THE RADIO AND ELECTRONIC ENGINEER

The Journal of the British Institution of 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 26 JULY 1963 NUMBER 1

RELIABILITY—THEORY AND PRACTICE

The problem of producing equipment of high reliability depends on many factors, including the economic need to replace craftsmen by process operators, especially in component manufacture. So many variables are involved that it is incumbent upon the engineer constantly to ensure that proved theories of reliability are applied at all stages of development and production.

In an editorial commenting on the report—" Reliability of Electronic Equipment "†—the Chairman of the Technical Committee pointed out that reliability depends essentially on the proper mental approach. Its achievement is an aim which must be borne in mind continually by *everyone* concerned with the design and production of electronic equipment if that equipment is to operate to the entire satisfaction of the user. The Institution's Report, meetings held by other professional bodies in recent years, Ministry of Aviation conferences and publications, textbooks, papers and articles have all pointed, in general or particular terms, to ways of designing reliability into equipment. Recently a British Standards Institution Committee (TLE/16), on which the Brit.I.R.E. is represented, has been set up to ensure that all future standards take into consideration the question of reliability. For over five years this intensive propaganda has been poured out, but has it really made any impact on the engineer who is not compelled by contractual requirements (such as apply in many military projects) to attain a certain degree of reliability in the performance of his equipment?

The Services and manufacturers of especially complex equipment (e.g. computers) and devices such as submarine telephone repeaters and communications satellites, lay down specifications for construction and testing. Basically, nevertheless, there is no reason—except cost—why *all* equipment, whether for domestic or industrial use, should not be designed for reliable operation and long life—and the growth of experience in design techniques must surely lead eventually to achievement of reliability more easily and cheaply.

The need for reliability in industrial equipment is linked irrevocably with the increased use of such equipment—manufacturers in other industries will not be disposed to acquire new devices, however versatile, if the services of a resident engineer are required. Similar caustic comments were made some years ago in connection with the critical adjustments required for the first colour television receiver in the United States; now that technical advances indicate that this is not likely to be a major problem with future designs, it is vital that the introduction of colour television in Great Britain should not be marred by unreliability of equipment.

The ultimate aim must therefore be to co-ordinate theory and practice of reliability throughout all branches of the radio and electronics industry and through all levels—the technician and the assembly line operatives must be as aware of its prime importance as the development engineer and the board of directors. Many different approaches, by education and other means, will be necessary if the philosophy and 'know-how' are to permeate through all the strata of industry. As far as the Institution is concerned, the subject is high on the agenda when future activities are being planned.

F. W. S.

† *J.Brit.I.R.E.*, **23**, No. 4, pages 287–95, April 1962.

INSTITUTION NOTICES

Birthday Honours List

The Officers and Council of the Institution have congratulated the following members whose names appear in Her Majesty's 1963 Birthday Honours List. Both appointments are to the Military Division of the Most Excellent Order of the British Empire.

Captain A. J. B. Naish, R.N., M.A. (Member) appointed Ordinary Commander of the Order. (Captain Naish is on the staff of the Director of Electrical Engineering, Ship Department, Admiralty, Bath; he served on the Institution's Council from 1954–59 and serves on the Membership Committee of which he was for several years chairman.)

Lieutenant Colonel J. T. R. Sylvester-Bradley, M.A., R. Sigs. (Associate Member) appointed Ordinary Officer of the Order. (Colonel Sylvester-Bradley, who was formerly at the School of Signals, Catterick, is now overseas with the 16th Signals Regiment.)

Institution Co-operation with National Committees

Whenever a new activity impinges on the work of a number of professional bodies, it is sometimes desirable to co-ordinate their activities in the field and to provide a forum where views can be exchanged. The Institution is associated with four such organizations, the British Conference on Automation and Computation, the British National Committee for Non-Destructive Testing, the National Council for Quality and Reliability and the British Nuclear Energy Society. The main function of these bodies is to organize meetings which members of the individual societies can attend and, probably more important, to provide a central body which can co-ordinate British participation in international conferences and conventions.

As an example, the British Conference on Automation and Computation has set up two committees to organize the British contributions to the analogue computer conference to be held in London in 1964 and to the International Measurements Conference to be held in Stockholm in 1964.

Institution representatives on these Committees are:

British Conference on Automation and Computation. A. St. Johnston (Member) and W. Renwick (Member).

Committee for International Federation on Analogue Computation. K. H. Simpkin (Member).

Committee for International Measurements Conference (IMEKO). A. G. Wray (Member).

British National Committee for Non-Destructive Testing. A. Nemet (Member).

National Council for Quality and Reliability. F. G. Diver, M.B.E. (Member).

British Nuclear Energy Society. Col. G. W. Raby, C.B.E. (Member), and R. J. Cox (Associate Member).

The Institution's Office in India

Members in India are advised that correspondence in connection with the activities of the Indian Division should be addressed to the Administrative Secretary of the Division, Captain G. Swaminathan, 33/3 Infantry Road, Bangalore 1.

The Administrative Secretary holds stocks of Regulations, forms, copies of past *Journals*, etc.

Standard Frequency Transmissions

The Council is pleased to announce that arrangements have been made with the Director of the National Physical Laboratory for the publication in the Institution's *Journal* of measurements made by the Laboratory's Standards Division on the frequency of certain standard frequency transmissions. It is believed that this service will be particularly valuable to radio engineers concerned with precise frequency measurements.

The first set of values relating to the month of June 1963 are published on page 34 of this issue and subsequent tables will appear each month. A short contribution from N.P.L. explaining the methods of measurement is given on page 78.

Index to Volume 25

The index to Volume 25 of the *Journal* (the first half of 1963, January to June) has now been prepared and copies are being sent with this issue to all members and subscribers.

Members are reminded that they may send their *Journals* (six issues plus index) to the Institution for binding. The charge for this service is 16s. 6d., postage extra (Great Britain 3s.; other countries 4s.).

Corrections

The following amendments should be made to the paper "Visual Detection in Intensity-modulated Displays" which was published in the March 1963 issue of *The Radio and Electronic Engineer*:

Page 226, line 18: The factor *should be* $\sqrt{N}$.

Page 231, Table 2: The value of S/N ratio for brightness level of 1·0 ft-lambert and target diameter 14·7′ *should be* −13·6 dB.

Page 232, Table 3: The heading *should read* The effect of non-linearity on signal/noise ratio.

Page 236, left-hand column, 9 lines from the bottom, *should read* 4·4 deg.

Page 238, left-hand column, 13th line and equation 29 (first part): the factor *D should be inserted.*

Page 238, right-hand column, lines 12 and 13 *should read* Fig. 13.

U.D.C. 621.321.37/8

Mass Flow Measurement with Turbine-Type Flowmeters— An Electronic Method of Density Correction

By

I. C. HUTCHEON, M.A.†

AND

L. S. DUFFY†

Presented at the Convention on "Electronics and Productivity" in Southampton on 17th April 1963.

Summary: The temperature of the fuel delivered to an aircraft engine is measured by a silicon resistor, and the volume flow rate is measured by a turbine-type flowmeter fitted with a saturable magnetic pick-off. Electronic circuits convert the transmitted pulse rate and temperature signal into a direct current and into a slightly different pulse rate, both representing the mass rate of flow of fuel. The current operates a rate-meter, and the pulses drive a counter which displays the total mass of fuel consumed. The principles could be applied to industrial mass flow measurement.

1. Introduction

Turbine and positive-displacement type flowmeters have been widely used for many years for measuring the flow of water, oil and other clean liquids in numerous types of process. They are simple robust devices, usually comprising a rotor or an eccentric piston mounted on bearings in a suitable body, and driving a mechanical counter directly. Their accuracy is typically ±1% or better over a flow range of at least 10 to 1, and because of their simple construction they are relatively inexpensive and reliable.

If the measurement has to be transmitted over a distance, however, or used for control purposes, a direct mechanical readout is inadequate. Systems have been developed in recent years, therefore, in which the rotor speed is sensed by an electromagnetic pick-off and transmitted as a pulse rate to an electronic receiver or computer. This method has three main advantages:

(a) negligible load is imposed on the rotor,

(b) the pick-off is external to the flowmeter, and

(c) various operations can be performed on the signal.

The paper describes a system of this type which is suitable for fuel flow measurement in aircraft. The flow rate and the temperature of the fuel are measured and used to compute the flow rate and total flow in mass units. The principles of the system are equally applicable to mass flow measurement under industrial conditions.

2. System

Figure 1 shows the three units of the mass flow measuring system which is illustrated diagrammatically in Fig. 2. The flowmeter is mounted in the fuel line of the aircraft, and carries a small probe unit containing a saturable magnetic pick-off and a temperature-sensitive voltage divider.

The pick-off modulates a carrier signal, generated in the receiver, at a frequency proportional to volume flow rate. This is demodulated in the receiver and used to drive a diode pump whose stroke is the difference between two voltages V_a and V_m.

Voltage V_m is pre-set manually according to the specific gravity of the fuel at some nominal temperature, while voltage V_a is derived from the divider in the probe. Since specific gravity and temperature are linearly related, V_a follows changes in specific gravity with temperature, and $V_a - V_m$ can be made to represent the specific gravity of the fuel at its operating temperature. The output of the pump then is a pulsating direct current which is proportional to the product of volume flow rate and actual specific gravity, i.e. to mass flow rate. This current is passed through a moving-coil rate indicator and into a large reservoir capacitor.

It is extracted again from the reservoir capacitor by another diode pump whose stroke is constant. A sensitive error amplifier monitors the voltage across the capacitor, and controls the rate of operation of the pump so that the error voltage remains small. The

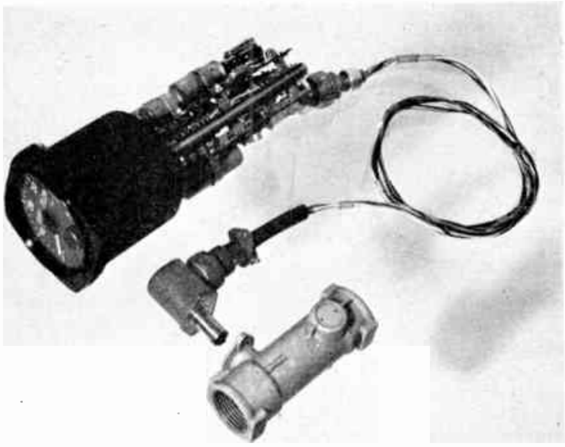


Fig. 1. Flowmeter, probe, and receiver with the case removed.

† Research and Development Department, George Kent Ltd., Luton, Beds.

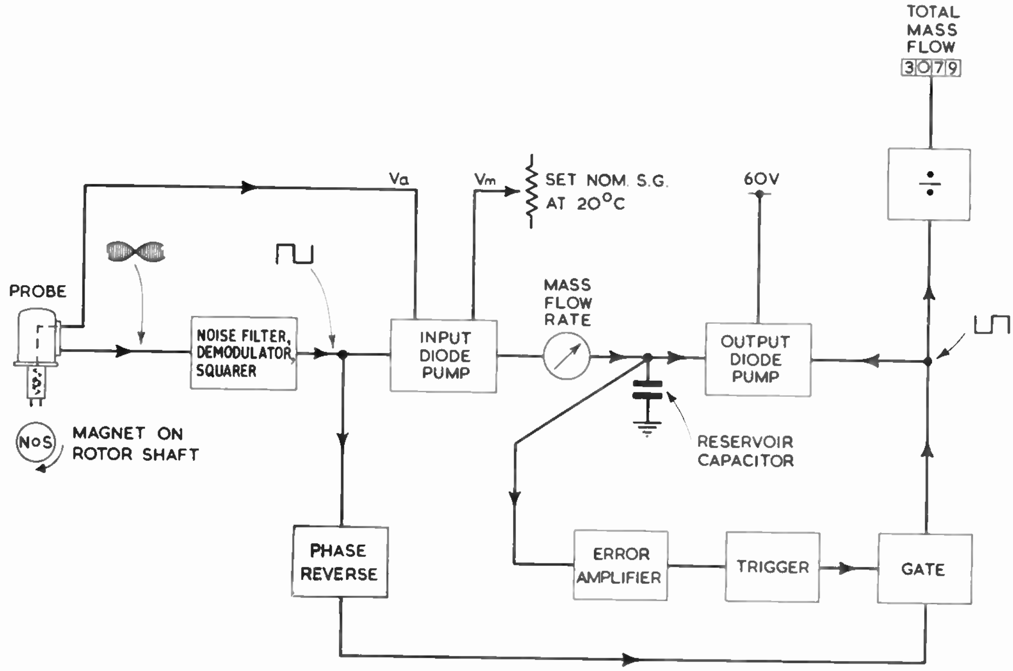


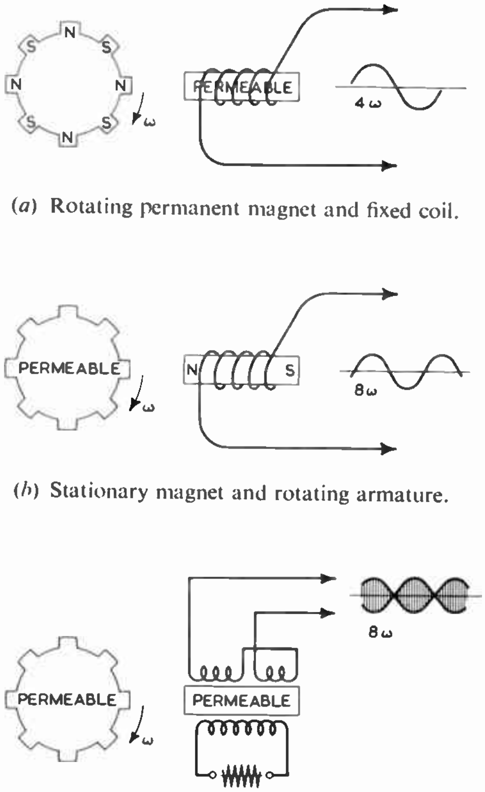
Fig. 2. System for mass flow measurement.

current drawn by the amplifier is negligible, hence virtually all the current reaching the capacitor leaves it again in the form of equal quantities of charge, each representing a known mass of fuel. The total number of these increments in any period is a measure of the total mass flow and, after suitable frequency division, is displayed by the counter. The symmetry of the electronic system confers several advantages. For example, variations in h.t. supply have almost no effect on the accuracy of the counter reading, since both pumps are operated from the same supply rail. Similarly, errors caused by the finite voltage drop, which remains across the pump diodes at the end of a pump or reset period, are largely cancelled out; the effects of variations in value of the pump capacitors with temperature are also cancelled out.

3. Magnetic Pick-off

The pick-off is required to deliver pulses at a rate proportional to rotor speed, while imposing negligible load torque on the rotor. It has to operate through the wall of the flowmeter or of a probe inserted in it, and should not be too critically dependent on the distance between itself and the operating magnet or armature. The output signal should be immune to interference of any type likely to be encountered. Three types of probe commonly used are shown in Fig. 3.

The first uses the rotation of a permanent magnet to induce an e.m.f. in a fixed coil (Fig. 3(*a*)); the second has the magnet stationary and the permeable part rotating (Fig. 3(*b*)). The amplitude of the output



(*a*) Rotating permanent magnet and fixed coil.

(*b*) Stationary magnet and rotating armature.

(*c*) Position-sensitive pick-off.

Fig. 3. Three types of magnetic pick-off.

signal is proportional to the speed of rotation in both cases, and, for operation at low speeds, the coil usually must have a large number of turns. To minimize this difficulty, the armature, whether magnetized or permeable, frequently has more than two poles, and the rotor speed is kept as high as possible.

The third type of pick-off is position-sensitive, relying on the armature to unbalance a differential transformer or similar device[1] which is supplied with a carrier signal of relatively high frequency (Fig. 3(*c*)). Although it is rather more complicated, this method has the advantage that the output signal has the same amplitude at all frequencies, and can be detected in the presence of noise by the use of tuned filters, synchronous demodulation, etc. Also, low rotor speeds can be used, with consequent increase in bearing life.

The type of pick-off used in the system described is shown in Fig. 4. This design[2] is also position-sensitive, but relies on magnetic saturation for its operation, being analogous to the flux-gate magnetometer. It has the additional advantages that only two wires are needed to connect it to the receiving unit, and the coil requires few turns.

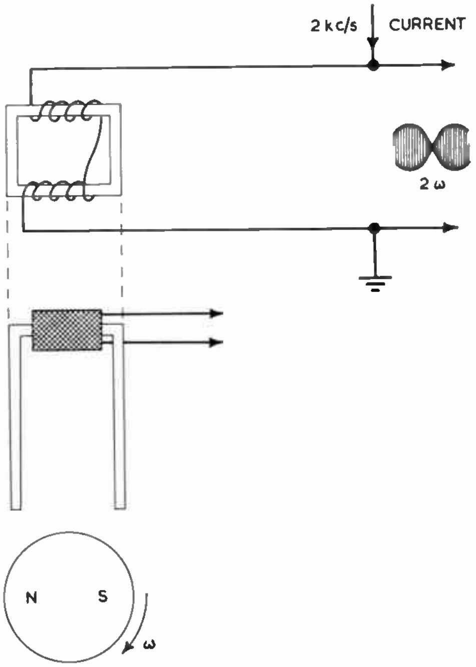
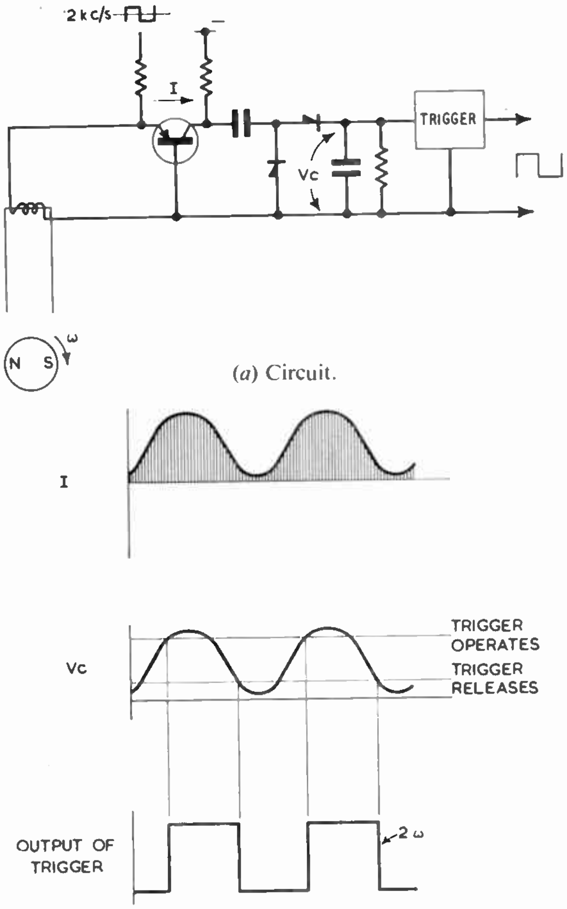


Fig. 4. Saturable magnetic pick-off.



(*a*) Circuit.

(*b*) Waveforms.

Fig. 5. Noise filter and demodulator.

As the magnet rotates, it periodically saturates the two thin cross members of a mumetal armature, causing the impedance of the coil to fluctuate widely. The unit can therefore be considered and used as a switch to control the path taken by a carrier frequency alternating current.

4. Noise Filter and Demodulator

Figure 5 shows the method chosen for energizing the pick-off, demodulating the signal which it delivers, and eliminating the effects of stray pick-up.

