shown a video amplitude/frequency characteristic of a Nyquist demodulator measured in this way. This demodulator is designed also for compatible colour television. To avoid excessive 'colour cross-talk', the response is made as flat as

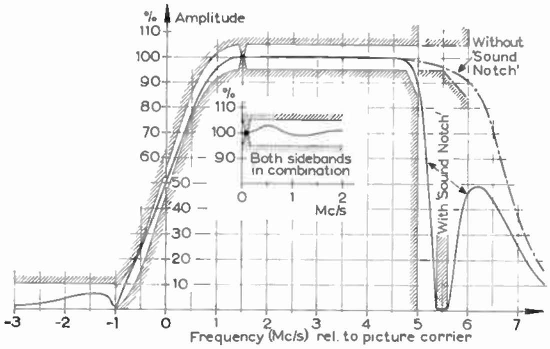


Fig. 13. Measured overall video amplitude/frequency characteristic for a Nyquist demodulator designed also for compatible colour television. For the proposed tolerance limits, see text.

possible in the upper video frequency range where the additional colour information is interlaced.[38]

When a definite phase or group delay characteristic is specified for the Nyquist demodulator—which implies that recourse must be made to phase correcting networks (e.g. 'all-pass networks') to keep within the C.C.I.R. recommendations[1]—it is not necessary to specify the amplitude response *outside* the nominal channel. In fact, no restriction on the amplitude response in the upper adjacent channel is necessary (Fig. 13). In the lower adjacent channel, however, excessive response lobes should not be allowed, in order to make it possible to measure the Nyquist demodulator through a double sideband laboratory generator.

The tolerance limits proposed in Fig. 13 should not be unduly strict. For monochrome television, and in case of PAL compatible colour television,[39] the lower tolerance limit might be relaxed for frequencies in the range 4–5 Mc/s. For the N.T.S.C. and SECAM systems,[40,41] however, the lower limit should preferably be as shown in Fig. 13—unless a special 'high-boost' circuit is applied in the colour picture monitor.

3.4. *Overall Video Phase/Frequency Response; Each Sideband Separately*

The single sideband procedure might also be used for overall video phase measurements. A carrier/sideband voltage ratio of at least 30 dB is preferable to render the influence of quadrature distortion negligible. A phase-linear demodulator supplies the reference phase to the phase-meter, as depicted in Fig. 14. The two signal sources are connected to the Nyquist demodulator and the broadband demodulator through a resistive network providing correct impedance matching.

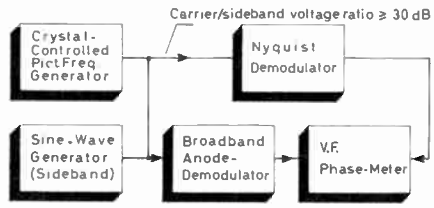


Fig. 14. The single-sideband method is also advantageous for overall video phase measurements. A demodulated sample o the input signal is used as reference signal for the phase-meter.

It was found that a simple diode detector could not be used because the input signal to the Nyquist demodulator was thereby distorted. Instead, a broadband anode demodulator, shown in Fig. 15, was used. The input impedance of this demodulator is constant, approximately equal to the terminating resistor R1. Anode demodulation is performed by the pentode E180F which is adjusted for Class BC operation by the cathode resistor R2. L1 and R3 are adjusted for flat amplitude response to beyond 6 Mc/s. Phase linearity in the nominal video frequency range is

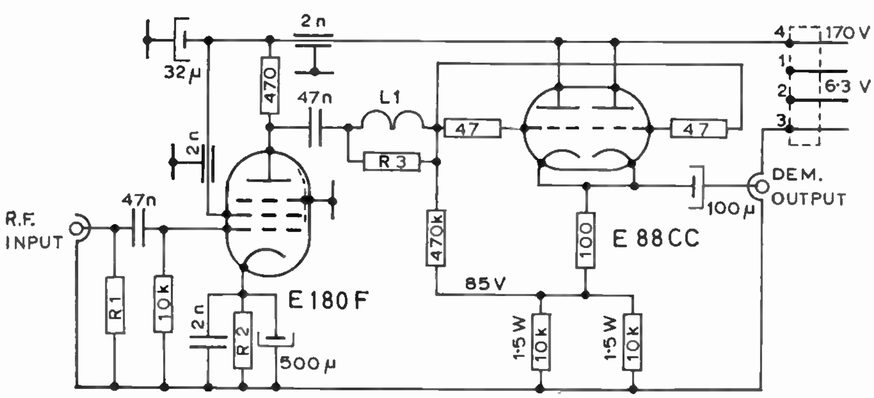


Fig. 15. The broadband anode demodulator used for supplying the reference phase in Fig. 14. A simple diode detector could not be used, because the input signal to the test object was then distorted.

thereby assured. A cathode follower E88CC provides a low output impedance (approximately 50 ohms).

The phase meter used is of our own design and consists basically of a phase meter for the frequency range 10 c/s–250 kc/s and a converter for the range 100 kc/s–10 Mc/s. Automatic frequency control of the local oscillator is applied to maintain an intermediate frequency of about 30 kc/s in the two separate channels of the converter. The two channels, which have identical phase characteristics, are also equipped with automatic gain control. Once the frequency control is switched in, no adjustment of the measuring equipment is necessary as the frequency and level of the input signals vary. This greatly facilitates measurement of phase/frequency characteristics. Automatic recording on a X-Y recorder is also made possible. (A second converter which extends the measuring range to 250 Mc/s, is now being made.) A description of this phase measuring equipment is to be published later.