A 2 kc/s energizing current is alternately shunted to earth via the pick-off, and allowed to flow into a grounded-base transistor amplifier which drives a subsidiary diode pump. The pump charges a capacitor which is resistively loaded so that its voltage closely follows the envelope of the modulated carrier signal, and a trigger circuit develops the desired square-wave output signal. Circuit values are chosen so that several pumping operations are required to operate the trigger circuit, hence the probability of its being operated by extraneous signals is extremely small.

5. Computation of Mass Flow Rate

5.1. *Voltage Analogue of Fuel Specific Gravity*

Figure 6 shows how the specific gravity of jet-aircraft fuels varies with temperature. The lines are straight and approximately, though not exactly, parallel.

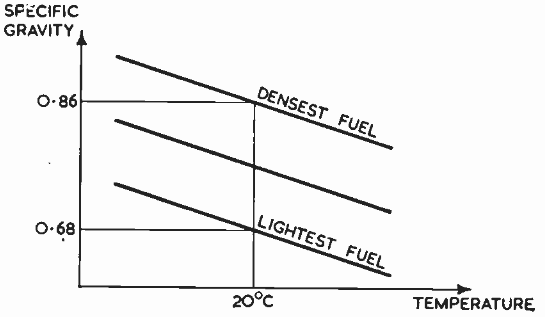


Fig. 6. Specific gravity of aircraft fuels vs temperature.

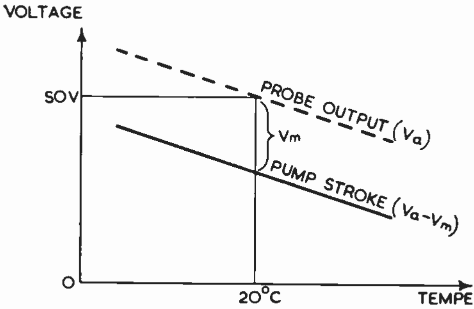


Fig. 7. Voltage analogue of fuel specific gravity.

As shown in Fig. 7, therefore, it is possible to represent with good accuracy the specific gravity of any particular fuel at any particular operating temperature by a voltage difference $V_a - V_m$ where

(a) V_a at 20° C represents the specific gravity at 20° C of the densest fuel (0·86).

(b) The temperature coefficient of V_a represents the average slope of all the curves.

(c) V_m represents the difference between the specific gravity of the densest fuel at 20° C and that of the fuel actually in use at 20° C.

If an aircraft always uses the same type of fuel, greater accuracy can be obtained, of course, by choosing the temperature coefficient of V_a for that particular fuel.

5.2. *Input Diode Pump*

Figure 8 shows the circuit of the diode pump which multiplies the voltage difference $V_a - V_m$ by the input frequency. The square wave signal from the demodulator drives a transistor switch on and off at this frequency (2ω). Whenever the switch is open (reset condition), capacitor C1 charges to a voltage $60 - V_m - V_{D1}$ where V_{D1} is the voltage remaining across the diode D1 at the end of the period. (The reason for this is discussed in Section 5.3.)

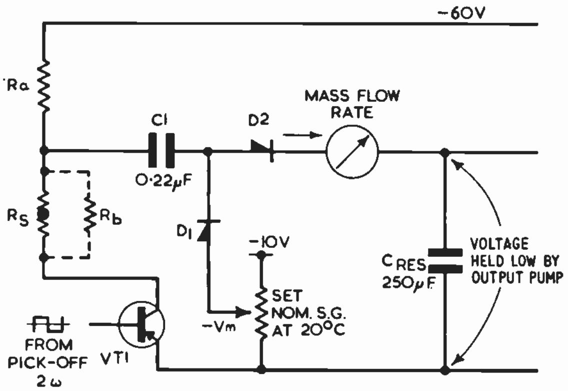


Fig. 8. Circuit of input pump.

Whenever the switch is closed (pump condition), C1 discharges to a voltage $60\left(\frac{R_s}{R_a+R_s}\right) + V_{D2}$ where V_{D2} is the voltage remaining across D2. The bottoming voltage of the transistor switch is neglected, since it is both small and stable. The voltage across the reservoir capacitor is also neglected since it is kept negligibly small by the circuits which follow. Hence the voltage range or stroke through which C1 charges and discharges is

$$\text{pump stroke} = 60\left(\frac{R_a}{R_a+R_s}\right) - (V_{D1}+V_{D2}) - V_m \qquad \text{......(1)}$$

and the (pulsating) direct current delivered to the rate-meter is

$$I = 2\omega C_1\left[60\left(\frac{R_a}{R_a+R_s}\right) - (V_{D1}+V_{D2}) - V_m\right] \qquad \text{......(2)}$$

i.e.
$$I = 2\omega C_1(V_a - V_m) \qquad \text{......(3)}$$

where
$$V_a = 60\left(\frac{R_a}{R_a+R_s}\right) - (V_{D1}+V_{D2}) \qquad \text{......(4)}$$

Thus the circuit behaves as required, provided that V_m is properly adjusted, R_s is the correct nominal value and temperature coefficient, variations in V_{D1} and V_{D2} can be compensated, and variations in the value of C1 and the 60 V supply are either small or compensated. These points will now be considered.

5.3. *Errors Due to Diodes*

The time-constants formed by the circuit resistors and C1 in both the pump and reset conditions are chosen to be less than one seventh of the time available at maximum rotor speed. Were it not for the diodes, therefore, C1 would charge and discharge to within 0·1% of the theoretical limits, whatever the flow rate.

In fact, however, the forward resistance of the relevant diode rises rapidly as the current falls, and eventually takes complete control of the decay operation. It is shown in Appendix 1 that this occurs after a time $t = 7RC$ approximately, at which point the diode supports about 0·5 to 0·6 V. From then on, the voltage across the diode falls extremely slowly, following a logarithmic law. For typical silicon diodes the fall is about 100 mV for every tenfold increase in time, and it is impossible in practice to allow sufficient time for the voltage to drop to a negligible level.

Over a 30 to 1 range of flow rates, the change in the voltage drop $V_{D1}+V_{D2}$ across two diodes is about 300 mV, which, if not compensated, would cause an appreciable error. However, the same effect occurs in the output pump and, provided that the two pumps operate for most of the time at the same frequency, with the same mark/space ratio, and through approximately the same voltage range, compensation is almost perfect in respect of the indication of total mass. The rate indication is not compensated, but this is less important.

The diode characteristics and hence the voltage drop are affected slightly by ambient temperature also, but again there is almost complete compensation.

5.4. *Temperature Coefficient of V_a*

Since specific gravity falls with rising temperature, V_a must also fall, and R_s is required to have a positive temperature coefficient. A silicon resistor meets this requirement, and also provides the relatively large output which is required. Measurements over nearly two years (Appendix 2) show that both the absolute value and the temperature coefficient of this type of device are sufficiently stable for the purpose.

As Fig. 7 shows, V_a is required to vary linearly with temperature, and it can be seen from eqn. (4) that this will occur only if the R_s-temperature relationship is non-linear (to a degree determined by the ratio R_s/R_a). Fortunately this happens to be the case for silicon resistors and, contrary to what even the most sanguine engineer might expect, the curvature is in the right direction. Thus it is possible to obtain the desired linear law by choosing a suitable value for the ratio R_s/R_a.

In practice this cannot be done if only two resistors are used, since R_s/R_a has to be adjusted to give the desired nominal value of V_a. Hence a third resistor R_b, shown dotted in Fig. 8, is connected in shunt with R_s. The procedure for constructing a probe unit is as follows.

The nominal value and temperature coefficient of R_s are measured. R_a and R_b are calculated to give precisely the desired nominal value and temperature coefficient for V_a, and approximately the right amount of correction for curvature. The coils then are wound and potted in the probe together with the pick-off. The spread in R_s and its coefficient make exact curvature correction impossible, but the final error is negligible.

Since V_a is less than the output of the probe circuit by an amount $V_{D1}+V_{D2}$, this quantity affects the coefficient which the probe circuit must have. It is quite sufficient, however, to assume a nominal value of 1 V in the calculation.

It is worth noting that routine checking of the voltage divider (if required) is a simple matter, involving the measurement of a resistance ratio only, and not requiring an accurate voltage source. Slight self-heating effects make it necessary to apply a voltage which gives approximately the same heating effect (2 deg C) as obtained in normal use: this is $60/\sqrt{2}$ volts if the test voltage is continuous.

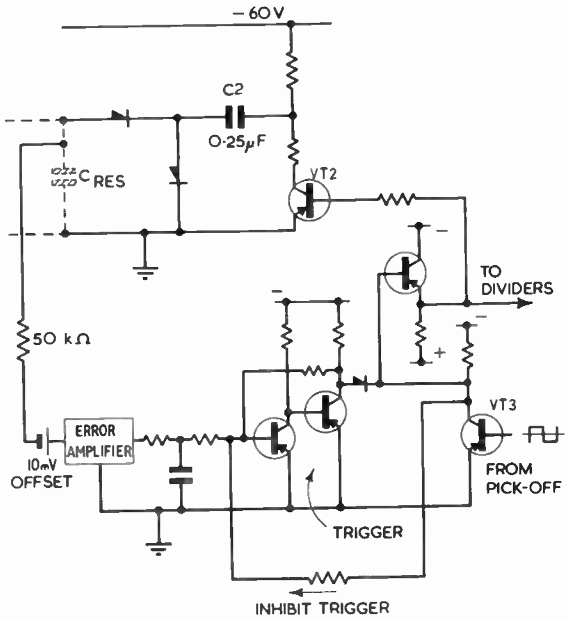


Fig. 9. Output pump and drive circuits.

6. Computation of Total Mass Flow

All the current flowing into the reservoir is extracted again by the output diode pump shown in Fig. 9.

This pump is supplied from the 60 V h.t. rail, and its capacitor C2 is slightly larger than the corresponding capacitor C1 in the input pump. It is normally driven synchronously and in anti-phase with the input pump so that both pumps re-set together and pump together. Hence the upper plate of the reservoir capacitor is driven slightly negative at the start of each pumping period, as shown in Fig. 10.

A simple chopper-type d.c. amplifier monitors the voltage across the reservoir capacitor, and has an offset voltage of about 10 mV introduced in series with its input. Thus when the upper plate reaches a level of about −10 mV, the output of the amplifier changes polarity and operates the trigger circuit which follows it. This clamps the collector of the phasereversing transistor VT3 to earth, and holds transistor VT2 off. Thus the output pump is held in the 'just-having-pumped' condition while the input pump re-sets and pumps again. As a result, the upper plate of the reservoir capacitor is driven sharply positive by about 20 mV, the trigger releases, and the circuits operate as before with both pumps working together.

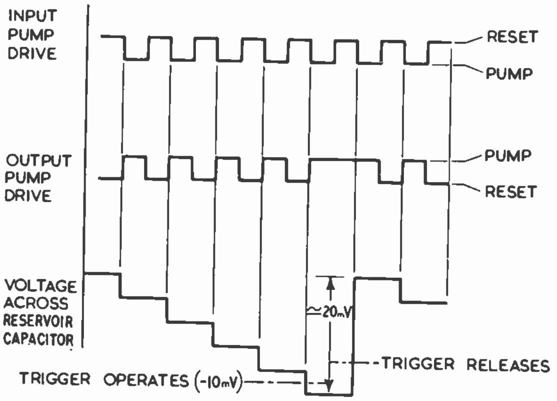


Fig. 10. Waveforms showing operation of pumps and trigger.

The net result is that the two pumps work in synchronism for most of the time, but every now and again the output pump is inhibited for one operation. The frequency at which this occurs depends on the specific gravity of the fuel.

Since the output pump has a constant stroke, the number of operations which it performs represents the total mass of fuel which has passed through the meter. This is displayed on a counter driven through a binary divider from the pulses applied to VT2. The intermittent nature of this train is no disadvantage.

Practical precautions which must be observed in this circuit include the use of a reservoir capacitor with suitably high internal shunt resistance, and an amplifier which draws negligible current. These requirements are not too difficult to meet. The full-scale rate current is 500 μA, hence if 0·1% accuracy is required at 1/10 full flow, the lost current must not exceed about 0·05 μA. The leakage current of a 250 μF tantalum capacitor with 10 mV applied to it is much less than this figure, and in fact it is only the unbalance between the positive and negative excursions of this voltage which matter. A similar argument applies to the current drawn by the amplifier.

One further precaution is necessary. Because of the smoothing circuit which follows the amplifier, the voltage applied to the trigger rises exponentially. Occasionally this voltage may not reach the triggering level during a pump half cycle but does so at some time during the subsequent reset period. This might cause VT2 to switch into the pump condition after a very short period in the reset condition, and is avoided by applying a small current from VT3 which inhibits operation of the trigger during a re-set period.

7. Power Supply Circuits

The system used is shown in Fig. 11 and is similar in principle to that used in the d.c. amplifier described in Ref. 3.

The 21–29 V d.c. input is first stabilized at 16 V ±1% by a transistor and zener diode regulating circuit. It is then converted into a 2 kc/s square wave by a relaxation oscillator containing two transformers, one of which saturates and defines the frequency. The other transformer feeds rectifying circuits delivering −60 V, +12 V and −12 V d.c., and also supplies 2 kc/s to the pick-off and the input and output choppers in the d.c. amplifier. The high stability of all these supplies simplifies the design of the main circuits.

Protection is included against the 80 V transients which occur in aircraft supplies. The counter is supplied directly from the unregulated d.c. supply, since its consumption is relatively large and intermittent. Total consumption of the system is approximately 0·25 A.

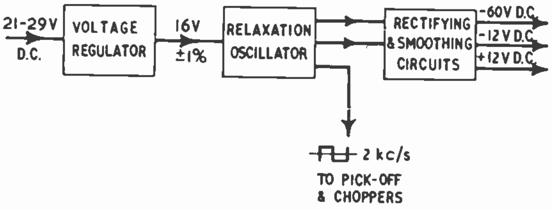


Fig. 11. Power supply system.

8. Performance

The temperature-sensitive voltage divider can be calibrated to have an output to input voltage ratio within ±0·1% of the design value and a temperature coefficient which is sufficiently accurate to cause negligible error. Hence probes can be treated as interchangeable items.

The most important characteristic of the receiver unit is the ratio of output pulse rate to input pulse rate. This can be readily adjusted initially within ±0·1% with the probe held at 20° C and the "set-nominal-s.g." control set at, for instance, 0·8. The effect on this ratio of variations in receiver temperature between −20° and +70° C are less than ±0·3%, and the effect of supply variations between 21 V and 29 V is not measurable. The ratio also does not vary by more than ±0·1% over a flow range of 10 to 1.

All the semiconductors (except one) are silicon, and the circuits are designed for operation in ambient temperatures between −20° and +70° C. The normal operating range of fuel temperature is −20° to +50° C but wider ranges can be accommodated if necessary. Mechanical construction of all units is to aircraft specifications.

9. Conclusions

A system[4] has been described which computes the mass flow rate and the total mass flow of a liquid from the signals delivered by a simple turbine-type flowmeter fitted with a combined magnetic pick-off and temperature sensitive probe. Manual setting of the nominal specific gravity is required, and the relation between specific gravity and temperature must be linear. The system is simple, accurate, and inexpensive.

10. Acknowledgments

The authors thank the Directors of George Kent Ltd. for permission to publish this paper.

11. References

1. G. S. T. Hudson, "Electronic aspects of a true-mass flowmeter system", *Brit. Commun. & Electronics*, 9, No. 4, p. 283, April 1962.
2. Brit. Pat. Application No. 15518/61.
3. I. C. Hutcheon and G. B. Marson, "Solid-state electronic instruments for measurement and control", *Instrum. Engr*, 3, No. 3, p. 49, April 1961.
4. Brit. Pat. Application No. 33613/62.

12. Appendix 1

Voltage Decay in Diode Pump Circuit

The forward characteristic of a silicon diode is represented approximately by

$$i = A\,e^{BV_d} \qquad \ldots\ldots(5)$$

where i is the forward current in amperes

$A \simeq 10^{-10}$

$B \simeq 23$

V_d is the forward voltage drop in volts, and exceeds 0·1.

Combining this expression with the differential equation for the circuit shown in Fig. 12, we find that the time, t, taken for the current to decay from an initial value i_0 at time $t = 0$ (when the switch is closed) to any value i, is given by

$$t = RC\log\left(\frac{i_0}{i}\right) + \frac{C}{B}\left(\frac{1}{i} - \frac{1}{i_0}\right) \qquad \ldots\ldots(6)$$

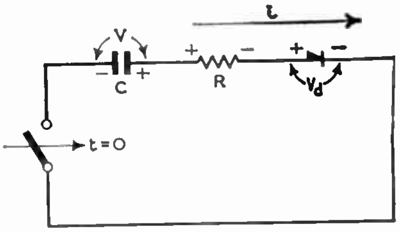


Fig. 12. Circuit with C, R, and diode.

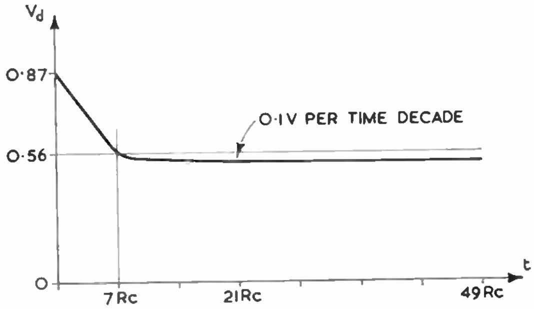


Fig. 13. Voltage across diode. $V_0 = 50$ V, $R = 1$ kΩ, $i_0 = 50$ mA.

Initially, when iR is large, the first term predominates and describes the normal exponential decay which then occurs. Neglecting the second term and combining the remainder of eqn. (6) with eqn. (5) we find that the voltage across the diode falls linearly with time during this period according to the expression

$$V_d = \frac{1}{B}\left[\log\left(\frac{i_0}{A}\right) - \frac{t}{RC}\right] \qquad \ldots\ldots(7)$$

As the current falls still further, the slope resistance of the diode rises until it takes charge and the second term in eqn. (6) predominates. The transition between the two states occurs approximately when the slope resistance is equal to R, i.e. when

$$R = \frac{\mathrm{d}V_d}{\mathrm{d}i} = \frac{1}{Bi} \qquad \ldots\ldots(8)$$

and $$t = RC\log BRi_0 \simeq RC\log V_0 \qquad \ldots\ldots(9)$$

Thus if $V_0 = 50$ V, for example, the transition occurs when $t \simeq 7RC$; eqn. (7) shows that the diode then supports about 0·56 V. During the subsequent period, the first term in eqn. (6) can be neglected and, since $i_0 \gg i$, we find

$$t = \frac{C}{B}\frac{1}{i} = \frac{C}{AB}\,\mathrm{e}^{-BV_d} \qquad \ldots\ldots(10)$$

which implies that V_d falls logarithmically by approximately 100 mV for every tenfold increase in time. Figure 13 shows the decay of V_d during the whole period for $V_0 = 50$ V, $R = 1000\,\Omega$, and $i_0 = 50$ mA.

Experimental results agree closely with the above deductions.

13. Appendix 2

Long Term Stability of Silicon Resistors

Twenty resistors of nominal value 1 kΩ were checked at intervals over a period of 20 months. Ten were temperature cycled between −20° C and +20° C:

ten were cycled between +20° C and +60° C. In all cases the measured value of resistance at 20° C remained within ±0·5% of the nominal value throughout the tests and there was little evidence of drift in one direction. Since the nominal temperature coefficient is +0·8%/deg C, the spread in results corresponds to ±⅔ deg C approximately. Some of this spread was probably due to small errors in measurement. No change in temperature coefficient could be detected.