The phase characteristic shown in Fig. 6 was measured using these techniques.

3.5. *Overall Video Amplitude Response*

It is necessary to know how the two sidebands combine in the resultant response for video frequencies

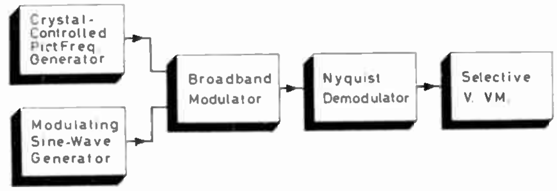


Fig. 16. To measure the overall amplitude response for video frequencies below 1·5 Mc/s, a double sideband modulator is used. This response is dependent on symmetry of Nyquist flank and the carrier location.

below 1·5 Mc/s. This could be measured using the test set-up shown in Fig. 16, in which a double sideband modulator is used. The correct picture carrier frequency is supplied by a crystal-controlled generator (or, if possible, by the transmitter). For frequency response measurements, a sine-wave generator is used for modulation with a low modulation index. The output voltage from the Nyquist demodulator is measured with a broad-band voltmeter and the result plotted against frequency. For measurement near cut-off, it might be necessary to use a selective valve voltmeter (Fig. 16).

The amplitude characteristics shown in the insets to Figs. 5 and 13, are measured by the procedure

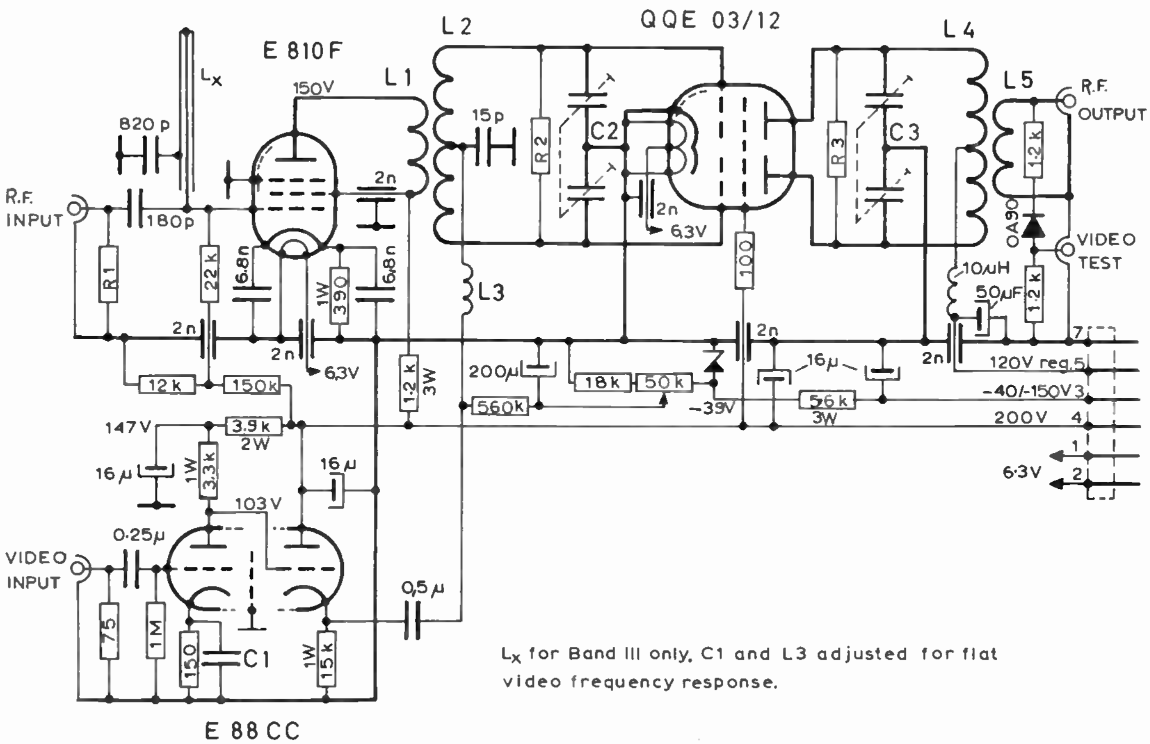


Fig. 17. Double sideband modulator designed for measurements on Nyquist demodulators (amplitude, phase and transient response). Operation as broadband power amplifier is also possible.

described here. To keep within the tolerance (±0·5 dB), it might be necessary to re-adjust the Nyquist flank close to the carrier frequency, because the low-frequency response is very much dependent on carrier location and symmetry of the Nyquist flank.

Figure 17 shows a broad-band modulator designed for measurements on Nyquist demodulators. The r.f. output voltage is 5 V into a load of 50 ohms. A source impedance of approximately 50 ohms is provided by the additional damping resistor R3. To obtain a correct matching over a broad frequency range, a 10 dB/50 ohm coaxial attenuator is normally used at the output.

The grid circuit of the QQE03/12 (the modulated r.f. amplifier) is also additionally damped, by the resistor R2, to avoid excessive incidental phase modulation. For the same reason, the QQE 03/12 is operated only in the negative grid region, and effective screening between the anode circuit and the grid circuit is provided.

For true double sideband modulation, the grid-circuit and the tank-circuit should be tuned for maximum response for the standard picture carrier of the television channel in which measurements are to be undertaken. This tuning for maximum is made by the trimmers C2 and C3 in Fig. 17.

(a)

(b)

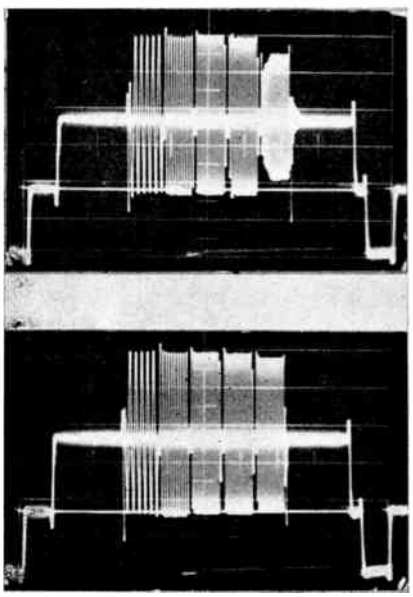


Fig. 18. A frequency burst signal checks the overall video response of the Nyquist demodulator with the characteristics shown in Fig. 9 and Fig. 13.

(a) 'Sound notch' switched in.

(b) 'Sound notch' switched out. Frequency bursts of 1, 2, 3, 4 and 5 Mc/s.

A 1-V signal to the 'video input' of the modulation amplifier (E88CC), modulates the r.f. carrier to a depth of approximately 50%. To reduce quadrature distortion, a lower modulation depth must be used for all measurements of amplitude, phase and transient response.

Cathode compensation by means of capacitor C1 and series-peaking by means of inductor L3 are applied for obtaining a flat modulation frequency response to beyond 6 Mc/s. This response might be controlled at the 'video test' output (Fig. 17).

The modulator might also be used as a broadband amplifier. For Band III, the fairly high input capacitance of the E810F is tuned out with a short piece of coaxial cable L_X (Fig. 17).

A frequency burst signal might conveniently be used to check the video response of the Nyquist demodulator. This signal is applied to the 'video input' of the modulator, and the signal at the output of the Nyquist demodulator is displayed on an oscilloscope. The result for the Nyquist demodulator with the characteristics previously shown in Fig. 9 and Fig. 13, is illustrated in Fig. 18.

3.6. *Overall Video Phase/Frequency Response*

The overall video phase characteristic might also be measured through the broadband modulator, as indicated on Fig. 19. This procedure gives additional information on the resultant phase distortion in the double sideband region, below approximately 1 Mc/s.

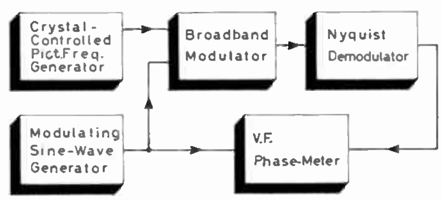


Fig. 19. The phase characteristic of the Nyquist demodulator might alternatively be measured through a double sideband modulator if this modulator is sufficiently linear.

Figure 20 shows the result for a Nyquist demodulator intended to be practically phase-linear. In (A) the group delay is derived from the measured phase characteristic by numerical derivation. The group delay variations for a Nyquist demodulator should preferably be less than 30 ns relative to the specified ideal characteristic, to facilitate group delay measurement of the transmitter.

Valuable information of the distorting effects on the transient response is gained if the deviation from the ideal phase characteristic is plotted, as in Fig. 20 (B). According to the 'theory of paired echoes',[42-44] the phase-deviation characteristic gives rise to an array of distorting 'echo pairs'. The magnitude and spacing of one such pair of echoes are determined by a corresponding term in the Fourier-expansion of the phase-deviation/frequency characteristic. A maximum phase-deviation of ± 4 deg, varying sinusoidally with frequency, gives a single echo-pair with magnitude 3·5% of the undistorted signal. The proposed tolerance limits of ± 4 deg (Fig. 20, B) are therefore not unduly strict.

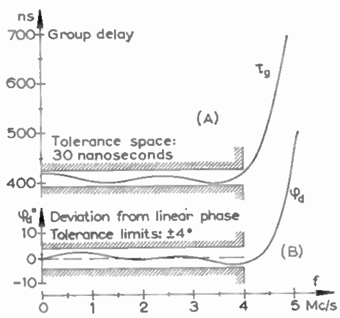


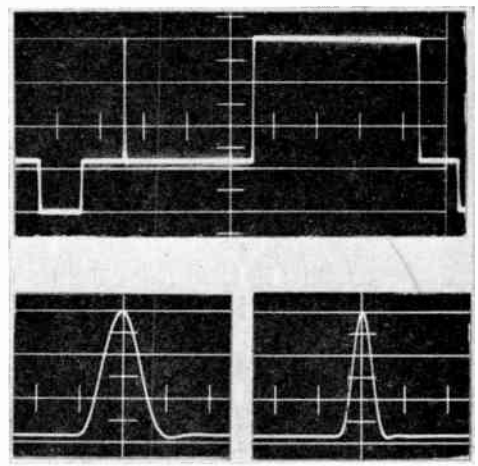
Fig. 20. Measured characteristics of a Nyquist demodulator intended to be practically phase-linear. For the proposed tolerance limits, see text.

The phase distortion in the frequency range 4–5 Mc/s, which is caused by the cut-off in the amplitude response, is not particularly harmful for black-and-white television. In case of N.T.S.C. compatible colour television, however, it would be necessary to introduce a definite precorrection at the transmitter for the receiver cut-off delay to reduce colour cross-talk and restore time coincidence of chrominance and luminance information.[45–46] The ideal cut-off precorrection characteristic would eventually have to be decided upon internationally. The current practice in the U.S.A. is to switch in a *fixed* receiver cut-off delay equalizer after the transmitter has been adjusted for phase-linearity. A phase-linear Nyquist demodulator is then used ('sound notch' switched off during measurement, cf. Fig. 13).[8]

3.7. *Transient Response and Picture Quality*

The picture quality is closely related to the transient response. It is advisable to measure this response, because the combining effects of the amplitude and phase characteristics on the picture reproduction cannot be easily predicted.