Similar tests by the manufacturers on 60 units of three nominal values, stored at 25° C for 12 months gave maximum excursions of ±0·6% in nominal value and an average drift of −0·33%.

Manuscript received by the Institution on 14th December 1962 (Paper No. 825).

DISCUSSION

(Under the chairmanship of Mr. J. R. Halsall)

Mr. W. H. Topham: The system described in the paper cannot truly be described as a "mass meter" because an independent measurement of density (i.e. specific gravity) must be made, and fed into the system. The equipment, although most elegant and suited for aircraft use, is comparable with temperature compensated pd meters, already widely used in industry.

The petroleum industry has a need for true mass meters, since specific gravities are generally not known accurately, and may vary during a particular loading or transfer. Mass meters such as the gyroscopic type are probably too complex for industrial use, and have been reported to be unreliable.

The authors (*in reply*): We agree with Mr. Topham that the system is, in fact, inferential. In many aircraft applications, however, the disadvantage of having to measure the density at some nominal temperature is far outweighed in our opinion by the relative simplicity, reliability, and low cost of this type of meter.

Mr. P. Wood: The approach proposed by the authors pre-supposes a knowledge of the specific gravity of the fuel. In practice this can vary by as much as ±10% and fuels of different specific gravities can be included in a tank at any one time. Estimates of mean specific gravity are liable to error due to stratification etc.

Have the authors considered measuring the permittivity of the fuel to provide an indication of its specific gravity? A large number of measurements on fuels of different types has shown a close relationship between permittivity and specific gravity.

The authors (*in reply*): We have not yet been able to investigate the possibilities of permittivity measurement, but are grateful to Mr. Wood for his suggestion.

Mr. Topham: I consider it likely that, for a particular type of aircraft using a particular type of fuel, a more probable figure for specific gravity variation would be 1%, i.e. 0·008 on a s.g. of 0·800.

The authors (*in reply*): We do not agree that density variations in flight would normally be as small as 1%. Our information from the aircraft industry is that 4% to 6% is not uncommon.

Mr. R. E. Ross: In testing large liquid fuelled rocket engines it is necessary to measure flow rates of various liquids, e.g. water, kerosene and liquid oxygen. The measurement of mass flow of liquid oxygen presents many difficulties. We have experienced disappointingly short bearing lives in this fluid. Bearings also appear to have a poor life in simpler liquids such as water and kerosene; that is, poorer than the manufacturers would have us believe. I consider that much more effort is needed in bearing design before the primary measuring element can match up to the associated electronics in accuracy, range and reliability.

Mr. Topham quoted recent experience of total failure of a turbine meter rotor after only 16 hours of intermittent operation (totalized) at maximum revolutions. This was an imported meter of U.S. origin.

The authors (*in reply*): Both Mr. Ross and Mr. Topham mention the problem of bearing life, and we agree that this has not yet been solved for industrial applications, where 10 000 hours might be regarded as an absolute minimum. Our meter was designed to operate at very low rotor speeds, and we believe this contributes to its 3000-hour life which is quite adequate for aircraft applications. It is hoped to extend this life in the future.

U.D.C. 621.382.3.072.3

High Resistance Transistor Circuits

By

H. C. BERTOYA

(*Associate Member*)†

Summary: The paper reviews the way in which transistors may be used to increase circuit resistance. The subject is discussed under three main headings: the use of the transistor itself as a high resistance; the increase of the effective value of a resistor; and the design of signal transmission circuits possessing high input resistance. Practical examples are given, together with performance data. All the circuits employ silicon transistors mainly of the low-cost industrial type.

1. Introduction

In addition to their function as signal amplifiers, it is well known that active elements such as transistors may be used to increase or decrease the effective value of an impedance. It is proposed in this paper to discuss this phenomenon. The discussion will be confined to circuits which increase a given circuit impedance and the term 'impedance' will be further restricted to imply simply 'resistance'. Practical examples of circuits are described which illustrate the principles involved.

Apart from the straightforward increase of resistance, an important application is the simultaneous presentation by a circuit element of a high dynamic (a.c.) resistance and a low steady-state (d.c.) resistance. This phenomenon is of particular value in low distortion class-A circuits when the total alternating signal swing must not extend beyond the linear regions of the active element. This requires a large standing current developing the *smallest* possible standing voltage (thus implying low circuit impedance), on which is superimposed a small alternating current developing the *largest* possible voltage (thus implying high circuit impedance).

It should be remarked at this point that this effect could be obtained with an inductor, which also has a low d.c. and a high a.c. impedance. However, the reactance is proportional to frequency and the signal suffers a phase change; both phenomena may thus make the inductor unsuitable for many applications.

1.1. *Modes of Operation*

The ways of increasing circuit resistance will be discussed under three headings:

(a) Use of the transistor itself as a high resistance,

(b) The increase of the value of a resistance,

(c) Signal transmission devices having a high input resistance and a signal gain of unity or greater.

1.2. *Conventions adopted in Circuit Analysis*

In the circuit analysis the following conventions are adopted: E_{ac} is the voltage existing between points a and c; c is more positive than a.

$$E_{ac} = -E_{ca}, \qquad E_{ab}+E_{bc} = E_{ac}$$

Current flows out of the positive terminal of a generator. The transistor equivalent circuit is shown in Fig. 1 where the constants are r_e, r_b, r_c, β and α, $1+\beta = 1/(1-\alpha)$ and $r_c' = r_c/(1+\beta)$. Capital letters refer to external resistances R_e, R_b, etc.

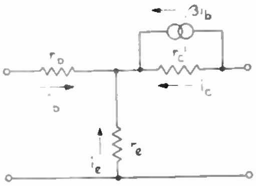


Fig. 1. Transistor low frequency equivalent circuit.

1.3. *Choice of Transistor Type*

All the practical circuits given use silicon transistors. Most of the circuits employ low-cost industrial types which are suitable for the desired frequency of operation. The tests of input resistance and other parameters were carried out at 1 kc/s.

2. The Transistor as a High Resistance

The output characteristic of a transistor allows it to have a large dynamic resistance and a low static resistance simultaneously (Fig. 2). This property enables a transistor to be used as the d.c. load for another transistor.

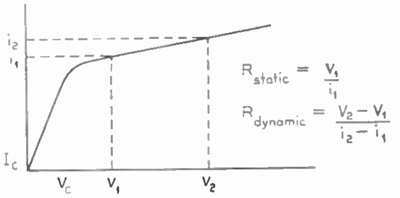


Fig. 2. Output characteristic of a transistor.

2.1. *The Transistor as an Emitter Load*

Consider the circuit of Fig. 3; an emitter follower is required to work into a load R_L via a capacitor. In order to obtain the required amplitude of a.c. signal in R_L, R_e may have to be of comparable value. A considerable amount of power is wasted in R_e and a

† British Scientific Instrument Research Association, 'Sira', South Hill, Chislehurst, Kent.

transistor must be chosen of which the power dissipation is greater than that required if the power in R_L alone were considered. The d.c. voltage across R_e is given by $i_{e1} = E_{ab}/R_e$.

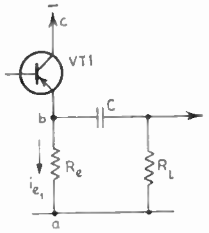


Fig. 3. Emitter-follower circuit.

The circuit may be modified as in Fig. 4. R_{b2} is adjusted to give the same d.c. conditions as for those given in Fig. 3, that is $i_{e2} = i_{e1}$. The effective value of the load R_e at a.c. is now the output impedance of VT2.

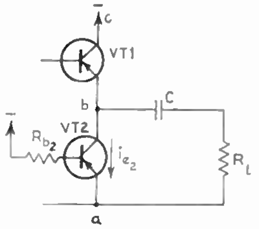


Fig. 4. Emitter-follower with grounded emitter d.c. load.

If a positive as well as a negative supply is available a preferable method is shown in Fig. 5. The value of i_{e2} is then virtually independent of variations in β of VT2. In addition the effective a.c. resistance of VT2 is higher in this grounded base configuration than it is in the grounded emitter configuration of Fig. 4.

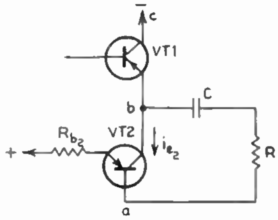


Fig. 5. Emitter-follower with grounded base d.c. load.

A further consideration regarding the choice of configuration depends on the method of biasing VT1. If the bias circuit is that of Fig. 6, then the arrangement of Fig. 4 is satisfactory. Changes of current through VT2 will not affect the voltage at b since this voltage depends largely on the voltage at d. This presumes that the base current of VT1 is small in comparison with the current through the chain R_1, R_2.

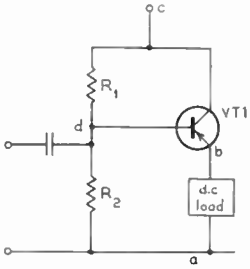


Fig. 6. Constant voltage bias circuit.

If the bias circuit is that of Fig. 7 then the best arrangement is that of Fig. 5. The voltage at b now depends on the current through VT2, since a fraction $1/\beta$ will flow through the bias resistor R_{b1}. The voltage at b will be given approximately by

$$E_{ab} = E_{ac} - \frac{i_e R_{b1}}{\beta_1}$$

These circuits only avoid the use of a more powerful transistor by the use of two lower-powered ones, but it does mean that in a given circuit more transistors of identical type may be employed. In addition, the difference in power dissipation between different types of transistor is very large—three types of Texas *npn* transistor, for example, dissipate 0·15, 4 and 37·5 W at 25° C. Exceeding the maximum rating by only a small percentage means the use of a type whose capabilities are far in excess of requirements.

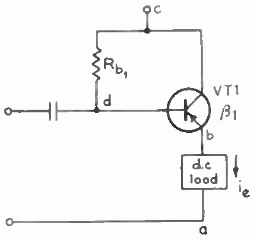


Fig. 7. Constant current bias circuit.

2.2. *The Transistor as a Collector Load*

npn and *pnp* transistors can be connected as in Fig. 8, so that VT2 forms the load for VT1. The arrangement is not satisfactory since the constant-current devices are placed back-to-back and the voltage at b is indeterminate. The arrangement of Figs. 4 and

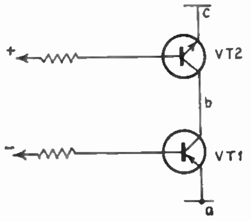


Fig. 8. Transistor as a collector load.

5 is more satisfactory due to the fact that the two constant-current generators (VT2) feed into a constant-voltage point (Fig. 6) or into an effective load R_{b1} (Fig. 7) which is in parallel with the high output impedance of VT1.

2.3. *Transistor Output Resistance*

Both the methods described so far involve the transistor output resistance. This resistance may be determined from the actual and equivalent circuit of Fig. 9, where the input to the transistor is short-circuited, a generator is placed between ground and collector and the output resistance R_o is found from the ratio E_{ac}/i_g.

The circuit equations are:

$$i_g = i_c + \beta i_b$$

$$i_b(1+\beta) + i_c + i_e = 0$$

$$i_b(R_b + r_b) = i_e(R_e + r_e)$$

$$E_{ac} = E_{ab} + E_{bc} = -i_b(r_b + R_b) + i_c r_c'$$

if we find i_e and i_c in terms of i_g then we may express the ratio

$$\frac{E_{ac}}{i_g} = R_o = \frac{r_b + R_b + r_c'\left(1 + \beta + \dfrac{r_b + R_b}{r_e + R_e}\right)}{1 + \dfrac{r_b + R_b}{r_e + R_e}} \qquad \ldots\ldots(1)$$

(i) suppose the source impedance is very high ($R_b \rightarrow \infty$)

then $$R_o = r_e + R_e + r_c/(1+\beta) \qquad \ldots\ldots(2)$$

(ii) suppose the source impedance is very low ($R_b \rightarrow 0$)

then $$R_o = \frac{r_b + r_c\left(1 + \dfrac{r_b}{(r_e + R_e)(1+\beta)}\right)}{1 + \dfrac{r_b}{r_e + R_e}} \qquad \ldots\ldots(3)$$

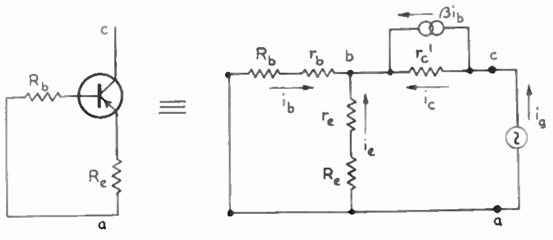


Fig. 9. Circuits for determination of output resistance.

Typical values for the constants are $r_b = 500\,\Omega$, $r_b = 20r_e$, $\beta = 30$, $r_c = 300 \times 10^3\,\Omega$.

In eqn. (3) the value of R_o depends to a large extent on the ratio $r_b/(r_e + R_e)$. In this case, when $R_e = 0$, the numerator is divided by approximately 20. If the resistor R_e is inserted the ratio can be reduced and the output resistance increased accordingly.

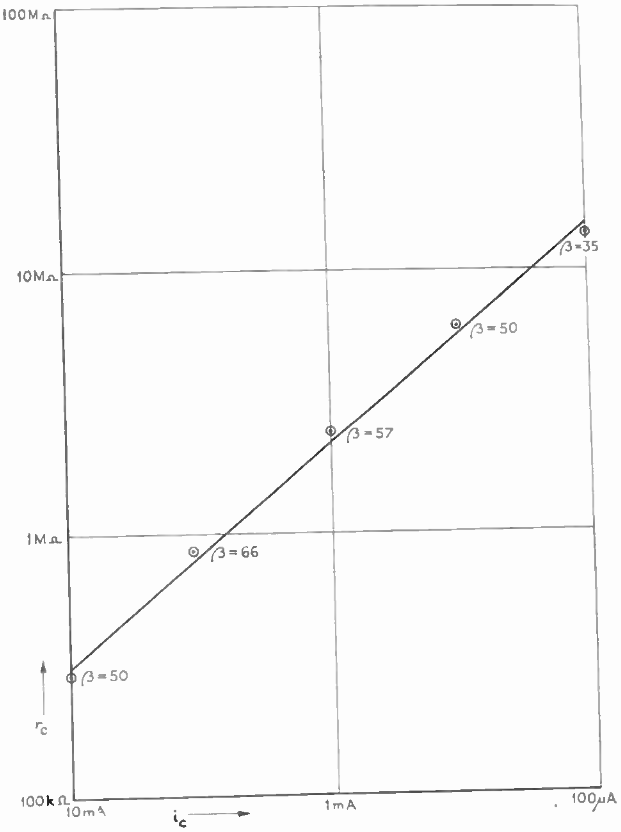


Fig. 10. Variation of r_c with I_c.

In eqn. (2) the output resistance is increased simply by the addition of R_e to $r_c/(1+\beta)$.

The value of r_c for a given transistor varies greatly with standing current and this variation is shown in Fig. 10 for a particular BCZ11 transistor. The value of β is given alongside the values of r_c.

3. Increase of the Effective Value of a Resistance

Suppose we have a resistor R. We may deduce its value by applying a voltage E_{ab} across it and measuring i, the current through it. Suppose now we place the resistor in the circuit of Fig. 11.

Now $$E_{ab} + E_{ca} = iR$$

suppose $$E_{ac} = \gamma E_{ab}$$

then $$E_{ab}(1-\gamma) = iR$$

Therefore $$R\,(\text{effective}) = \frac{R}{1-\gamma} \qquad \ldots\ldots(4)$$

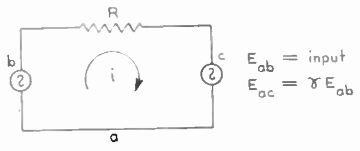


Fig. 11. Increase of effective value of R.

3.1. *Application of Method*

Two applications are described. The first is useful in emitter-follower circuits. If we examine the emitter-follower circuit of Fig. 18 we see that input resistance is shunted by the biasing resistor R_b. The effective resistance of R_b may be increased by connecting the resistor in the manner shown in Fig. 19. The circuit is discussed further in Section 4.4.

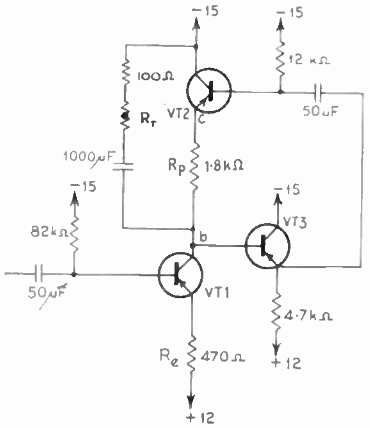


Fig. 12. Application of principle of increase of effective resistance of R. All transistors are Mullard BCZ11. R_T is an S.T.C. thermistor, type R14.

Another example is shown in Fig. 12, where it was required to increase the effective value of the resistor R_p. The problem in this case was that it was decided to use a directly-heated thermistor as the control element in an a.g.c. amplifier.† The best basic arrangement is shown in Fig. 13.

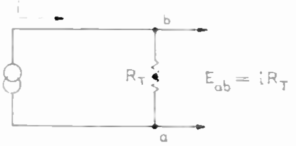


Fig. 13. Basic thermistor circuit.

If *i* flows from a constant-current source, the output E_{ab} is proportional to iR_t (where R_t is the thermistor resistance). The simplicity of this arrangement is spoilt by the fact that it is not permissible to pass d.c. through the thermistor. The arrangement must then take the form of Fig. 14 where R_p must be many times the maximum value of R_t, so that the output voltage will depend primarily on R_t. Sufficient d.c. must flow through R_p in order to allow the required undistorted alternating signal E_{ab} across the minimum value of R_t. For the required current a large value of R_p would require a very large supply voltage.

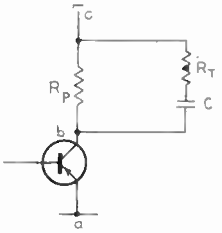


Fig. 14. Basic practical thermistor circuit.

The solution to this problem is found by increasing the effective value of R_p, and the circuit is given in Fig. 12, where the points b and c correspond to those shown in Fig. 11. The output impedance of VT1 was not high enough at the standing current required and so the impedance was increased by the resistor R_e as discussed in Section 2.3.

4. Transmission Devices with High Input Resistance

4.1. *Base Current Bias Arrangements*

The circuits in this section rely on the selection of a base current bias resistor in order to give the correct value of emitter and collector currents. Consideration was given to this point because of the desirability of specifying a given resistor rather than selecting according to the value of β.

However, even simple biasing systems which tolerate variations in β add considerably to circuit complexity since they seriously reduce the input resistance. It was decided, therefore, to tolerate resistor selection (to nearest value in 10% range). This is normally a simple matter and since the transistor characteristics are stable with time, the selected value does not require subsequent alteration.

4.2. *Measurement of Input Resistance*

The devices described should be used as high input resistance circuits fed from low resistance sources. If the source resistance is very high it would probably be better to treat it as a constant-current source and use a current amplifier. Nevertheless the circuits will all perform satisfactorily when fed from medium to high resistances; the measurement of input resistance was as follows.

The measuring circuit took the form of Fig. 15. The voltage E_{ab} was measured at the output of the signal

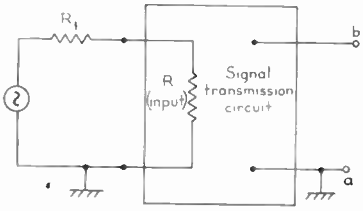


Fig. 15. Measurement of input resistance.

† H. C. Bertoya, "An a.g.c. circuit using a thermistor and transistors", *Electronic Engineering*, 35, p. 236, April 1963.

transmission circuit in order to obviate the effect of voltmeter input resistance.