The transient response of a Nyquist demodulator might be measured through a double sideband modulator, e.g. by using the 'pulse-and-bar' test signal shown in Fig. 21. Either half amplitude duration of 0·1 μs (T) or 0·2 μs ($2T$) might be chosen for the sine-squared impulse. The '$2T$-pulse' has a frequency spectrum limited to 5 Mc/s and would therefore not be distorted in the ideal 5-Mc/s low-pass system. The 'T-pulse' has frequency components to 10 Mc/s and would normally be subject to distortions caused by cut-off ('Gibbs ringing'). Reference must be made to the literature[9–19] for a full description of the 'pulse-and-bar' test technique.



(a)

(b) (c)

Fig. 21. 'Pulse-and-bar' test signal[9–19] used for measurement of transient response of Nyquist demodulators.

(a) Composite signal,

(b) '$2T$-pulse' expanded (0·2 μs/division),

(c) 'T-pulse' expanded.

Figure 22 shows the 'pulse-and-bar' response of the Nyquist demodulator with the amplitude and phase characteristics previously shown in Figs. 5 and 6, respectively. In (a) the $2T$-response is displayed, and in (b) the T-response. On an expanded scale is shown (0·2 μs per division) the $2T$-pulse response (c) and the T-pulse response (d), with the 'bar' superimposed ('double triggering') for assessing the pulse/bar ratio. The step response, (e) and (f), is simply the integral of the impulse response and presents therefore no additional information. However, it should be observed whether both transitions are of the same shape. This would be the case if quadrature distortion is not present.

The picture quality related to this transient response is illustrated in Fig. 23 (C). Considerable preshoot and smear is evident in the electronic test pattern (c), also the 'ringing' is excessive. On the reproduced half-tone picture† (d), the preshoot is particularly disturbing.

Figure 23 (B) shows transient response and picture quality for a practically phase-linear Nyquist demodulator. The group delay and phase characteristic of

† Part of B.B.C. Test Transparency No. 52.

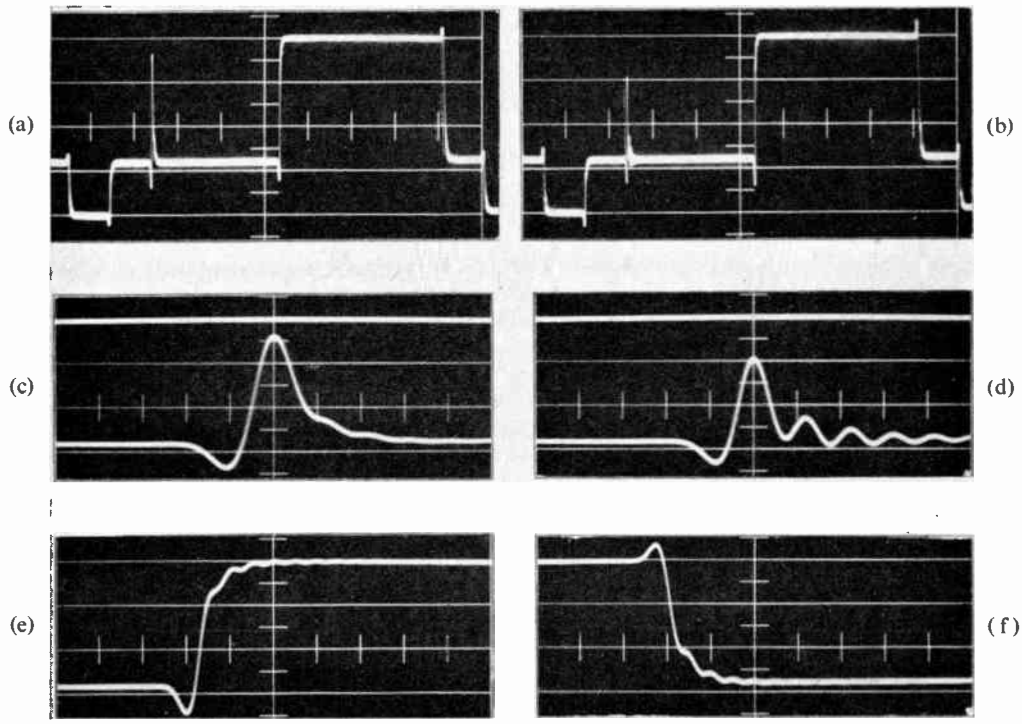


Fig. 22. Pulse-and-bar response of the Nyquist demodulator with the characteristics shown in Fig. 5 and Fig. 6.

(a) 2*T*-response, (b) *T*-response, (c) 2*T*-pulse response, (d) *T*-pulse response, (e) and (f) Step response.

this demodulator is depicted in Fig. 20. As seen, the preshoot and smear are completely eliminated. The 'ringing' is reduced, now appearing more symmetrical (b). The improvement in picture resolution, as judged from the 5-Mc/s striation on the reproduced electronic test pattern (c), is partly due to a residual response lobe in the upper adjacent channel.

For comparison, the undistorted test signals and pictures are shown in Fig. 23 (A).

Some restrictions should be set on the transient response of the Nyquist demodulator, because the phase equalizers of the transmitter are adjusted while observing the overall transient response. Because the Nyquist demodulator is not required to have a definite high-frequency cut-off characteristic (cf. Fig. 13), the ideal *T*-pulse response is not defined. The 2*T*-pulse, however, has a frequency spectrum limited to the nominal video bandwidth, and should ideally not be distorted in the Nyquist demodulator. It is therefore proposed to tolerate the 2*T*-pulse response, e.g. by using the *K*-rating system.[10–19] This particularly applies to Nyquist demodulators intended to be phase-linear. If a definite predistortion is prescribed, it is possible to compute the 2*T*-pulse response for the specified delay characteristic. This response could then be tolerated, cf. ref. 30.

3.8. *Linearity, Differential Gain and Differential Phase*

For measurement of linearity, the same test set-up as in section 3.3 is used (Fig. 12). The sideband is held constant, e.g. corresponding to a modulation frequency of 1 Mc/s, and the carrier level is varied within the region assigned for black-to-white transition. If output voltage is plotted against carrier level, a measure of the non-linearity is obtained.

The linearity of the Nyquist demodulator should be as good as possible, although the exact requirement would depend on the test procedures applied at the transmitter station, cf. Sect. 2.2. (*d*) and (*f*).

Differential gain and differential phase have significance only in case of compatible colour television. To measure the differential gain, the same procedure as described above is used; however, the sideband frequency should then correspond to a modulation frequency of 4·43 Mc/s (the chrominance subcarrier).

Differential phase measurement is carried out in an analogous way, using the test set-up in Fig. 14.

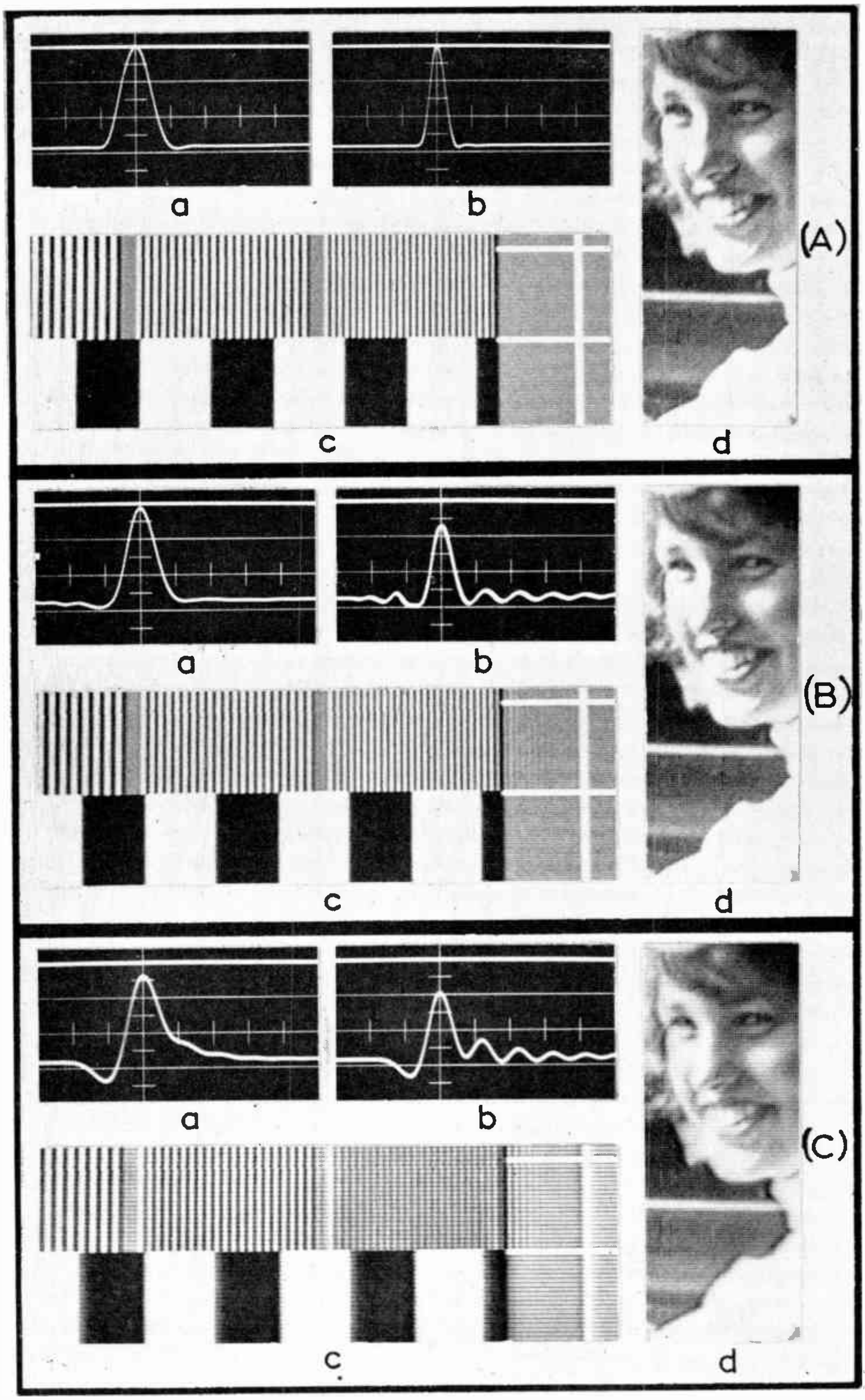


Fig. 23. (A) $2T$-pulse (a), T-pulse (b), electronic test picture (c) and half-tone picture (d) in case of no distortions.
(B) Same, but now at the output of a practically phase-linear Nyquist demodulator (cf. Fig. 20).
(C) Same, but in this case at the output of the non-phase-linear demodulator with the characteristics of Fig. 5 and Fig. 6.