Let E_{ab1} be the value of output voltage with R1 short-circuited.

Let E_{ab2} be the value with R1 in circuit.

Then
$$R_{in} = \frac{E_{ab2} R_1}{E_{ab1} - E_{ab2}}$$

The value of R1 was chosen to be about a tenth of the expected input resistance. In particular the circuit of Fig. 22 was measured with R1 = 10 kΩ and that of Fig. 27 with R1 = 100 kΩ.

4.3. *The Simple Emitter Follower*

First let us consider the simple emitter follower and its equivalent circuit. This is shown in Fig. 16.

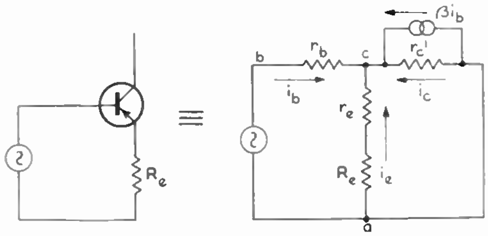


Fig. 16. Equivalent circuit of emitter follower.

The circuit equations are

$$E_{ab} = E_{ac} + E_{cb}$$
$$i_b(1+\beta) + i_e + i_c = 0$$
$$i_b(1+\beta) = E_{ac}\left(\frac{1}{r_e + R_e} + \frac{1}{r_c'}\right)$$
$$E_{cb} = i_b r_b$$

Carrying out the various substitutions yields the expression

$$E_{ab}/i_b = R_{in} = r_b + \frac{(r_e + R_e) r_c}{r_c/(1+\beta) + r_e + R_e} \quad ......(5)$$

The equation may be used with good accuracy by assuming r_b, r_e to be very small. This gives eqn. (6).

$$R_{in} \simeq \frac{R_e r_c}{r_c/(1+\beta) + R_e} \quad ......(6)$$

A further simplification may be made if r_c is large in comparison with R_e and if β is $\gg 1$. Equation (6) then reduces to

$$R_{in} \simeq R_e \beta \quad ......(7)$$

This serves as a useful "rule of thumb".

The maximum input resistance is obtained when either R_e or β approach infinity. In this case

$$R_{in\,(max)} = r_b + r_c \simeq r_c$$

Normally, however, the input resistance falls far short of r_c. Table 1 gives an idea of the order of input resistance for the simple emitter-follower of Fig. 16. The circuit was operated with various values of standing current, the transistor being a Mullard BCZ11. The Table neglects the shunting effect of bias resistors. Measurements were carried out at 1 kc/s.

Table 1

Standing current	r_c	β	Emitter load R_e	Input resistance	
				Measured	Calculated from eqn. (6)
10 mA	176 kΩ	31	270 Ω	9·4 kΩ	8·6 kΩ
1 mA	2·5 MΩ	39·2	2·7 kΩ	100 kΩ	104 kΩ
100 μA	13·0 MΩ	25	27 kΩ	650 kΩ	665 kΩ

4.4. *Emitter-follower Stages in Cascade*

A greater transformation ratio of input to output resistance can be obtained by putting two or more such stages in cascade as shown in Fig. 17.

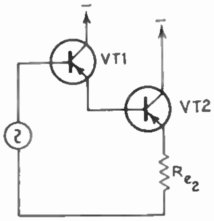


Fig. 17. Emitter followers in cascade.

The equation for this circuit may be obtained from eqn. (5), where R$_e$ is replaced by the input resistance of the following stage.

If we let $r_b \rightarrow 0$, and $\beta \gg 1$, $R_e \gg r_e$, then the input resistance of one stage alone is

$$R_{in} \simeq \frac{R_e r_c}{r_c/(1+\beta) + R_e} \quad ......(6)$$

If we now replace R$_e$ by the input resistance of VT2 and assume $\beta \gg 1$,

$$R_{in} \simeq \frac{\left(\dfrac{R_{e2} r_{c2}}{r_{c2}/\beta_2 + R_{e2}}\right) r_c}{r_c/\beta + \dfrac{R_{e2} r_{c2}}{r_{c2}/\beta_2 + R_{e2}}} \quad ......(8)$$

Rearranging eqn. (8) and putting in the form of eqn. (6) gives

$$R_{in} \simeq \frac{R_{e2} r_c}{r_c/\beta\beta_2 + R_{e2}(r_c/\beta r_{c2} + 1)} \quad ......(9)$$

The increase in input resistance will be brought about more by the fact that the new value of r_c is larger, since VT1 passes less current than VT2, than by the new factor $r_c/\beta\beta_2$. However R_{in} may still be

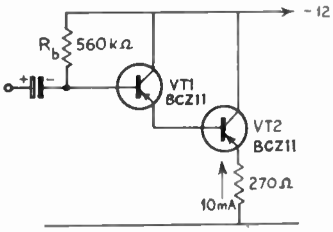


Fig. 18. Practical cascade emitter follower circuit.

much less than r_c. In the example shown (Fig. 18) the value of r_c was 6·6 MΩ and the input resistance (neglecting R_b) was about 300 kΩ. VT2 operated under the same conditions as it did in the first example quoted in Table 1, when the input resistance was 9·4 kΩ. The resistor R_b reduces the total input resistance to 195 kΩ.

The input resistance of the above circuit may be increased by the method of Section 3. Practical examples are shown in Fig. 19. where the shunting effect of the bias resistor R_b is reduced. Figure 19(*a*) shows a circuit suitable for d.c. operation where the input resistance is 250 kΩ. Figure 19(*b*) shows a circuit suitable for a.c. operation where the input resistance is 275 kΩ.

Note that no resistor is necessary in the emitter of VT1. The standing base current in VT2 forms the emitter current of VT1.

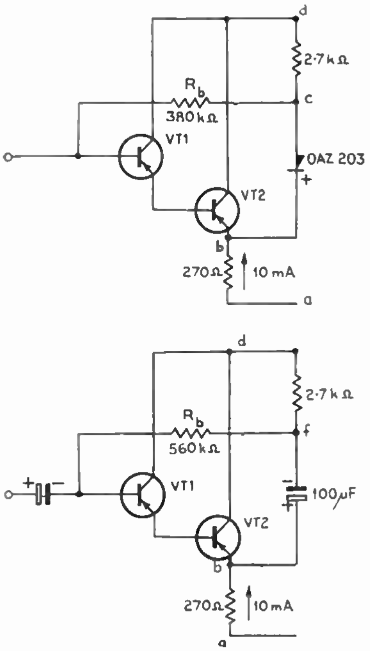


Fig. 19. Practical circuits with increase in effective value of R_b. VT1 and VT2 are Mullard BCZ11.

D.c. voltages $E_{ba} = 2·7$ V $E_{da} = 12$ V
$E_{ca} = 9$ V $E_{fa} \simeq 12$ V

Since the value of input resistance depends on both β and r_c, any correlation between the two quantities could have a bearing on transistor selection and circuit performance limits. None of the manufacturers approached on the subject were able to give information, so a batch of 19 BCZ11 transistors was measured at a standing collector current of 1·5 mA and a collector voltage of 10 V. There is no correlation.

It is possible, however, to find ways of increasing the input resistance other than by cascading stages. Such methods, based on the emitter-follower configuration, fall into two classes. Those which follow the analysis leading to eqn. (5), make the effective R_e tend to approach infinity. The maximum input resistance is then r_c. By employing feedback in the collector branch, the other class of device allows the input resistance to become greater than r_c.

4.5. *Stage with Emitter and Collector Loads*

Before proceeding to the next part of the survey it is proposed to investigate a stage with emitter and collector loads. This is a useful circuit for stages requiring a low gain or for phase splitting. The circuit and its equivalent is given in Fig. 20.

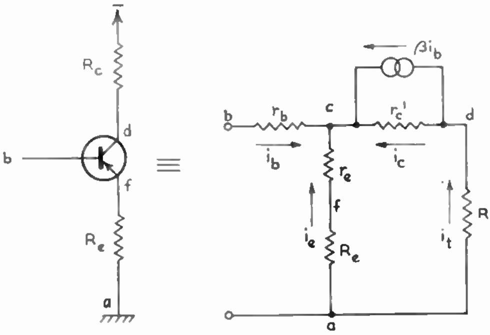


Fig. 20. Stage with emitter and collector loads.

If we note that $E_{cb} = i_b r_b$, $E_{ac} = E_{ab} + E_{bc}$ and $E_{da} = E_{dc} + E_{cb} + E_{ba}$, then the circuit equations are

$$i_b\left(1+\beta+\frac{r_b}{R_e+r_e}\right) = E_{ab}/(R_e+r_e)+E_{dc}/r_c' \quad \ldots\ldots(10)$$

$$i_b(\beta - r_b/R_c) = -E_{ab}/R_c + E_{dc}\left(\frac{1}{R_c}+\frac{1}{r_c'}\right) \quad \ldots\ldots(11)$$

This yields the result

$$E_{ab}/i_b = R_{in} = r_b + \frac{(R_e+r_e)(R_c+r_c)}{R_e+r_e+R_c+r_c/(1+\beta)} \quad \ldots\ldots(12)$$

Now suppose $r_b \to 0$ and $R_e \gg r_e$ and $R_c = KR_e$

$$R_{in} \simeq \frac{R_e(KR_e+r_c)}{R_e(1+K)+r_c/(1+\beta)} \quad \ldots\ldots(13)$$

When $K = 0$ this reduces to eqn. (6).

The gain of such a stage is given approximately by K, so long as K is small—perhaps less than 10. In a phase splitter $K = 1$. If values are inserted into equation for $K = 0$ and $K = 1$ it will be seen that there is little difference in the value of R_{in}. This is a useful property. When $K = 10$ the input resistance will be reduced, but not catastrophically.

Table 2 shows values of input resistance for different values of K. The effect of the bias resistor is not considered.

Table 2

K	E_{ad}/E_{af}	β	r_c	Input resistance	
				Measured	Calc. from eqn. (13)
0	—	52	1·85 MΩ	53·5 kΩ	51·5 kΩ
1	0·99	52	1·85 MΩ	49·5 kΩ	50 kΩ
10	9·8	43	1·04 MΩ	31 kΩ	30 kΩ

Note: The fall in β and r_c when $K = 10$ is due to the very low collector voltage existing with this load. The practical circuit used a BCZ11 and required a bias resistor of 470 kΩ, R_e was 1 kΩ and the collector current was 1·5 mA.

4.6. *Transmission Devices in which $R_{\text{in (max)}} = r_c$*

An arrangement which uses an amplifier both to increase the input resistance and to enable the network to have a gain greater than unity is shown, with its equivalent circuit, in Fig. 21.

The circuit equations are:

$$i_b(1+\beta)+i_e+i_c = 0$$

$$i_t = \beta i_b + i_c, \qquad i_f + i_K = i_e$$

$$i_f + i_K + i_b + i_t = 0$$

$$R_T = \frac{R_c R_i}{R_c + R_i}, \qquad R_o \ll R_f$$

$$i_b(1+\beta) = E_{fg}/r_e + E_{cg}/r_c'$$

$$\beta i_b = E_{ca}/R_T + E_{cg}/r_c'$$

$$0 = E_{fd}/R_f + E_{fa}/R_e + E_{fg}/r_e$$

$$i_b = E_{df}/R_f + E_{af}/R_e + E_{ac}/R_T$$

$$E_{fg} = E_{fa} + E_{ab} + E_{bg}$$

$$E_{cg} = E_{ca} + E_{ag}$$

$$E_{fd} = E_{fa} + E_{ad}$$

$$E_{ad} = -AE_{ac}$$

If the various substitutions are carried out, we may write:

$$i_b\left[1+\beta+r_b\left(\frac{1}{R_e}+\frac{1}{r_c'}\right)\right] = E_{ab}\left(\frac{1}{r_e}+\frac{1}{r_c'}\right) - \frac{E_{ac}}{r_c'} - \frac{E_{af}}{r_e} \qquad \text{......(14)}$$

$$i_b\left(\beta+\frac{r_b}{r_c'}\right) = \frac{E_{ab}}{r_c'} - E_{ac}\left(\frac{1}{R_T}+\frac{1}{r_c'}\right) \qquad \text{......(15)}$$

$$i_b\frac{r_b}{r_e} = \frac{E_{ab}}{r_e} - \frac{AE_{ac}}{R_f} - E_{af}\left(\frac{1}{R_f}+\frac{1}{R_e}+\frac{1}{r_e}\right) \qquad \text{......(16)}$$

$$i_b = E_{ac}\left(\frac{1}{R_T}+\frac{A}{R_f}\right) + E_{af}\left(\frac{1}{R_e}+\frac{1}{R_f}\right) \qquad \text{......(17)}$$

If we find E_{ac} from eqn. (15) and E_{af} from eqn. (17), and substitute in eqn. (14) we may find the ratio.

$$E_{ab}/i_b = R_{\text{in}} = \frac{\dfrac{1+\beta+r_b(1/r_e+1/r_c') + \dfrac{1}{r_e(1/R_e+1/R_f)} - \dfrac{\beta+r_b/r_c'}{r_c'(1/R_T+1/r_c')}}{1/R_T+A/R_f} + \dfrac{\beta+r_b/r_c'}{r_e(1/R_T+1/r_c')(1/R_f+1/R_e)}}{\dfrac{(1/r_e+1/r_c') - \dfrac{1}{(r_c')^2(1/R_T+1/r_c')}}{1/R_T+A/R_f} + \dfrac{1}{r_e r_c'(1/R_T+1/r_c')(1/R_f+1/R_e)}} \qquad \text{......(18)}$$

Now suppose $A \rightarrow \infty$, then $R_{\text{in}} = r_b + \beta r_c'$.

If $\beta \gg 1$ and $r_c \gg r_b$, then:

$$R_{\text{in}} \simeq r_c \qquad \text{......(19)}$$

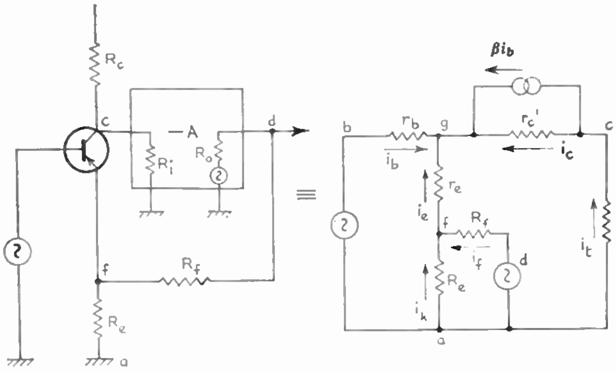


Fig. 21. Circuit with r_c as the maximum input impedance. In the equivalent circuit $R_0 \ll R_f$, $R_T = \dfrac{R_c R_i}{R_c + R_i}$

We may make a number of simplifying assumptions in eqn. (18); these are:

$$r_e \ll r'_c \qquad R_e \ll R_f$$

$$\frac{r_b}{r'_c} \ll \beta \qquad R_T \ll r'_c$$

$$\frac{R_e}{r'_c} \ll 1 \qquad r_b \ll R_e\beta$$

$$\frac{1}{r_e} \gg \frac{R_T}{(r'_c)^2} \quad \text{and} \quad \beta \gg 1$$

The equation then reduces to

$$R_{\text{in}} \simeq \frac{\beta R_e\left(\dfrac{1}{A} + \dfrac{R_T}{R_f}\right)}{\dfrac{1}{A} + \dfrac{R_T R_e}{r'_c R_f}} \qquad \ldots\ldots(20)$$

Note that if $R_f \to \infty$ the equation reduces to that of eqn. (7), and if $A \to \infty$ and $\beta \gg 1$ it reduces to that of eqn. (19). Equations (18) and (20) neglect the effect of biasing resistors.

4.6.1. An example of the method

A practical example of such a circuit is given in Fig. 22. The values of the constants were

$$\beta = 31 \qquad R_e = 2{\cdot}2\,k\Omega \qquad r'_c = 36\,k\Omega$$

$$A = 125 \qquad R_f = 27\,k\Omega \qquad R_T = 560\,\Omega$$

If these values are inserted in eqn. (20) then the calculated input resistance is about 210 kΩ. This resistance in parallel with the 180 kΩ bias resistor (R_{b1} in Fig. 22) gives a total input resistance of 97 kΩ which agrees fairly well with the measured resistance of 100 kΩ.

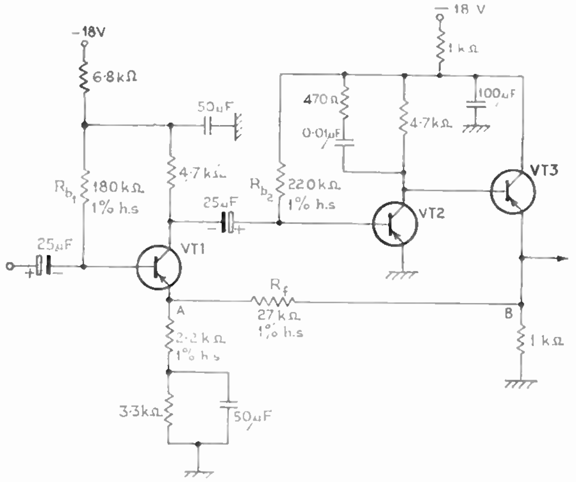


Fig. 22. Practical example of Fig. 21. VT1, VT2, VT3—all Mullard BCZ11. All resistors are 10% tolerance except where noted.

Setting up: With feedback line broken, select R_{b1} to give 5 V ± 0·25 V at point A, and R_{b2} to give 5 V ± 0·25 V at point B. Connect and adjust R_f to give overall gain of 10.

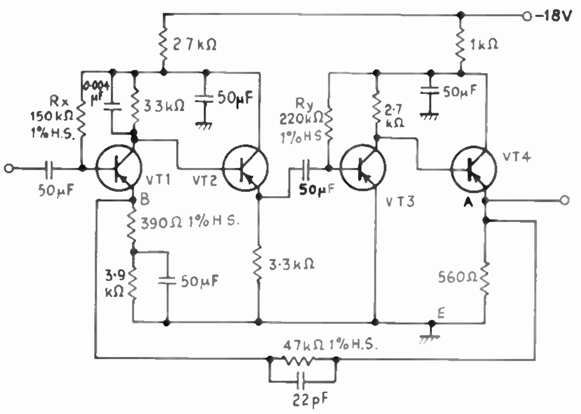


Fig. 23. Another practical example of Fig. 21. All transistors Mullard BCZ11.

Setting up: Disconnect feedback, select Rx in 10% range to give $V_{be} = 4$ V ± 10%. Select Ry to give $V_{ae} = 4$ V ± 10%. Connect feedback again.

4.6.2. Performance data for circuit of Fig. 22

Input resistance and gain vs. temperature.

Temperature °C	Input resistance	Gain
25	105·2 kΩ	10·02
35	107·7 kΩ	10·15
50	109·2 kΩ	10·17
55	107·7 kΩ	10·17
60	106·4 kΩ	10·16

Bandwidth. The response was 3 dB down at 15·5 kc/s. At 15 c/s the response was 0·967 of the maximum.

Output resistance. With the input short-circuited the output resistance was 45 Ω.

Maximum output. 1 V r.m.s.

4.6.3. Another example

An amplifier with more gain than that in the previous example was required. An extra active element was necessary so that the increase in gain was not obtained at the expense of a proportional decrease in input resistance. The amplifier circuit is given in Fig. 23, where it is seen to consist of two similar stages. The first transistor in each stage provides gain and the second transistor is used as an emitter follower. The high input resistance of the first emitter follower enables a high value of R_T to be maintained in eqn. (20). The emitter follower in the second stage enables that stage to have a high voltage gain and assists in the provision of a low amplifier output resistance.

4.6.4. Performance data for circuit of Fig. 23

Input resistance and gain vs. temperature.

Temperature °C	Input resistance	Gain
24	84 kΩ	100
29	84·5 kΩ	100
35	85 kΩ	99·7
45	85 kΩ	99·2
54	85 kΩ	99·1
62	85 kΩ	99·0
75	85 kΩ	98·9

Bandwidth. The response was 3 dB down at 10 c/s and 75 kc/s.

Output resistance. With input short circuited the output resistance was 10 Ω.

Maximum output. 2 V r.m.s.

Feedback stability margin. 15 dB.

D.c. levels. The change in direct voltages at points a and b were as follows:

Temperature °C	V_{ae}	V_{be}
27	3·59	3·74
76	2·9	4·0

The transistor β under operating conditions was:

$$\beta_1\beta_2\beta_3\beta_4 = 52\times 30\times 50\times 37 \simeq 2{\cdot}9\times 10^6$$

At 25° C the calculated input resistance was 93 kΩ which included the effect of R_{b1}

Variation of performance with β. The transistor in the first stage was replaced by one with a β of 31.

$$\beta_1\beta_2\beta_3\beta_4 = \simeq 1{\cdot}7\times 10^6$$

The input resistance and gain of the amplifier was then:

Temperature °C	Input resistance	Gain
24	49 kΩ	100
44	50 kΩ	98·5
74	51 kΩ	98·0

4.7. *Transmission Devices in which $R_{\text{in (max)}} > r_c$*

This condition is achieved by employing feedback. The effect may be described with reference to the equivalent circuit of Fig. 24 where the simple emitter-follower configuration is modified by the insertion of generators in the branches containing R_e and r_c'.

Let the voltage from the two generators be a function of the input voltage, such that

$$E_{af} = mE_{ab} \quad \text{and} \quad E_{ad} = nE_{ab}$$

then the circuit equations are

$$E_{ab} = E_{ad}+E_{dc}+E_{cb}$$
$$E_{ab}(1-m) = E_{fc}+E_{cb}$$
$$E_{ab}(1-n) = E_{dc}+E_{cb}$$
$$E_{dc} = -i_c r_c'$$
$$E_{fc} = -i_e(r_e+R_e)$$
$$E_{af}+E_{fc} = E_{ad}+E_{dc}$$
$$i_b(1+\beta)+i_e+i_c = 0$$

Carrying out the various substitutions yields

$$\frac{E_{ab}}{i_b} = R_{\text{in}} = \frac{r_b(r_e+R_e+r_c')+r_c'(1+\beta)(r_e+R_e)}{(1-n)(r_e+R_e+r_c')+r_c'(n-m)} \quad \ldots\ldots(21)$$

it is seen that if $n = m\rightarrow 1$ then $R_{\text{in}}\rightarrow\infty$.

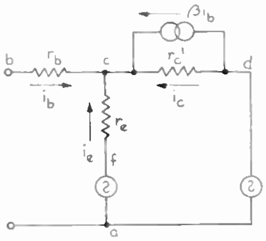


Fig. 24. Equivalent circuit in which the input resistance can be greater than r_c.

Now suppose we consider the case where $n = 1$, then substituting this value in eqn. (21) and rearranging gives:

$$R_{\text{in}} = \frac{\dfrac{(r_e+R_e)(r_b+r_c)(1+\beta)}{r_c}+r_b}{1-m} \quad \ldots\ldots(22)$$

If $R_e \gg r_e$, $r_c \gg r_b$, $R_e(1+\beta) \gg r_b$, then we may write

$$R_{\text{in}} \simeq \frac{R_e(1+\beta)}{1-m} \simeq \frac{R_e\beta}{1-m} \quad \ldots\ldots(23)$$

We can see from eqn. (23) that the generator E_{af} is performing the function, noted in Section 3, of increasing the effective value of R_e. We may further rewrite eqn. (23) with $R_{\text{(effective)}} = R_e/(1-m)$

$$R_{\text{in}} \simeq R_{e\,\text{(effective)}}\cdot\beta \quad \ldots\ldots(24)$$

Compare this equation with eqn. (7).

The input impedance is seen to vary directly with β, which distinguishes it from the arrangement noted in Section 4.6, where the input impedance (so long as β, A or R_e were high) was r_c. No advantage is to be gained in selecting either method from the point of view of repeatability of performance from transistor to transistor. The spread in r_c for a given type is about the same as the spread in β.

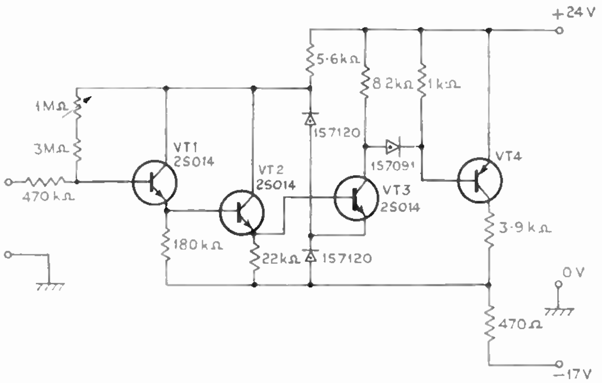


Fig. 25. High input resistance stage—practical example of Fig. 24 (Texas instruments).

It should be noted that if either n or m are greater than unity, then the input resistance can be negative and the system unstable.

4.7.1. An example

A practical example (taken from Texas Instruments Application Note No. 6) is given in Fig. 25. The amplifier consists of two series emitter followers (VT1, VT2).

The amplifier (VT3, VT4) is used both to amplify the signal and to inject voltages corresponding to E_{af} and E_{ad} in the circuit of Fig. 24.

The shunting effect of the first bias resistors (R2 and R3) is reduced in the manner described in Section 3.

The performance data given by Texas Instrument for the circuit of Fig. 25 are as follows:

Input resistance	= 200 MΩ
Input capacitance	= 2 pF
L.f. voltage gain	= 20 dB
3 dB bandwidth	= 230 kc/s
Drift with temperature referred to input	= 4 mμA/°C
Noise referred to input	= 1 mμA
Output resistance	= 30 Ω

4.7.2. Another example

Another approach to the problem was made because there was need for a device with an input resistance in the order of 10 MΩ and which need not operate at d.c. Furthermore, the circuit was to be as simple as possible and to employ low-cost industrial transistors. The schematic and equivalent circuits of this further design are given in Fig. 26.

So far as the author is aware, this is a novel circuit and differs in principle from the one given in Section 4.7.1. The high input resistance is obtained by injecting a current into the emitter of the input transistor in opposition to that which would flow due to the voltage applied to its base.

The circuit equations are

$$i_b(1+\beta)+i_c+i_e = 0$$

$$i_2+i_o = i_e \qquad E_{cb} = i_b r_b$$

$$i_2 = KE_{ad}$$

where K is a factor which will be referred to later.

$$E_{cd} = E_{ca}+E_{ad}$$

$$i_b(1+\beta)+E_{ca}\left(\frac{1}{r_c'}+\frac{1}{r_e}\right)+\frac{E_{ad}}{r_e} = 0$$

$$E_{ad} = \frac{E_{ca}(r_o r_e)}{r_e(r_o r_e K - r_e - r_o)}$$

Carrying out the various substitutions yields

$$\frac{E_{ab}}{i_b} = R_{in} = r_b + \frac{r_c(r_o+r_e)-Kr_c r_e r_o}{r_o+r_e+r_c/(1+\beta)-Kr_o[r_e+r_c/(1+\beta)]} \qquad \text{......(25)}$$

Suppose $K = 0$ then eqn. (25) reduces to

$$R_{in} = r_b + \frac{r_c(r_o+r_e)}{r_o+r_e+r_c/(1+\beta)} \qquad \text{......(26)}$$

which agrees with eqn. (5) except that the impedance r_o replaces R_e.

In the circuit shown in Fig. 6 VT2 is a grounded-base transistor fed via a resistor R_f from the output of the complete system

$$i_2 = \alpha i_{e(VT2)}$$

and $$i_{e(VT2)} = E_{ad}\Delta/R_f$$

(The gain from the emitter of VT1 to the output may be less than unity. This fractional loss is denoted by Δ.)

Thus $$i_2 = \frac{\alpha\Delta E_{ad}}{R_f} = KE_{ad}$$

Therefore $$K = \frac{\alpha\Delta}{R_f}$$

If we assume $r_o \gg r_e$ and replace K in eqn. (25) we obtain

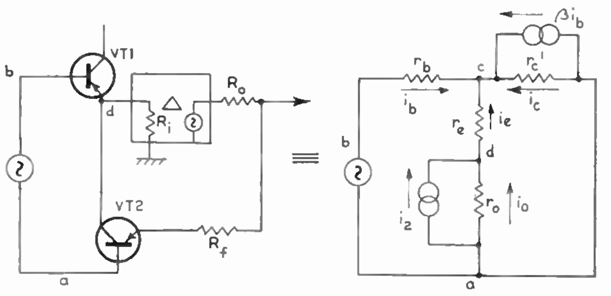


Fig. 26. Schematic and equivalent circuit giving R_{in} greater than r_c.

$R_o \ll R_f$. Output resistance of VT2 $= r_{VT2}$

$$r_o = \frac{r_{VT2}R_i}{r_{VT2}+R_i}$$

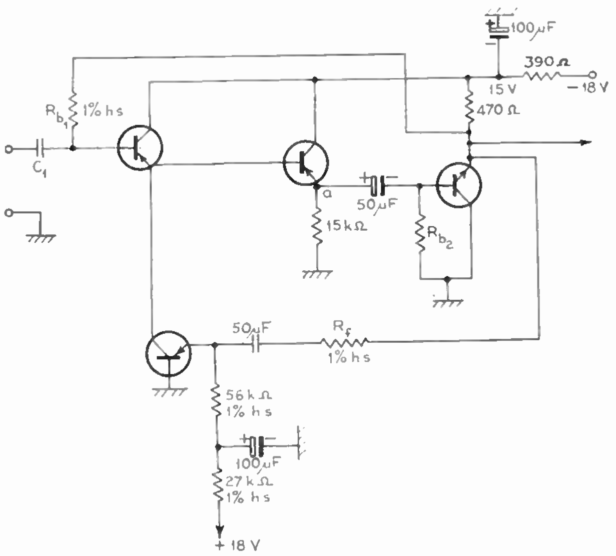


Fig. 27. Practical example of Fig. 26. VT1, VT2, VT3—Mullard BCZ11. VT4—Texas 2S701. For values of C_1, R_{b_1}, R_{b_2} and R_f, see text.

Setting up: (*a*) Select R_{b_2} (in 10% range) so that point 'b' = 12 V approx. (*b*) Select R_{b_1} (in 10% range) so that point 'a' = 6 V approx.

$$R_{\text{in}} = r_b + \frac{r_o r_c\left(1 - \dfrac{\alpha\Delta r_e}{R_f}\right)}{\dfrac{r_c}{1+\beta}\left(1 - \dfrac{\alpha\Delta r_o}{R_f}\right) + r_o} \qquad \text{......(27)}$$

Now, $\alpha.\Delta$ is of the order of unity so that $\dfrac{\alpha\Delta r_e}{R_f} \ll 1$.

If $\beta \gg 1$ and r_b is small we may re-write eqn. (27)

$$R_{\text{in}} \simeq \frac{r_c}{1 + \dfrac{r_c}{\beta}\left(\dfrac{1}{r_o} - \dfrac{\alpha\Delta}{R_f}\right)} \qquad \text{......(28)}$$

For a given value of r_c, β, etc., we may adjust R_f in order to achieve the desired value of input resistance. If $\alpha\Delta/R_f > 1/r_o$ then $R_{\text{in}} > r_c$. The curve of input resistance against R_f is similar to $1/x$. Over the final portion of the range (before the input resistance becomes negative) the rate of change of R_{in} with R_f is large. It is therefore recommended that the value of R_f be set by experiment rather than by the results of calculation.

A practical circuit is given in Fig. 27 when the value of R_{in} obtained in eqn. (28) must be modified by considering a parallel resistor of value $R_{b1}/(1-\Delta)$.

In the example given the values of the various constants are

$$r_c = 9\,\text{M}\Omega \qquad \beta = 40 \qquad \Delta = 0{\cdot}97$$
$$R_f = 180\,\text{k}\Omega \qquad r_o = 300\,\text{k}\Omega \qquad R_{b1} = 680\,\text{k}\Omega$$
$$\text{and } \alpha = 0{\cdot}975$$

If we insert these values in eqn. (28) we obtain

$$R_{\text{in}} = 16\,\text{M}\Omega.$$

This is in parallel with

$$R_{b1\,(\text{effective})} = R_{b1}/(1-\Delta) = 22{\cdot}5\,\text{M}\Omega$$

Therefore, the total input resistance is 9·4 MΩ approximately, which compares with a measured value of 9·9 MΩ.

Because of the form of the curve R_{in} vs. R_f (Fig. 28) the agreement is probably better than might be expected.

If we refer to eqn. (28) we see that if

$$\frac{r_c}{\beta}\left(\frac{1}{r_o} - \frac{\alpha\Delta}{R_f}\right) = -1 \qquad \text{......(29)}$$

the input will be infinite.

Rearranging eqn. (29) gives

$$R_f = \frac{r_o r_c \Delta\alpha}{\beta r_o + r_c} \qquad \text{......(30)}$$

Putting values into eqn. (30) gives

$$R_f = 120\,\text{k}\Omega$$

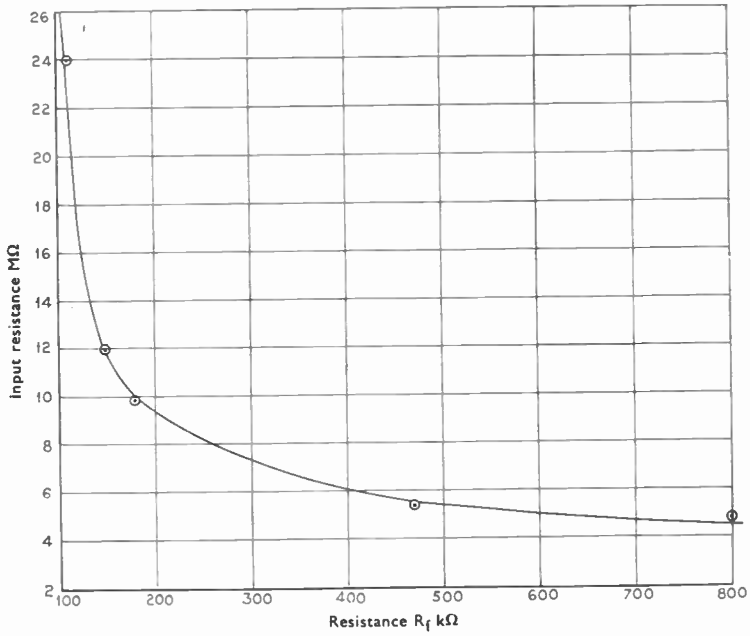


Fig. 28. Relation of R_{input} vs R_f.

In the example given the input resistance became infinite at about 100 kΩ.

4.7.3. Performance data for circuit of Fig. 27

In the example given below the value of R_{b1} = 680 kΩ, R_{b2} = 47 kΩ, R_f = 180 kΩ, and C1 was a 1 μF 150 V working paper capacitor. Care must be taken to observe that C1 is a low leakage capacitor; an electrolytic capacitor is unsuitable since the leakage alters the potential of the base of VT1.

Input resistance and gain vs. temperature.

Temperature °C	Input resistance	Gain
20	9·9 MΩ	0·97
30	9·9 MΩ	0·97
40	9·9 MΩ	0·97
50	11·0 MΩ	0·97
60	11·0 MΩ	0·97
66	12·4 MΩ	0·97

The fact that R_{in} vs. temperature does not follow a smooth curve is due to the difficulty of accurate measurement.

Frequency response. From a low resistance source the bandwidth is greater than 10 c/s to 100 kc/s (the range of the available oscillator).

Output resistance (measured with a load of 1 kΩ).

With input short circuited : 25 Ω

With 100 kΩ in input : 31 Ω

Maximum output voltage. Into 1 kΩ load : 1 V r.m.s.

Variation of performance with β. In the example above the overall current gain of the circuit was

$$\beta_1\beta_2\beta_3\beta_4 = 40\times40\times45\times38 = 2{\cdot}74\times10^6$$

The transistors were removed and replaced by another series. R_{b1} was altered to 390 kΩ. R_f remained unchanged at 180 kΩ. The new overall value of β was

$$\beta_1\beta_2\beta_3\beta_4 = 25\times25\times25\times17 \simeq 0{\cdot}266\times10^6$$

The gain was 0·97 and the input resistance was 3·6 MΩ.

5. Conclusions

Various methods of increasing circuit resistance have been surveyed and suggested. The principal points to note are stated briefly below.

5.1. *Input Resistance*

Although the maximum input resistance of the circuits discussed in Section 4.6 is r_c, R_{in} usually falls far short of r_c.

The value of R_{in} is normally calculated without consideration of the bias resistor R_{b1}. The complete input resistance is then calculated using the effective value of R_{b1}. The limiting value of input resistance is often due to the resistor R_{b1} as much as to any other cause.

5.2. *Circuit Stability*

In devices which have a limiting input resistance greater than r_c, the term $1/(1-K)$ forms part of the circuit equations. This involves differences between two large quantities and attention must be paid to this. The difference must be large enough to prevent variation in circuit values leading to unstable conditions.

In the practical circuits given, the capacitors across collector loads and feedback resistors were inserted to prevent loop oscillations.

5.3. *Circuit Components*

When very large resistances are involved, the fact that electrolytic capacitors have appreciable d.c. conductance must not be overlooked.

5.4. *Effect of Frequency*

As the frequency of operation is increased, the forward gain and the circuit impedance will decrease. The effect of this on the input resistance must be borne in mind. For example, the input resistance of the amplifier of Fig. 23 fell to 0·71 of its maximum value at 9 kc/s.

6. Acknowledgments

The author would like to acknowledge the considerable amount of measurement work carried out by Messrs. P. Hider, A. Ramsay and J. Doggett, which was an important part of the investigation. The author wishes to thank the Director and Council of The British Scientific Instrument Research Association for permission to publish this paper and Messrs. Texas Instruments Limited for permission to include the circuit given in Section 4.7.

7. Bibliography

1. R. Stampel and R. A. Hanel, "Transistor amplifier with extremely high input impedance", *Proc. Natl Electronics Conf.*, **11**, 1955.
2. R. D. Middlebrook and C. A. Mead, "Transistor a.c. and d.c. amplifiers with high input impedance", *Semiconductor Products*, **2**, March 1959.
3. I. Levine, "High impedance transistor circuits", *Electronics*, **33**, p. 50, 2nd September 1960.
4. A. G. Boyle, "Transistor matching impedance", *Electronic Technology*, **37**, p. 28, January 1960.
5. D. A. Smith and T. M. A. Lewis, "A buffer stage for piezo-electric strain gauges", *Electronic Engineering*, **34**, p. 99, February 1962.

Manuscript first received by the Institution on 13th August 1962 and in final form on 3rd October 1962. (Paper No. 826.)

U.D.C. 621.317.331 : 66-932

An Electrodeless Conductivity Meter for Process Control

By

E. HARRISON†

AND

P. F. ROACH *(Graduate)*†

Presented at the Convention on "Electronics and Productivity" in Southampton on 17th April 1963.

Summary: An inductive form of electrodeless conductivity meter suitable for dirty and radio-active solutions is described. It gives a continuous indication, corrected for temperature variations, in the ranges 0·01 to 1 mho-cm with a reproducibility of 0·15% of range.