3.9. *Noise and Spurious Signal Components*

The noise level of the Nyquist demodulator is measured at the output, using a wide-band oscilloscope. Noise and spurious signals should at least be 40 dB lower than the nominal video output level to make it possible to assess the additional noise of the transmitter.

4. Conclusions

To decide whether the transmitted signal is in compliance with agreed standards or not, it is necessary to use a measuring demodulator with accurately known characteristics. Ideal and practical characteristics of the system, as well as the test methods used at the transmitter, were analysed to find the preferable characteristics for this demodulator. Valuable information from measured demodulators of different manufacture made it possible to propose tolerance limits which could be maintained in practice.

The described test procedures provide some 'overlap', and not all of them would generally be necessary. However, because the tolerances to be met are small, different test methods might be required as a control.

5. Acknowledgments

This work was carried out as part of a research project concerning picture quality in television, sponsored by the Royal Norwegian Council for Scientific and Industrial Research and the Norwegian Broadcasting Corporation. The author wishes to express his appreciation to these institutions.

Thanks are also due to the Norwegian Telecommunication Administration for their co-operation and for permission to use some of the measured characteristics for illustrating this paper.

6. References

1. C.C.I.R. Recommendation No. 266, Los Angeles 1959, "Phase Correction of Television Transmitters Necessitated by the Use of Vestigial Sideband Transmission".
2. D. Bünemann, "Der Laufzeitausgleich eines Fernsehsystems", *Archiv Elektr. Übertragung*, **10**, pp. 10–18, 1956.
3. D. Bünemann and W. Händler, "Der Schwarz-Weiss-Sprung bei der Restseitenband-Fernseh-Übertragung", *Archiv Elektr. Übertragung*, **10**, pp. 457–66, 1956.
4. H. Hopf, "Laufzeitausgleich für die Restseitenbandübertragung im Fernsehen", *Rundfunktech. Mitt.*, **2**, pp. 180–3, 1958.
5. H. Borkan, W. C. Morrison and J. G. Reddeck, "Video test generator", *Electronics*, **25**, pp. 139–41, September 1952.
6. W. C. Morrison, K. Karstad and W. L. Behrend, "Test instruments for color television", *Proc. I.R.E.*, **42**, pp. 247–58, 1954.
7. J. F. Fisher, "Alignment of a monochrome TV transmitter for broadcasting N.T.S.C. color signals". *Proc. I.R.E.*, **42**, pp. 263–70, 1954.
8. T. M. Gluyas, Jr., "Television transmitter considerations in color broadcasting", *R.C.A. Review*, **15**, pp. 312–34, 1954.
9. I. F. Macdiarmid, "A testing pulse for television links", *Proc. Instn Elect. Engrs*, **99**, Part III A, pp. 436–44, 1952.
10. N. W. Lewis, "Waveform responses of television links", *Proc. Instn Elect. Engrs*, **101**, Part III, pp. 258–70, 1954.
11. R. C. Kennedy, "Sine-squared pulses test color-TV systems", *Electronics*, **27**, pp. 138–9, December 1954.
12. I. F. Macdiarmid and B. Phillips, "A pulse-and-bar waveform generator for testing television links", *Proc. Instn Elect. Engrs*, **105**, Part B, pp. 440–8, 1958.
13. I. F. Macdiarmid, "Waveform distortion in television links", *P.O. Elect. Engrs J.*, **52**, pp. 108–14 and pp. 188–95, 1959.
14. I. F. Macdiarmid, "Waveform distortion in television links", *J. Brit. I.R.E.*, **20**, pp. 201–16, 1960.
15. R. C. Kennedy, "Sine-squared pulses in television system analysis", *R.C.A. Review*, **21**, pp. 253–65, 1960.
16. H. Springer, "Betrachtungen über die zweckmässige Messung und Tolerierung des Einschwingverhaltens von Fernsehgeräten", *Nachrichtentech. Z.*, **15**, pp. 57–62, 1962.
17. H. Springer, "Die Vorteile des Impuls- und Sprung-Signales bei der Messung linearer Übertragungsverzerrungen", *Rundfunktech. Mitt.*, **7**, pp. 25–41, 1963.
18. K. H. Potts, "The performance evaluation of television line and microwave links using sine-squared test techniques", *The Radio and Electronic Engineer*, **25**, pp. 357–76, 1963.
19. F. G. Johannesen, "Pulsmåling av nettverk og transmisjonssystemer" ("Pulse testing of networks and transmission systems"), *Elektroteknisk Tidsskrift*, **77**, pp. 205–16, 1964.
20. J. A. Bauer, "A versatile approach to the measurement of amplitude distortion in color television", *Proc. I.R.E.*, **42**, pp. 240–6, 1954.
21. G. L. Fredendall and W. C. Morrison, "Effect of transmitter characteristics on N.T.S.C. color television signals", *Proc. I.R.E.*, **42**, pp. 95–105, 1954.
22. J. Piening, "Über Entstehung und Messung Pegelabhängiger Phasen- und Amplitudenänderungen bei der Hilfsträgerübertragung im Farbfernsehen", *Nachrichtentech. Z.*, **11**, pp. 70–7, 1958.
23. J. Ruston, "Improving TV system transient response", *Electronics*, **25**, pp. 110–13, August 1952.
24. H. J. Griese, "Möglichkeiten zur Entzerrung der Restseitenbandübertragung des Fernsehens", *Fernmeldetech. Z.*, **8**, pp. 94–103, 1955.
25. H. Breimer, "The influence of main- and vestigial-sideband widths on picture quality" (I.E.E. Paper No. 3928E, June 1963). Proceedings of the International Television Conference 1962, pp. 5–11.
26. A.R.D. (Arbeitsgemeinshaft der Rundfunkanstalten der Bundesrepublik Deutschland), Pflichtenheft Nr. 5/21, Ausgabe Nr. 1.
27. G. Coldewey, "Kurze Darstellung der zusätzlichen Forderungen an Schwarz-Weiss-Fernsehsender bei der Übertragung von farbigen Bildern", *Nachrichtentechn. Z.*, **13**, pp. 175–8, 1960.
28. P. Klopf, "Modulation von Fernsehsendern im Zwischenfrequenzbereich", *Rundfunktech. Mitt.*, **5**, pp. 79–87, 1961.
29. H. W. Bode, "Network Analysis and Feedback Amplifier Design" (D. van Nostrand, New York, 1946).