List of Symbols

A	Effective area of liquid loop
A_1	Cross-sectional area of T1 core
A_2	Cross-sectional area of T2 core
B_{m1}	Maximum flux density in T1
B_{m2}	Maximum flux density in T2
C	Cell constant
f	Frequency
G_B	Balancing loop conductance
G_L	Liquid loop conductance
ΔG_L	Incremental change in conductance from the balanced condition
I	Input current to winding N1 (Fig. 13)
I_L	Liquid loop current
I_2	Current flowing in winding N2
I_3	Balancing loop current
K	Specific conductivity
L	Effective length of liquid loop
l	Effective length of the cores of T1 and T2
M	Current gain of the temperature compensating amplifier
N1	Exciting winding
n_1	Number of turns on winding N1
N2	Balancing loop winding on T1
n_2	Number of turns on the winding N2
N3	Balancing loop winding on T2
n_3	Number of turns on winding N3
n_4 and n_5	Single turn liquid loops on T1 and T2 respectively
N6	Signal output winding on T2
n_6	Number of turns on winding N6
R_1	Resistance of the balancing loop potentiometer
R_1'	Resistance of the read-out potentiometer
R_2	Total resistance of the balancing loop
R_2'	Read-out circuit loading resistance
T1	Primary transformer
T2	Secondary transformer
V_b	Voltage across winding N6 for incremental change in conductance G_L
V_L	Voltage induced into the single turn liquid loop (Fig. 13)
V_S	Read-out supply voltage
V_{out}	Read-out voltage
V_1	Energizing or excitation voltage
x	Fractional position, from the common end, of the balancing loop and read-out potentiometer sliders
α	Temperature coefficient of liquid conductance
β	Temperature coefficient of resistance for the resistance thermometer
θ	Temperature departure from a nominal set-point
μ_2	Permeability of T2
Φ_{T1}	Flux in T1 (Fig. 13)
Φ_{T2}	Flux in T2 (Fig. 13)

1. Introduction

The efficient control of a chemical separation plant for the recovery of plutonium and uranium requires a knowledge of the normality of acid solutions in which irradiated reactor fuel elements have been dissolved. For the purpose of control, continuous in-line monitoring has obvious advantages.

A system devised to meet the above requirements necessitated the separate measurement of density and electrical conductivity. However, the measurement of electrical conductivity is complicated by the following four difficulties encountered in working with these solutions:

† Reactor Engineering Laboratory, U.K. Atomic Energy Authority, Risley, Lancashire.

(a) radio-activity,

(b) corrosion,

(c) suspension of solids,

(d) tendency to precipitate impurities (largely silica-gel).

These plant conditions, together with the high conductivity of the solutions, rule out the use of all conventional electrode type cells.

An instrument, based on the electrodeless induction type conductivity meter, was developed,[1, 2] which was temperature compensated and self balancing to give a continuous reading. It therefore lends itself to process control applications.

2. Principle of Operation

2.1. *Single Loop Method*

The basic principle of operation is illustrated in Fig. 1. The liquid path, defined by the geometry of the system, forms a coupling winding between the energizing and measuring transformers, T1 and T2. The e.m.f. at the output winding, for a constant input voltage and frequency, is proportional to the conductivity G_L of the liquid.

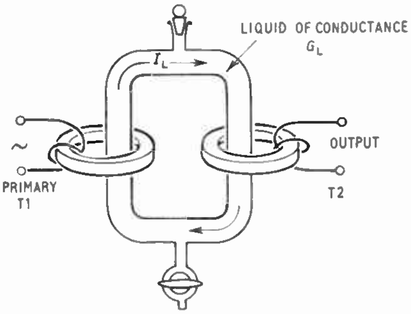


Fig. 1. Simple single loop coupling.

2.2. *Balanced System*

An improved system, basically a transformer bridge, is shown in Fig. 2. The flux induced into the secondary core by the liquid loop is in anti-phase to the flux induced by the external balancing loop. Assuming the external loop to be a single turn only, a null is obtained when $G_L = G_B$. The balance point is now independent of circuit parameters common to both loops.

If a linear variable resistance is used for G_B, the relationship between G_L and the physical position of the wiper of such a resistance is a reciprocal function. If a conductance is used, it must comprise several decade boxes in parallel, depending on the accuracy required. Neither the simple series resistance (requiring variable gain) nor variable conductance is suitable for a servo balance system.

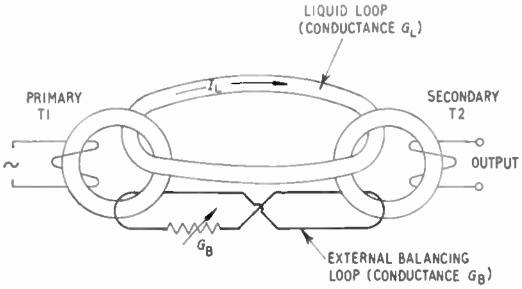

LIQUID LOOP (CONDUCTANCE G_L)
I_L
PRIMARY T1
SECONDARY T2
~
OUTPUT
G_B
EXTERNAL BALANCING LOOP (CONDUCTANCE G_B)

Fig. 2. Balanced system.

The basic measuring system actually adopted is shown in Fig. 3. Comparing this system with that shown in Fig. 2, the balancing loop current I_3 is now determined by R_2 and x. x is the physical position of the wiper of R1, measured from the common end, and varies from 0 to 1. R2 defines the maximum conductance that can be measured for any range and the accuracy with which the balancing loop current can now be set is limited by the resolution of R1. By using a multi-turn potentiometer this can be made better than 0·1% of the maximum value of current.

3. Design of the Basic Measuring System

When an alternating voltage V_1 is applied to the exciting winding N1, the voltage induced into each turn of all the windings on T1 will be (neglecting leakage flux) V_1/n_1 volts per turn. The current I_L flowing in the single-turn liquid loop is therefore

$$I_L = \frac{V_1}{n_1} G_L$$

and since there is just one liquid loop turn coupling with T2, the liquid loop ampere-turns ($I_L n_5$), will be

$$(I_L n_5) = \frac{V_1}{n_1} G_L$$

The ampere-turns on transformer T2 due to the external balancing loop, ($I_3 n_3$), are derived in

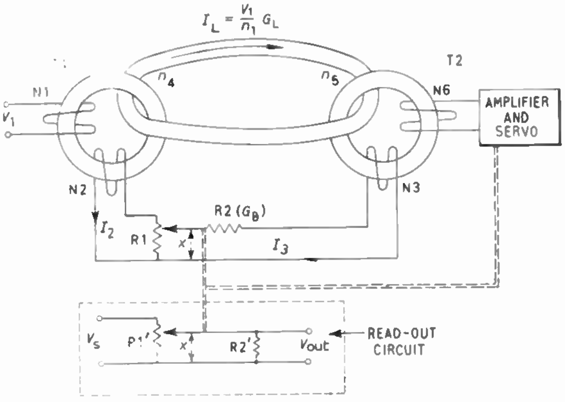


Fig. 3. Basic measuring system.

Appendix 1 and are shown to be

$$(I_3 n_3) = \frac{V_1 n_2 n_3 x}{n_1 [R_2 + R_1(x-x^2)]} \qquad \ldots\ldots(1)$$

where n_3 is the number of turns of the balancing loop winding coupling with the secondary core.

Evidently, at balance (I_L n_5) and (I_3 n_3) must be equal, hence

$$\frac{V_1}{n_1} G_L = \frac{V_1 n_2 n_3 x}{n_1 [R_2 + R_1(x-x^2)]} \qquad \ldots\ldots(2)$$

$$G_L = \frac{n_2 n_3 x}{R_2 + R_1(x-x^2)} \qquad \ldots\ldots(3)$$

If K is the specific conductivity, $K = G_L L/A$ where A is the cross-sectional area and L the effective length of the liquid loop. In practice, the ratio of these two-dimensional constants can only be determined empirically and is denoted by C the cell constant; we can therefore re-write eqn. (3) as

$$K = \frac{C n_2 n_3 x}{R_2 + R_1(x-x^2)} \qquad \ldots\ldots(4)$$

Equation (4) indicates that the relationship between specific conductivity and x (the physical position of the slider on R1) is not linear; the deviation from linearity is not great, however, if R_2 is made appreciably larger than R_1.

3.1. *Read-out Circuit*

A convenient method of obtaining a linear output voltage/conductivity relationship is as follows. A potentiometer R'1 loaded with a resistor R'2 is fed from a voltage source V_S (Fig. 3). The output signal is given by

$$V_{out} = \frac{V_S R'_2 x}{R'_2 + R'_1(x-x^2)} \qquad \ldots\ldots(5)$$

R'1 is ganged with R1 (twin-ganged helical potentiometer), hence x is common to eqns. (4) and (5). The combination of these two gives

$$\frac{K}{V_{out}} = \frac{C n_2 n_3 x R'_2 \left[1 + \frac{R'_1}{R'_2}(x-x^2)\right]}{V_S R'_2 x R_2 \left[1 + \frac{R_1}{R_2}(x-x^2)\right]}$$

If the ratio $R_1/R_2 = R'_1/R'_2$, the relationship between the specific conductivity and the output voltage from the additional ganged potentiometer, is given by

$$K = \frac{C n_2 n_3}{V_S R_2} V_{out} \qquad \ldots\ldots(6)$$

and is a linear relationship whereby C, n_2, n_3, V_S and R_2 are constant. In practice

$$R'_1 = R_1 \quad \text{and} \quad R'_2 = R_2$$

The voltage V_S can be any convenient a.c. or d.c. value and the output signal (V_{out}) is fed to a high impedance indicator. The indicator can be remote from the electronic unit, at any distance likely to be encountered in a chemical plant.

For most practical cases, the resistance of the interconnecting leads between conductivity cell and electronic unit can be neglected. The lead resistance however modifies eqn. (4) but by adding resistance to the read-out circuit in the appropriate place, eqn. (5) can be modified in a similar way; this preserves output linearity, in other words, eqn. (6) remains unaltered.

If the read-out circuit is not modified, a fixed error plus a degree of non-linearity is introduced, depending upon the lead resistance and relative values of R_1 and R_2.

4. General Description

The complete conductivity measuring system is shown schematically in Fig. 4. The exciting voltage V_1 is supplied by an oscillator and amplifier; a frequency of 10 kc/s is used. The instrument is designed to have three ranges 0–1 mho-cm, 0–100 mmho-cm and 0–10 mmho-cm. The range of the instrument is changed by

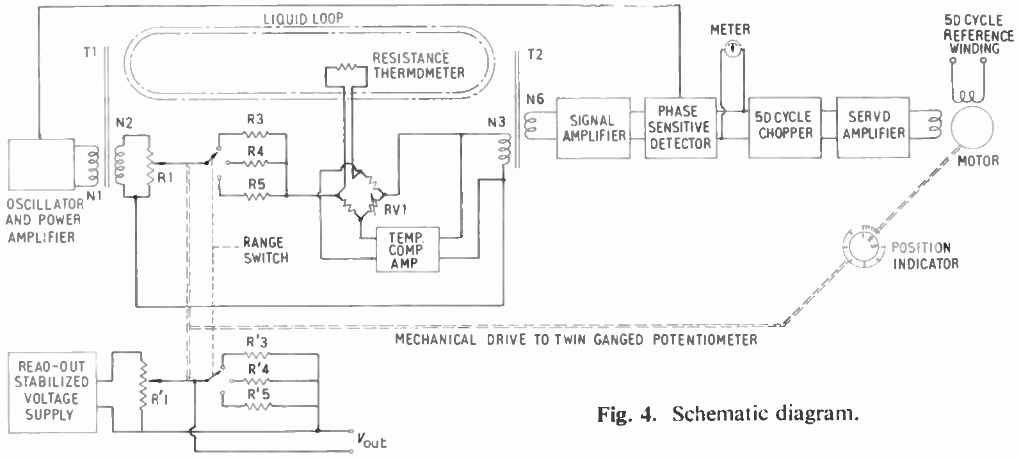


Fig. 4. Schematic diagram.

selecting values of R_3, R_4 or R_5; R'_3, R'_4 and R'_5 must also be changed to preserve output linearity (see section 3.1).

When the secondary-core liquid loop ampere-turns do not cancel the balancing loop ampere-turns, the out-of-balance signal is amplified by the signal amplifier and fed to the phase-sensitive detector. The latter ascertains the phase of the out-of-balance signal and develops the correct d.c. polarity for moving the servo towards balance. Also, it discriminates between the desired signal, which arrives via the resistive liquid and balancing loops, and quadrature signals due to direct magnetic coupling between the toroids, leakage inductance and stray capacitive coupling. The d.c. output from the phase-sensitive detector is chopped at 50 c/s and fed to the servo amplifier. The helical potentiometers are driven to balance by the servo motor via a reduction gear of 100 : 1; an extension of the helical potentiometer shaft drives a turns counting dial or position indicator mounted on the front panel of the electronic unit.

The inclusion of a switch to cut out the servo operation and a centre reading d.c. voltmeter (connected across the output of the phase-sensitive detector) makes manual balancing possible for maintenance and checking purposes.

5. Measuring Head Design

The effects of primary and secondary core parameters on sensitivity can be assessed from the following equation, which is derived in Appendix 2:

$$V_b \propto \frac{\mu_2 n_6 f^2 A_2 A_1 B_{m1} \Delta K A}{L . l} \qquad \ldots\ldots(7)$$

where V_b is the voltage developed across a signal output winding of n_6 turns when the bridge is put off balance by an amount

$$G_L = \Delta K A / L$$

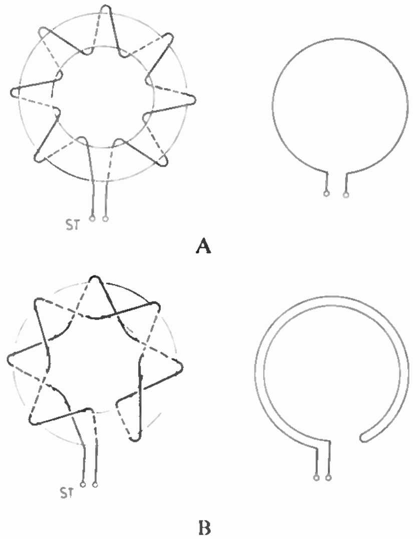


Fig. 5. Core windings. Astatic winding B eliminates single turn coupling possible with two cores wound as A.

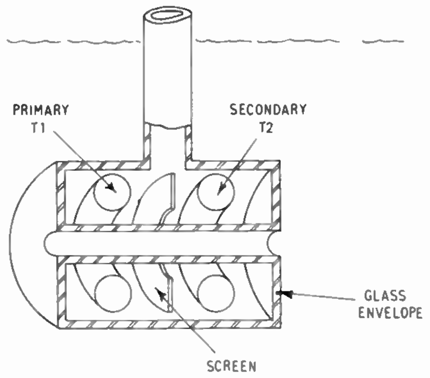


Fig. 6. Conductivity cell—simplified diagram.

The parameters effecting sensitivity are thus:

Primary core: B_{m1}, A_1, f

Secondary core: n_6, A_2, f, μ_2, l

The frequency f can only be increased up to the point where "skin effects" in the solution become significant. With fixed overall dimensions for the cell any advantage gained by increasing the primary and secondary core cross-sectional areas A_1 and A_2 or their lengths l is offset to some extent by the resulting physical changes in the area A and the length L of the liquid loop.

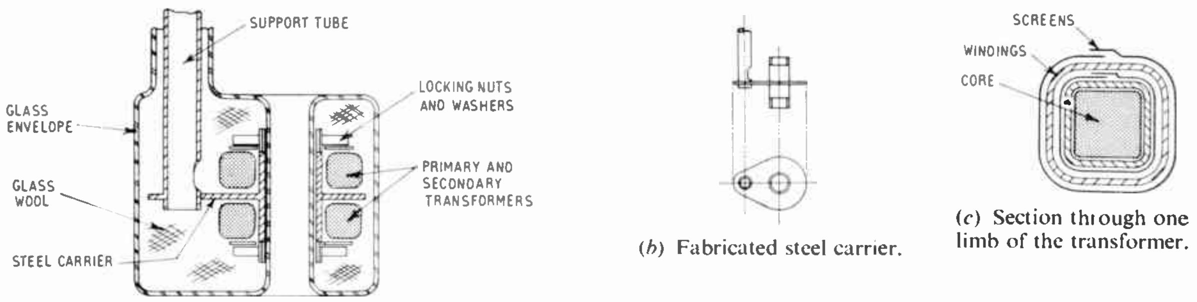
B_{m1}, the primary core flux and μ_2 the permeability of the secondary core, were in fact governed by the core material chosen for this particular application, i.e. a ferrite with good radiation-resisting properties. In the absence of radiation, advantage could be taken of the high-flux-density materials for the primary core and high μ materials for the secondary core.

It was soon discovered that the limitations of sensitivity were to be governed by the amplitude of the unwanted residual signal. The number of turns n_6 of the signal output winding was therefore made just as large as was convenient for winding.

To keep the stray pick-up between T1 and T2 to a minimum the following precautions are taken. The windings are wound astatically and the turns are very accurately spaced (see Fig. 5). Metallic screens are fitted round each winding and the leads to and from the transformer cores are screened. The ferrite cores are selected for minimum amount of ovality, and after assembly the wound cores are orientated with respect to each other, to give minimum stray coupling.

5.1. *Mechanical Form*

The mechanical form of the conductivity cell is shown, in a simplified form, in Fig. 6. By making the orifice through the centre of the toroids long compared



(*b*) Fabricated steel carrier.

(*c*) Section through one limb of the transformer.

(*a*) Section through cell. Fig. 7. Conductivity cell.

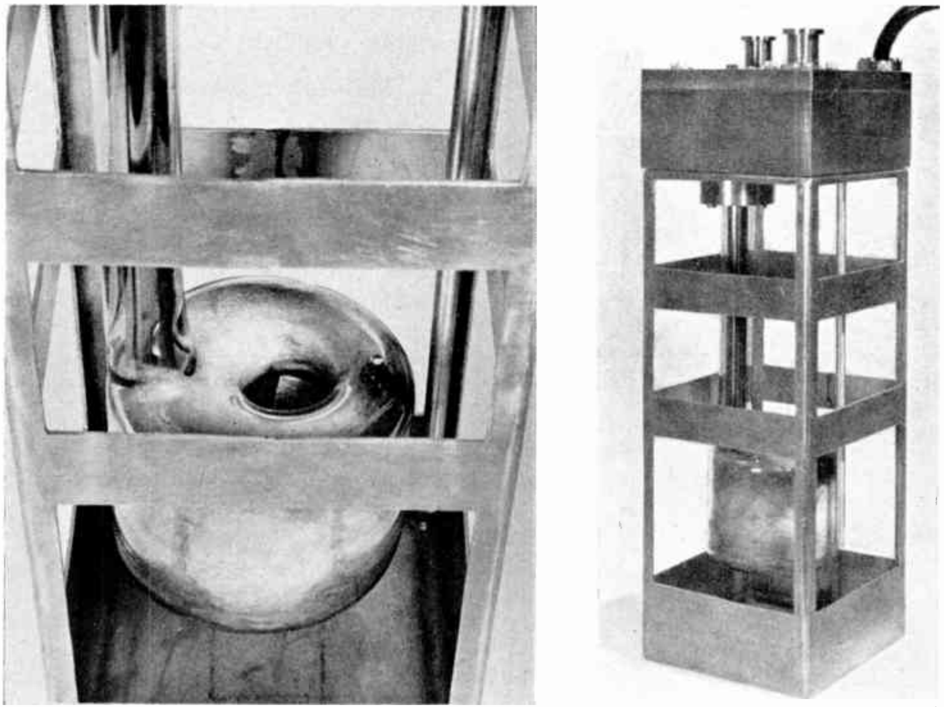
with its diameter, the bulk of the liquid loop resistance is concentrated in the central tubular portion. The effects of shunting current paths through adjacent metal surfaces and changes in the shape of the solution container are thus minimized.

There are no clearly defined optimum cell dimensions. The diameter of the orifice was made large in comparison to the size of solids likely to be suspended in the solution concerned and the volume of material likely to precipitate out in a given time. The problem of defining the current paths around the outside of the cell is difficult. An orifice L/A ratio high enough to make negligible the effects of the steel trough walls on the conductivity reading was established empirically. Ratios of 3 : 1 and above were found satisfactory and since (from eqn. (7)) the sensitivity of the system is adversely affected by increasing this figure, a ratio of 3 : 1 was adopted for the design.

5.2. *Mechanical Design*

A section through the conductivity cell is shown in Fig. 7. The two transformers mounted co-axially, are locked on to, and supported by, a stainless-steel carrier. The support tube, forming part of the carrier, is threaded at the end remote from the transformers and can be locked to any convenient supporting frame. The glass envelope is a loose fit over the carrier and coils. Glass wool, packed round the coil assembly prior to sealing the glass envelope, restricts the movement of the latter and reduces the risk of damage.

The only suitable encapsulating material for this application is silica glass. To relieve strains after fabrication, the completed envelope must be annealed at 560° C; hence the core, insulation and wire must be able to withstand this temperature. Nickel-plated copper wire, insulated with glass braiding, is used to wind the toroids, the metallic screens and cores being



(*a*) Glass encapsulated cell with resistance pocket. (*b*) Cell mounted on a carrier.

Fig. 8. Conductivity cell assembly for operating in radio-active solutions.

insulated with ruby mica and glass tape. Photographs of the assembled conductivity cell are shown in Fig. 8.

6. Temperature Compensation

Aqueous solutions have a high temperature coefficient of conductance, i.e. + 1·5 to 3% per deg C. Consider eqn. (4) when the temperature of the solution is altered; we now have

$$K(1+\alpha\theta)=\frac{Cn_2 n_3 x}{R_2+R_1(x-x^2)} \qquad \ldots\ldots(8)$$

where α is the temperature coefficient of conductance for the solution and θ is the temperature change. It is required that the value of conductivity indicated refers always to one fixed temperature T (set point temperature). Thus the right hand side of eqn. (8) requires a temperature sensitive term equal to $(1+\alpha\theta)$. This is achieved by using a temperature-sensitive resistance bridge network in series with and forming part of the resistance R2 in Fig. 3. The bridge is shown in detail in the schematic diagram Fig. 4, where R_3 plus the bridge resistance is equivalent to R_2.

One arm of the resistance bridge is a temperature-sensitive element (resistance thermometer) which is immersed in the solution. The bridge is balanced (by adjustment of RV1, Fig. 4) at the set point temperature T, and as the solution temperature varies the resulting out-of-balance voltage is amplified and fed into the winding N3.

The method of temperature compensation is best understood by referring back to Fig. 3. If the current in the liquid loop I_L changes by an amount ΔI_L due to temperature only, the secondary core has a change in ampere-turns of ΔI_L (since there is one liquid loop turn). For the read-out to remain unchanged, the liquid loop and balancing loop ampere-turns must remain equal. Thus the balancing loop current I_3 through N3, must be augmented by an externally supplied current of magnitude $\Delta I_L/n_3$. The output of the temperature-compensating amplifier is therefore connected across winding N3 as shown in Fig. 4.

The temperature compensating amplifier gain M is made variable to accommodate a range of temperature coefficients. In Appendix 3 an expression is derived relating M and α to the temperature coefficient of the resistance thermometer, β.

The temperature-compensating system can be made inoperative by simply switching off the power supply to the amplifier. This facility, included in the instrument developed, is essential for initial setting-up procedure in the field.

7. Circuit Details

7.1. *Oscillator and Power Amplifier*

The exciting voltage is supplied by a push-pull output stage driven from a 10 kc/s oscillator (Fig. 9). The collector load of the push-pull stage is in series with the primary of a small air-cored transformer T3. The transformer feeds the phase-sensitive detector switches, introducing a 90 deg phase shift, which is necessary owing to a corresponding change through the transformer bridge (see Appendix 4).

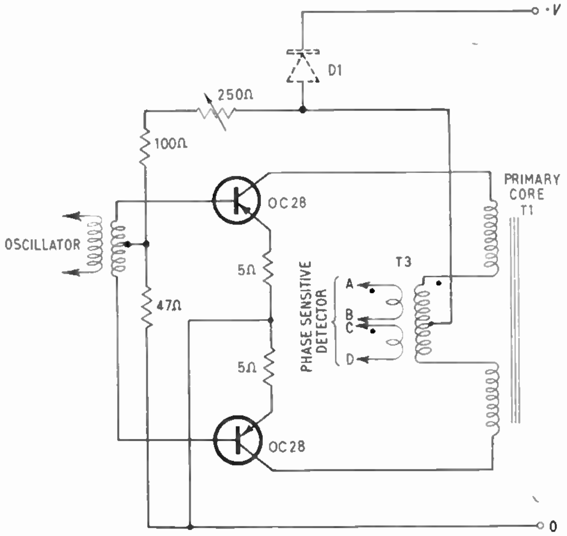


Fig. 9. Power amplifier and phase sensitive detector drive.

7.2. *Signal Amplifier*

The signal amplifier is conventional in design but its gain must be adjustable, the low conductivity range demanding greater sensitivity. The amplifier must not be saturated by the residual quadrature signal and the phase change through the amplifier must be small.

7.3. *Temperature Compensating Amplifier*

The temperature compensating amplifier requires care, since its output is in parallel with the balance signal in the external balancing loop; spurious outputs may alter the balance point. The d.c. supplies to the amplifier must be well filtered and the amplifier itself must be carefully screened. The output current from this amplifier must be either in phase or in anti-phase with the balancing loop current I_3 (according to whether the liquid loop temperature is above or below the set point temperature T). As the input to the amplifier is in phase with I_3, the phase shift through the amplifier must be kept to a few degrees and must not change appreciably with signal level, temperature, etc. To minimize spurious input signals to the amplifier, the resistance thermometer bridge and amplifier form a separate unit close to or mounted directly on top of the conductivity cell.

7.4. *Phase-sensitive Detector*

The phase-sensitive detector is of the conventional form as shown in Fig. 10 and comprises two pairs of symmetrical transistors. The switching voltage

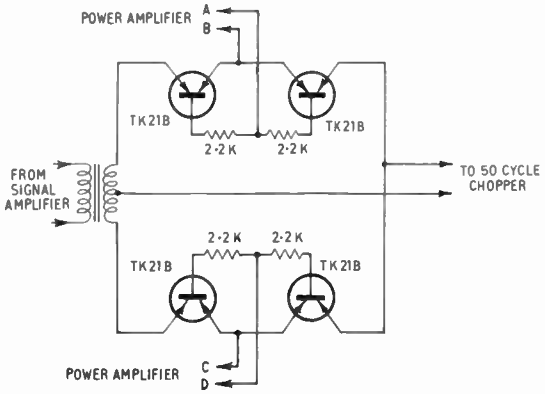


Fig. 10. Phase sensitive detector.

applied to one pair of transistors is in phase with the amplifier output signal voltage and in anti-phase to the voltage applied to the second pair of transistors.

8. Elimination of the 50 c/s Chopper

For high-conductivity measurements in the range 0·1 to 1 mho, the electronic circuits can be simplified. The d.c. supply to the push-pull output stage is replaced by a suitable 50 c/s voltage. A series diode D1 protects the transistors from the 'reversed' cycle of the supply (Fig. 9). The excitation voltage waveform is now similar to that shown in Fig. 11(*a*). The output from the phase-sensitive detector (Fig. 11(*b*)) is filtered and fed directly to the servo amplifier, eliminating the chopper.

With an excitation signal of this form the power fed into the liquid loop is approximately halved, reducing sensitivity. Also, because of harmonics generated during the build-up and decay periods of this type of waveform the residual output at balance is increased (Fig. 11(*c*)). These factors together make the 'modulated' system unattractive for low-conductivity measurements.

9. Setting-up Procedure

The magnitude and form of the read-out signal (it could be a.c. or d.c.) will be determined by external requirements. However, assuming an output of Y volts/mho-cm and a range of 0–1 mho-cm, $V_S = Y$ volts (Fig. 3). With the cell immersed in a solution of known specific conductivity K, the resistors R3 and R'3 are adjusted simultaneously, keeping their values the same, to give an output of KY volts (Fig. 4). Only R3 needs to be set accurately; the value of R'_3 can be within ±10% of that of R_3.

10. Cell Constant

The relationship between the liquid loop conductance G_L and the required specific conductance K is given in section 3 as $K = G_L\, L/A$. In the simple cell arrangement of Fig. 1, the dimensions L and A could be controlled quite accurately, making an absolute measurement possible. The difficulties of assessing the ratio L/A in an arrangement similar to Fig. 6 are obvious and a standard solution must be used for the initial calibration.

11. Performance

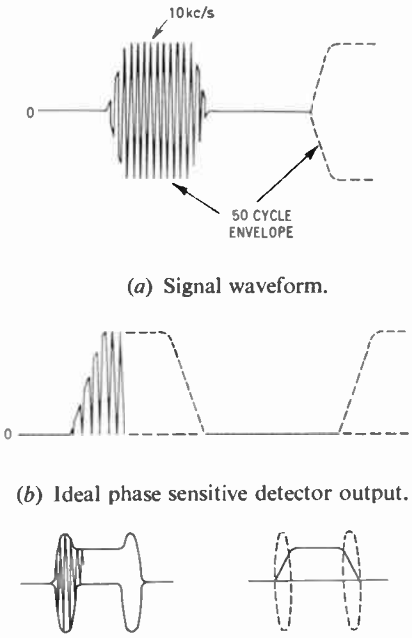
11.1. *Linearity*

The instrument is checked for linearity using fixed resistors connected to a wire loop threaded through the cell orifice. For the high-conductivity range, several wire loops in parallel are necessary to reduce the leakage inductance; these are connected to carbon resistors rather than a decade box, to reduce the inductance/resistance ratio. Without these precautions, a large residual signal can be expected. The percentage deviation from linearity did not exceed 0·1% of full-scale reading when balanced manually (eliminating servo dead-zone error). The error, including the servo dead-zone, did not exceed 0·15% of full-scale reading.

11.2. *Temperature Compensation*

Temperature compensation can be checked initially by using fixed resistors to simulate the liquid loop, and in addition by replacing the resistance thermometer by a decade resistance box. Any departure from linearity during tests carried out in this way were within ±0·15% of range.

In solutions, the temperature compensation was able to hold the instrument reading to within the



(*a*) Signal waveform.

(*b*) Ideal phase sensitive detector output.

(*c*) Actual phase sensitive detector output showing the effect of harmonics generated during build-up and decay periods of signal waveform.

Fig. 11. Waveforms—'modulated' system.

dead-zone of the servo for temperature changes of ± 10° C.

11.3. *Reference Standard*

An electrodeless conductance measuring system,† loaned by Wayne Kerr Laboratories, was used as a standard against which to check the stability of the self balancing system. The temperature and composition of the test solutions were varied over a period of four weeks. The conductivity of samples, taken daily, was measured at 25° C by the Wayne Kerr instrument.

The deviation from a mean calibration constant was less than $\pm 0{\cdot}3\%$.

12. Conclusions

Tests show that the instrument has a reproducibility of better than 0·15% of range. The accuracy depends upon the standard solution used and the selection of a balancing loop resistor during initial calibration; an overall measurement of specific conductivity, to within 1%, is easily achieved.

The instrument described has been designed for, and successfully worked under particularly stringent conditions. Considerable simplification is possible for less arduous applications, i.e. the use of metal toroidal cores and an encapsulating material more workable than glass. A small robust cell might be inserted directly into a process pipe line (Fig. 12).

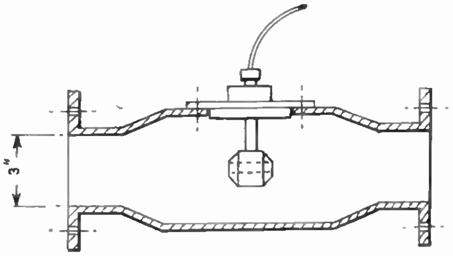


Fig. 12. Proposed conductivity cell of small dimensions, fitted into a process pipe line.

Regarding the electronic circuit, a higher R_2/R_1 ratio, giving better conductivity/slider position linearity, might enable the position indicator reading to be used as a measure of conductivity, removing the necessity for a read-out circuit. Manual control would eliminate the expensive servo and the use of a thermister in place of a resistance thermometer, simplifies temperature compensation.

13. Acknowledgments

The instrument described was developed successfully as the result of cumulative ideas and effort by the members of the Instrumentation Section, Reactor Engineering Laboratory, Risley. The paper is published by permission of Mr. R. V. Moore, Managing Director of the U.K.A.E.A. Reactor Group.

14. References

1. S. R. Gupta and G. J. Hills, "Precision electrodeless conductance cell for use at audio frequencies", *J. Sci. Instrum.*, **33**, p. 313, 1956.
2. M. Salamon and P. Svitok, "Low frequency conductometer for measuring the electrical conductivity of solution", *Chem. průmysl*, **6**, No. 31, pp. 10–14, 1956 (U.K.A.E.A. translation IGRL-T/CA-108, 1959).

Appendix 1

Derivation of Equation (1)

With reference to Fig. 4, the voltage V_2 induced into the primary core winding of n_2 turns is

$$V_2 = \frac{V_1}{n_1} n_2$$

Also
$$V_2 = I_2 R_1(1-x) + (I_2 - I_3)xR_1$$
$$= I_2 R_1 - I_3 x R_1 \qquad \ldots\ldots(9)$$

Also
$$I_3 R_2 = (I_2 - I_3)xR_1$$

and
$$I_2 = \frac{I_3}{xR_1}(R_2 + xR_1) \qquad \ldots\ldots(10)$$

Substituting eqn. (10) in (9):

$$V_2 = \frac{I_3}{x}(R_2 + xR_1) - I_3 x R_1$$

$$V_2 = I_3 \frac{R_2 + R_1(x - x^2)}{x}$$

Therefore
$$I_3 = \frac{V_2 x}{R_2 + R_1(x - x^2)}$$

The balancing loop ampere-turns ($I_3 n_3$) in T2 are

$$(I_3 n_3) = \frac{n_3 V_2 x}{R_2 + R_1(x - x^2)}$$

However,

$$V_2 = \frac{V_1 n_2}{n_1}$$

Therefore
$$(I_3 n_3) = \frac{V_1 n_2 n_3 x}{n_1 [R_2 + R_1(x - x^2)]} \qquad \ldots\ldots(1)$$

Appendix 2

Derivation of Equation (7)

Equation (2) for the balanced condition gives

$$\frac{V_1}{n_1} G_L = \frac{V_1 n_2 n_3 x}{n_1 [R_2 + R_1(x - x^2)]}$$

The out-of-balance ampere-turns in the secondary core of T2, when the bridge is out of balance by ΔG_L is

[†] The instrument was that used by Gupta and Hills[1]; its reproducibility is claimed to be better than 0·02%. British Patents 761,903 and 831,692 on the basic liquid loop cell are held jointly by Wayne Kerr Laboratories and R. Calvert.

$$n_5(I_L+\Delta I_L)-I_3 n_3 = \frac{V_1}{n_1}(G_L+\Delta G_L) - \frac{V_1 n_1 n_3 x}{n_1[R_2+R_1(x-x^2)]} = \frac{V_1}{n_1}\Delta G_L$$

Now

$$V_1/n_1 = 4{\cdot}44 B_{m1} f A_1 . 10^{-8} \text{ volts per turn} \quad ......(11)$$

The voltage output from the signal winding will be

$$V_b = 4{\cdot}44 B_{m2} f A_2 n_6 . 10^{-8} \quad ......(12)$$

Now

$$B_{m2} = \left(\frac{4\pi\mu_2}{10\,l}\right)\left(\text{out of balance ampere turns}\right) = \frac{4\pi\mu_2 V_1 \Delta G_L}{10\, l n_1} = \frac{1{\cdot}26\mu_2 \Delta G_L V_1}{l n_1} \quad ......(13)$$

Combining eqns. (12) and (13)

$$V_b = \frac{5{\cdot}59\mu_2 \Delta G_L n_6 f A_2 . 10^{-8}}{l} \frac{V_1}{n_1} \quad ......(14)$$

Substituting eqn. (11) in (14) and KA/L for G_L:

$$V_b = \frac{24{\cdot}8\mu_2 \Delta K n_6 f^2 A_1 A_2 A B_{m1}}{l L . 10^{16}} \quad ......(7)$$

Appendix 3

Derivation of the Relationship between M, α and β

The current feeding the thermometer bridge is I_3, the balancing loop current (Fig. 3). The output from the bridge (assuming the input impedance of the amplifier is matched to the impedance of the bridge and all arms of the bridge are equal) can be shown to be $I_4 = I_3\beta\theta/8$ when $\beta\theta \ll 8$. The amplifier output current fed into the winding N3 on the secondary core T2 will be

$$MI_4 = \frac{MI_3\beta\theta}{8}$$

and hence the temperature compensating ampere-turns will be

$$(n_3 MI_4) = \frac{n_3 MI_3 \beta\theta}{8} \quad ......(15)$$

The balancing loop ampere-turns is given in Appendix 1 as

$$(I_3 n_3) = \frac{V_1 n_2 n_3 x}{n_1[R_2+R_1(x-x^2)]} \quad ......(16)$$

The total balancing loop ampere-turns with temperature compensation becomes

$$I_3 n_3 + \frac{I_3 n_3 M\beta\theta}{8}$$

substituting for $I_3 n_3$ from eqn. (16) we have:

Total balancing loop ampere-turns

$$= \frac{V_1 n_2 n_3 x}{n_1[R_2+R_1(x-x^2)]}\left[1+\frac{M\beta\theta}{8}\right]$$

Inserting the temperature coefficient of conductance in eqn. (2) the expression for balance, with temperature compensation, becomes

$$\frac{V_1}{n_1} G_L(1+\alpha\theta) = \frac{V_1 n_2 n_3 x}{n_1[R_2+R_1(x-x^2)]}\left[1+\frac{M\beta\theta}{8}\right]$$

For balance at varying θ, $\alpha = M\beta/8$.

Appendix 4

Primary to Secondary Core Phase Shifts

Since the collector load of the push-pull stage (Fig. 9) is mainly resistive, the applied voltage V_1 and input current I are approximately in phase (see Fig. 13). The flux Φ_{T1} in the core of T1 lags the primary current by approximately 90 deg. The voltage V_L and the current I_L, induced into the liquid loop are in quadrature with the flux of $T1$.

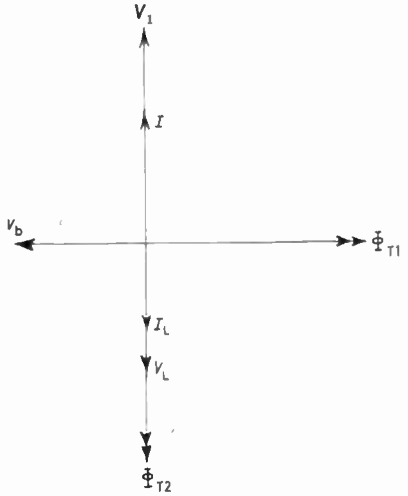


Fig. 13. Primary input—secondary output phase relationships. The magnetizing current has been neglected.