30. R. D. Kell and G. L. Fredendall, "Standardization of the transient response of television transmitters", *R.C.A. Review*, **10**, pp. 17–34, 1949.

31. A.R.D., Pflichtenheft Nr. 5/22, Ausgabe Nr. 1.

32. F. Schröter, "Fernsehtechnik. Zweiter Teil" (Sect. IX, 2.2) (Springer Verlag, Berlin, 1963).

33. A.R.D., Ergängung zum Pflichtenheft Nr. 5/22, Ausgabe Nr. 1.

34. A. van Weel, "Phase linearity of television receivers", *Philips Tech. Rev.*, **18**, pp. 33–51, 1956.

35. W. Strössenreuther, "Messgeräte für die hochfrequente Übertragung von Schwarz-Weiss- and Farbsignalen", *Rundfunktech. Mitt.*, **8**, pp. 212–9, 1964.

36. H. Nyquist and S. Brand, "Measurement of phase distortion", *Bell Syst. Tech. J.*, **9**, pp. 522–49, 1930.

37. C. J. Heuvelman and A. van Weel, "Group delay measurements", *Wireless Engineer*, **33**, pp. 105–13, 1956.

38. W. F. Bailey and C. J. Hirsch, "Quadrature cross talk in N.T.S.C. color television", *Proc. I.R.E.*, **42**, pp. 84–90, 1954.

39. W. Bruch, "Das PAL-Farbfernsehsystem–Prinzipielle Grundlagen der Modulation und Demodulation", *Nachrichtentech. Z.*, **17**, pp. 109–21, 1964.

40. H. Schönfelder, "Verzerrungen durch System- und Übertragungsfehler beim Farbfernsehen nach dem NTSC-Verfahren", *Rundfunktech. Mitt.*, **4**, pp. 158–72, 1960.

41. H. Schönfelder, "Der Einfluss von System- und Übertragungsfehlern bei einer Farbfernsehübertragung nach dem SECAM-Verfahren", *Archiv Elektr. Übertragung*, **16**, pp. 385–99, 1962.

42. H. A. Wheeler, "The interpretation of amplitude and phase distortion in terms of paired echoes", *Proc. I.R.E.*, **27**, pp. 359–85, 1939.

43. F. Strecker, "Über den Einfluss kleiner Phasenverzerrungen auf die Übertragung von Fernsehsignalen", *Elektrischen Nachrichtentechnik*, **17**, pp. 51–6, 1940.

44. F. Strecker, "Beeinflussung der Kurvenform von Vorgängen durch Dämpfungs- und Phasenverzerrung", *Elektrischen Nachrichtentechnik*, **17**, pp. 93–107, 1940.

45. R. C. Palmer, "System delay characteristics in N.T.S.C. color television", *Proc. I.R.E.*, **42**, pp. 92–5, 1954.

46. G. L. Fredendall, "Delay equalization in color television" *Proc. I.R.E.*, **42**, pp. 258–62, 1954.

Manuscript received by the Institution on 18th February 1965. (Paper No. 1002/T32.)

STANDARD FREQUENCY TRANSMISSIONS

(Communication from the National Physical Laboratory)

Deviations, in parts in 10^{10}, from nominal frequency for **August 1965**

August 1965	GBR 16kc/s 24-hour mean centred on 0300 U.T.	MSF 60 kc/s 1430–1530 U.T.	Droitwich 200 kc/s 1000–1100 U.T.	August 1965	GBR 16 kc/s 24-hour mean centred on 0300 U.T.	MSF 60 kc/s 1430–1530 U.T.	Droitwich 200 kc/s 1000–1100 U.T.
1	− 150·8	− 150·2	—	17	− 148·4	− 149·2	− 1
2	− 149·2	− 148·5	− 1	18	− 149·0	− 150·8	− 1
3	− 149·7	− 152·8	− 1	19	− 150·4	− 150·2	− 2
4	− 151·9	− 152·1	− 1	20	− 150·6	− 150·0	0
5	− 151·0	− 150·8	− 9	21	− 148·9	− 148·5	− 1
6	− 150·6	− 150·1	− 8	22	− 148·1	− 149·7	0
7	− 149·5	− 149·6	− 1	23	− 150·7	− 151·6	0
8	− 149·6	− 149·8	− 1	24	− 150·8	− 151·5	0
9	− 150·0	− 149·6	− 1	25	− 151·2	− 151·9	− 1
10	− 149·3	− 149·6	− 1	26	− 151·7	− 150·3	+ 1
11	− 149·1	− 149·8	− 2	27	− 150·2	− 151·2	0
12	− 149·3	− 150·9	− 1	28	− 150·3	− 151·1	+ 1
13	− 149·6	− 149·6	− 3	29	− 149·8	− 149·4	—
14	− 148·7	− 148·5	− 1	30	− 149·4	− 149·0	+ 1
15	− 150·0	− 149·7	− 2	31	− 148·2	− 149·3	+ 2
16	− 150·0	− 149·7	− 2				

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

Note: The phase of the GBR/MSF time signals was retarded by 100 milliseconds at 0000 U.T. on 1st September 1965.

Radio Engineering Overseas . . .

The following abstracts are selected from Commonwealth, European and Asian journals received by the Institution's Library. Abstracts of papers published in American journals are not included because they are available in many other publications. Members who wish to consult any of the papers quoted should apply to the Librarian, giving full bibliographical details, i.e. title, author, journal and date, of the paper required. All papers are in the language of the country of origin of the journal unless otherwise stated. Translations cannot be supplied.

MAGNETOSTRICTIVE DELAY LINE

A magnetostrictive delay line working as a pulse distributor for twelve independent outputs has been described by a Greek engineer; the line transmits pulses with a frequency of 8 kc/s and a duration of 1·5 microseconds approximately; the taps are each delayed by 9·6 microseconds. The paper gives an overall description with some technological comment and results of measurements. A special modification has done away with the necessity of providing a damping body on the input side of the line. A detailed analysis is given of the power load in the transistor circuit feeding the exciter coil, which is working under rather unusual conditions. The line is being used in a twelve-channel multiplex telephone system for a radio relay link with pulse position modulation.

" A magnetostrictive delay line for a pulse modulated multiplex system", M. Ohera, *Slaboproudý Obzor*, **25**, No. 11, pp. 631–8, November 1964.

SURFACE WAVE AERIALS

The different theories advanced to explain the performance of uniform and sinusoidally modulated surface wave aerials have been reviewed in an Australian paper. Precise measurements of the amplitude and phase of the near-field, together with radiation pattern measurements, show the precise mechanism of radiation from both uniform and sinusoidally modulated aerials. The results of these measurements demonstrate that the endfire array theory, which is often used to explain the performance of surface wave aerials, is invalid; and also that corrections must be made to the simplified treatment of the surface wave aerial as a non-radiating transmission line.

"The precise mechanism of radiation from surface wave aerials", B. MacA. Thomas, *Journal of The Institution of Engineers, Australia*, **36**, No. 9, pp. 225–38, September 1964.

MICROWAVE FREQUENCY DISCRIMINATOR

A new design of frequency discriminator of the waveguide type is described in a Czech paper; its error voltage is derived from the phase shifts of the electromagnetic field during the detuning of an impedance coupled cavity. The error voltage during the detuning of the oscillator in the 3000-Mc/s band is computed rigorously and the relations for the broad-band properties of the discriminator are derived. The design of a discriminator with a waveguide cavity coupled by an inductive slot is described briefly.

"A waveguide frequency discriminator", B. Novák, *Slaboproudý Obzor* (Prague), **25**, No. 6, pp. 353–61, 1964.

RADAR INVESTIGATION OF TROPOSPHERE

A radar radiating vertically into the atmosphere produces echoes even in the case of a completely clear sky. These echoes are in most cases produced by reflections or back-scattering at inhomogeneities of the refractive index of air. These phenomena, known as ghost returns or 'angels', have been investigated at the German central bureau for telecommunications (Fernmeldetechnisches-Zentralamt) by means of vertically-directed 10 cm radar. The aim of the investigation was to evaluate the phenomena and the possible application of the knowledge gained to problems of terrestrial radio links and links to satellites. The paper discusses the principle of measurement, accuracy of measurement, intensity and duration of ghost returns, as well as their frequency of occurrence as a function of the time of the day and height.

" An investigation of the structure of the troposphere by means of a vertically directed radar", L. Fehlhaber and J. Grosskopf, *Nachrichtentechnische Zeitschrift*, **17**, No. 10, pp. 503–12 October 1964.

BROADBAND TRANSISTOR AMPLIFIER

It has been shown that, for gain-peaking of broadband amplifiers, whose upper frequency limit lies in the u.h.f. region, and in the vicinity of the transit frequency of the chosen transistors, a transmission line with a wave impedance of several hundred ohms may be used as coupling element. In a German paper a procedure for dimensioning the transmission line is derived and applied to a sample problem.

A three-stage amplifier with transmission line coupling is described which has a gain of $19{\cdot}5 \pm 1{\cdot}5$ dB between 170 and 640 Mc/s.

"Broadband transistor amplifier" H. Klink, *Archiv der Elektrischen Übertragung*, **18**, No. 6, pp. 350–354, June 1964.

HALL-EFFECT PHASE DETECTOR

Hall effect detectors, although free from the necessity of sliding contacts, have the disadvantage of providing an output voltage whose amplitude is not only proportional to the Hall coefficient R, but also varies as a function of temperature, applied magnetic field and time.

The authors of this French paper describe a device in which the amplitude of the output voltage is independent of R but the relevant angle is given simply by the phase of the Hall voltage.

"Design of a phase detector based on the Hall effect", F. Perrot and R. Bonnefille, *L'Onde Électrique*, **45**, No. 455, pp. 234–9, February 1965.