The flux Φ_{T2} in the secondary core of T2 is produced by and in phase with I_L, whilst the output voltage V_b, developed across N6 (Fig. 3) is in quadrature with the flux of T2. Thus it can be seen that the output voltage is in quadrature with the input current I. As the phase-sensitive detector switching voltages are derived from the current I and as these voltages must be in phase with the signal output voltage, a 90 deg correction is necessary; this is obtained in the transformer T3.

Manuscript received by the Institution on 5th February 1963. (Paper No. 827.)

# of current interest . . .

Open Day at the Royal Radar Establishment

The Royal Radar Establishment (R.R.E.) at Malvern, Worcestershire, is the largest centre for electronics research in Great Britain. It is responsible, in broad terms, for research in physics and electronics and for the development, in collaboration with industry, of radar and electronic equipment for the three Defence Services. During May a series of Open Days were held, the first since 1948, and an impressive display was given of techniques and equipment which have been developed over recent years.

The Establishment is embarking on an extensive programme of work on air traffic control systems which obviously is closely linked with the Ministry of Aviation's civil aircraft activities. The increasing quantity of air traffic and the improved performance of aircraft, particularly in the upper air space above 25,000 ft in height, make an integrated air traffic control system essential. In the immediate future one of the main sources of information for such a system will be radar data, extracted, stored and manipulated in digital computers so as to present the working data to the controllers in the most convenient form. Of equal importance is flight plan information.

The first exhibit showed the potentialities and problems of ground based radar as an information source. Manual extraction of radar data must eventually be superseded by automatic tracking and the next exhibit showed the reasons for requiring automatic tracking, the advantages attainable and some of the problems encountered. A method of automatic detection and the initiation and maintenance of tracks was shown on a plan radar and the advantage of height measurement and the use of secondary radar in improving tracking capability were presented.

A control team demonstrated the principal functions of part of an A.T.C. system and showed how the processing of radar data by computer will enable controllers to perform their tasks with greater efficiency while maintaining a high standard of safety. A further extensive exhibit showed the processing of flight plan information with particular reference to the present procedural methods used on airways. This exhibit demonstrated how a computer can be used to process aircraft flight plans and how a controller can feed into a computer up-to-date information concerning aircraft delays etc. so that the computer produces revised flight plans to suit changing traffic situations.

In order to assess and improve the accuracy of modern tracking radars alternative and still more accurate methods of measuring the angular position of distant targets have been developed at R.R.E.

An interferometer, which measures the r.f. phase difference between two microwave signals, and which can detect angular changes of 10^{-6} radian, was demonstrated; its sensitivity corresponds to 0·02 in movement of the microwave source 500 yards away. The method of phase measurement makes the system virtually independent of the receiver components and the relative strength of the two signals being compared.

A television instrumentation tracker was also demonstrated which is capable of tracking and recording the true angular position of suitable light sources to an accuracy approaching 10^{-5} radian (3 seconds of arc). It can be used with aircraft, rockets or ships carrying suitable lamps and, with a very intense light, ranges up to 250 miles are possible on clear nights. The equipment will also lock on to stars and can be calibrated from their known courses.

The Physics and Electronics Department at R.R.E. is particularly well-known for applied physics research in such fields as thin magnetic films, superconducting thin films, semiconductor devices and infra-red photocells. The latter techniques were demonstrated in a research project in which a fast responding spectrometer recorded far infra-red radiation from the hot plasma of an electromagnetically driven shock tube. The Electronics Group of the Department is primarily concerned with circuit research including digital techniques, microminiaturization and reliability projects. The Group's low-noise masers are important contributions to systems work on ground radar.

STANDARD FREQUENCY TRANSMISSIONS

(Communication from the National Physical Laboratory)

Deviations, in parts in 10^{10}, from nominal frequency for **June 1963**

1963 June	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	− 130·4	− 130	− 4
2	− 130·3	− 130	—
3	− 130·3	− 130	− 3
4	− 130·4	− 131	0
5	− 130·1	− 131	+ 1
6	− 131·0	− 131	+ 3
7	− 130·8	− 132	− 1
8	− 131·3	—	0
9	− 130·8	—	0
10	− 129·7	—	+ 2
11	− 130·2	− 130	+ 1
12	—	− 129	+ 4
13	− 129·5	− 130	+ 4
14	− 130·9	− 130	+ 5
15	− 130·2	− 130	+ 5
16	− 129·4	− 130	+ 7
17	− 129·8	− 130	+ 7
18	− 129·9	− 132	+ 9
19	− 129·5	− 130	+ 6
20	− 129·8	− 130	+ 6
21	− 130·2	− 131	+ 3
22	− 129·5	—	+ 5
23	− 129·2	—	+ 11
24	− 129·4	− 129	+ 12
25	− 130·2	− 130	+ 14
26	− 128·9	− 130	+ 14
27	− 129·1	− 130	+ 14
28	− 129·4	− 131	+ 15
29	− 129·3	− 129	+ 16
30	− 129·7	− 130	+ 17

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 magnetic field.

U.D.C. 621.391.812.63: 621.396.65: 621.39

The Transmission of Digital Data over 500 km H.F. Radio Links

By

R. W. BROWN, B.Sc.(Eng.)†

Presented at a meeting of the Institution held in London on 14th November, 1962.

Summary: Experiments have been conducted to demonstrate the feasibility of using h.f. radio for the transmission of low grade digital data over 500 km paths with bit error rates of the order of 1 in 10^4. The paper includes a discussion of the factors leading to the poor reliability of short distance h.f. radio links. Equipment developed for the experiments is described and the results are presented.

Error rates of the order of 10^{-4} will probably be satisfactory for the transmission of raw seismic data, but should lower error rates be required, further work will be necessary to determine the cheapest and most reliable method of achieving this.

1. Introduction

Networks of stations are currently in use in areas subject to seismic activity for recording these phenomena. It is advantageous for the data to be recorded at a central station so that modern data processing methods may be applied. A network may therefore consist of a central recording station with data processing facilities and a number of auxiliary stations transmitting data to the central station. Many areas where such networks may be installed are characterized by remoteness and difficulty of access. The problem of transmitting data from auxiliary stations at a typical distance of 500 km was investigated. The remoteness and difficulty of the terrain increases the cost of either laying telephone lines or installing any radio system requiring repeaters, while maintenance problems become particularly severe. Of the alternative methods h.f. radio is the least expensive and this paper describes an experiment which was conducted to study the feasibility of using h.f. radio for the required data transmission links. Its use becomes a practical proposition since the extremely low error rates usually associated with data transmission are not necessary in the case of the raw data.

2. Propagation

The poor reliability of short distance h.f. communication links is well known. Such links are particularly susceptible to multi-path propagation[1] which causes a lengthening of signal elements due to the unequal travel times of the various components. At a transmission rate of 50 bits per second, at which the signal elements have a duration of 20 milliseconds, the synchronous regenerator was able to cope with lengthening of up to 9 ms without causing errors. Assuming the first arrival to be by single-hop E layer transmission with a path length of 540 km, the longest path length acceptable for a 9-ms delay would be 3240 km, corresponding roughly to 6-hop transmission via the F2 layer.

For a distance of 500 km using F2 layer transmission, the m.u.f. factor, i.e. the factor by which the critical frequency must be multiplied to give the m.u.f. is 1·16.[2] Hence, unless the frequency used is within 1/1·16 or 86% of the m.u.f., even vertical incidence radiation will be reflected and signals can be received via any number of hops within the limits of absorption and other path losses. The predicted m.u.f., however, is the mean of the expected monthly variations in m.u.f., and assuming these variations have a standard deviation of 0·12 × m.u.f. and a normal distribution, it can be seen that to use a frequency higher than 86% of the m.u.f. will result in a probability of less than 0·9 of reliable communication. If a frequency even closer to the m.u.f. were used, say 90%, the resulting reliability would be only 0·8 and there would still be a probability of 0·44 that vertical incidence radiation would be reflected. It is apparent, then, that protection against multi-path propagation cannot be realized by a choice of frequency sufficiently close to the m.u.f.

If a frequency well below the m.u.f. is used, the multiple hop components of the received signal will be absorbed to a greater extent than the single hop component and the latter will therefore be more likely to predominate. During the periods of dawn and dusk, however, all the signal components are likely to be subjected to Rayleigh fading, and ionospheric absorption is relatively low. It can be shown that if two signal components have mean field strengths $\bar{S}_1$ and $\bar{S}_2$ and have a Rayleigh distribution, the probability that S_1 will predominate is $\bar{S}_1{}^2/(\bar{S}_1{}^2+\bar{S}_2{}^2)$. If the path losses are negligible and the two components under consideration are the single-hop E layer component S_1, and the 7-hop F_2 layer component S_2, then the

† U.K. Atomic Energy Authority, Blacknest, Reading, Berkshire; now with the Canadian Marconi Company, Montreal, Canada.

mean field strengths will be proportional to the reciprocal of the path lengths and $\bar{S}_1 \simeq 7\bar{S}_2$, so that the 7-hop component can be expected to predominate for $\bar{S}_2{}^2/(49\bar{S}_2{}^2+\bar{S}_2{}^2)$ or 2% of the time. While it is recognized that this treatment is only very approximate, it can be seen that multi-path propagation will in fact cause a number of errors during the dawn and dusk periods and directional aerials must be used if the number of errors is to be reduced. Further, the aerials design should be such that it effects a compromise between the two requirements of (a), a high ratio of power radiated in the required direction for single-hop transmission to the power radiated at higher angles of elevation, and (b), the maximum possible gain in the required direction.

To use too low a frequency is of course as bad as using one that is too high, since the increased absorption will reduce the signal-to-noise ratio at the receiver and therefore increase the error rate. Reference 3 provides a method of choosing the optimum frequency for a required signal/noise ratio and describes a method of determining the probability that this signal/noise ratio will be realized. It fails to take account of multi-path propagation, however, so that the optimum frequency found by this method will probably be rather high, especially during the dawn and dusk periods.

3. Test Arrangement

A suitable data transmission link might take the form of the twelve-channel system shown in Fig. 1. Each channel has a capacity of 100 bits per second. Twelve two-tone frequency-shift baseband modulators feed a 600-watt independent sideband transmitter, giving six channels per sideband. Rhombic aerials would be used at both transmitting and receiving ends.

The experiment was carried out on a single-channel link equivalent to one channel of the above system. It consisted of a determination of error rate as a function of transmitter power and frequency, transmission rate, time of day and location of receiving site. Although facilities were also provided for the analysis of error distribution, no such analysis has yet been made. The locations of the transmitting and receiving sites are shown in Fig. 2. For reasons of simplicity and economy it was decided to use dipole aerials at both transmitting and receiving sites and to increase the radiated power per channel to compensate for the decrease in gain. This did mean, however, that the protection afforded by directional aerials against multi-path propagation was lost.

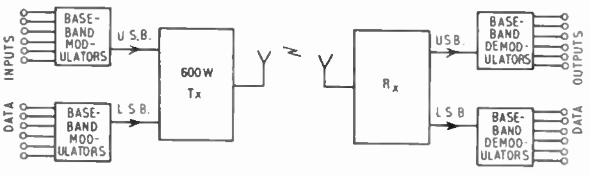


Fig. 1. Proposed 12-channel system.

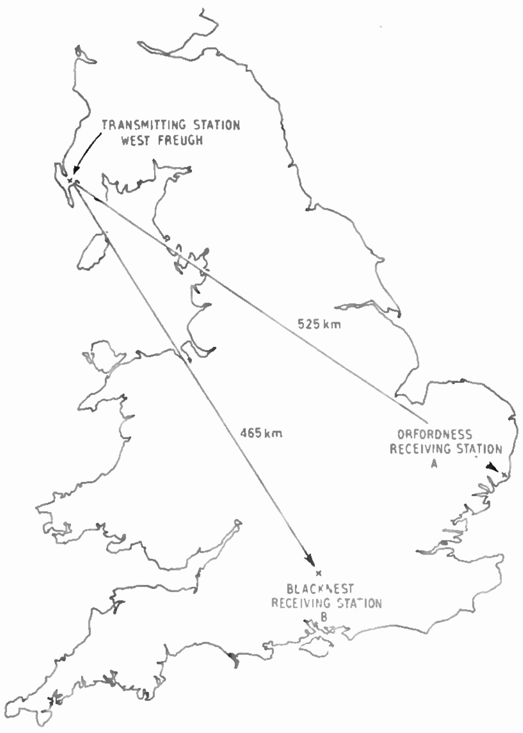


Fig. 2. Locations of transmitting and receiving sites.

Before the start of each test a reversals signal, i.e. a stream of alternate 1's and 0's, was transmitted to enable the receivers to be tuned and any necessary adjustments made to the receive terminal. At a predetermined time a synchronizing pattern was transmitted, followed by a test message which comprised the 1024 ten-bit binary numbers from 0 to 1023 in sequence. At the receiver the received message was compared with an identical internally generated message. If an error occurred in transmission the comparator generated an error pulse which was fed to an error counter. The received messages, including the synchronizing patterns, were also recorded simultaneously on magnetic tape for more detailed analysis on playback.

3.1. *Summary of Nomenclature*

Throughout this and subsequent sections, and in Figs. 3 to 6, the following abbreviations have been used:

A*i*	AND gate No. *i*.
BC	Binary counter.
C*i*	Counter No. *i*.
D	Delay.

E	Exclusive or gate-comparator.
ERA	Error rate analyser.
F*i*	Flip-flop No. *i*.
G*i*	Pulse generator No. *i*.
O*i*	OR gate No. *i*.
P	Phase splitter.
RC	Ring counter.
SPG	Synchronizing pattern generator.
SPR	Synchronizing pattern recognizer.
SR	Synchronous regenerator.
TMG	Test message generator.
TR	Tape recorder.
XO	Crystal oscillator.

3.2. *Send Terminal*

3.2.1. Digital equipment

A block diagram of this equipment is shown in Fig. 3. The output from a 10 kc/s crystal oscillator, XO, was divided by 100 (or 200) and used to trigger the clock pulse generator G1. The switch S1 was initially depressed setting flip-flops F1 and F2 and routing 100 (or 50) bits per second reversals via AND gate A1, OR gate O1, AND gate A6 and OR gate O2 to the output. When the test message was to be sent switch S2 was depressed, opening gate A2 so that the next clock pulse re-set F1 which:

(a) closed gate A1;

(b) after a 2 ms delay, D, caused F3 to set and open gate A3, thereby admitting clock pulses to the synchronizing pattern generator, SPG; and

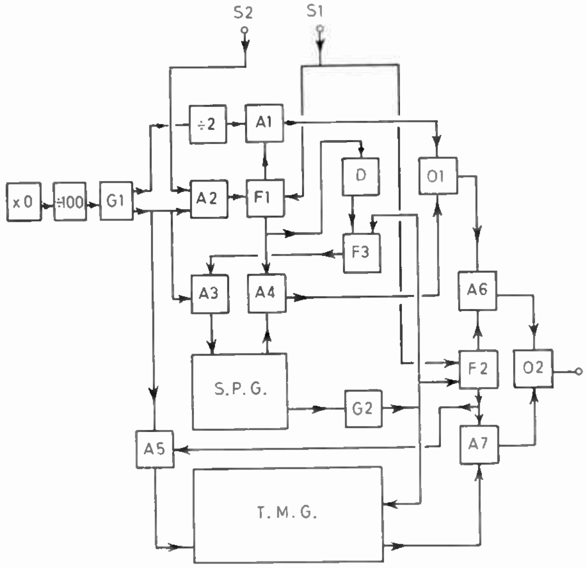


Fig. 3. Send terminal digital equipment.

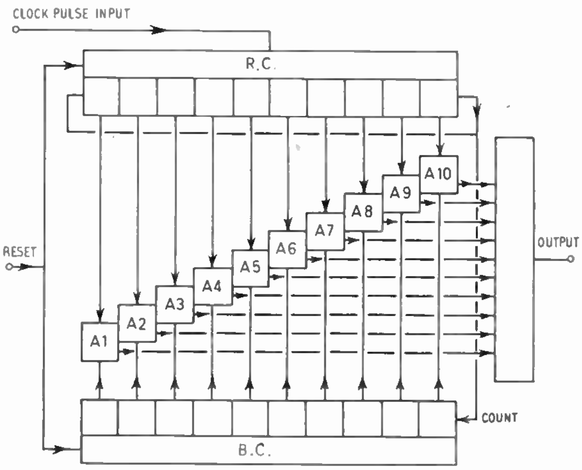


Fig. 4. Test message generator.

(c) opened gate A4, thereby routing the synchronizing pattern via O1, A6 and O2 to the output.

The SPG comprised a shift register containing the required synchronizing pattern and was stepped by the clock pulses.

At the end of the synchronizing pattern, the SPG provided a re-set pulse which triggered G2, causing:

(a) the test message generator, TMG, to be set to its initial condition;

(b) F2 to re-set, thereby opening gate A5 and admitting clock pulses to the TMG and opening gate A7, thus routing the test message to the output; and

(c) caused F3 to re-set.

A block diagram of the TMG is shown in Fig. 4. The ring counter, RC, had one stage in the 1 condition with all other stages in the 0 condition. It was stepped by clock pulses and as it stepped it opened each of the AND gates A1 to A10 in turn. It thereby sampled the condition of each stage of the binary counter, BC, in sequence, the number stored in the binary counter appearing serially at the output. As the 1 condition stepped from stage 10 to stage 1 of the ring counter it added 1 to the number stored in the binary counter, and hence the numbers 0 to 1023 appeared in sequence at the output.

The d.c. levels which appeared at the output modulated the two tones of the baseband modulator, the baseband then modulating the transmitter carrier.

3.2.2. Baseband modulator

The baseband modulator used was a single channel of a six-channel two-tone amplitude-modulated voice-frequency telegraph system (S.T.C. type TA–5).

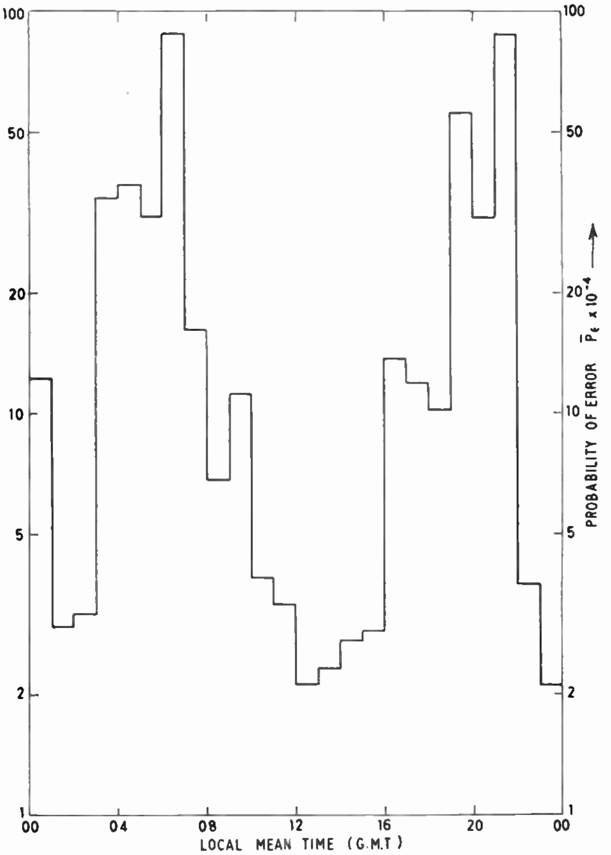
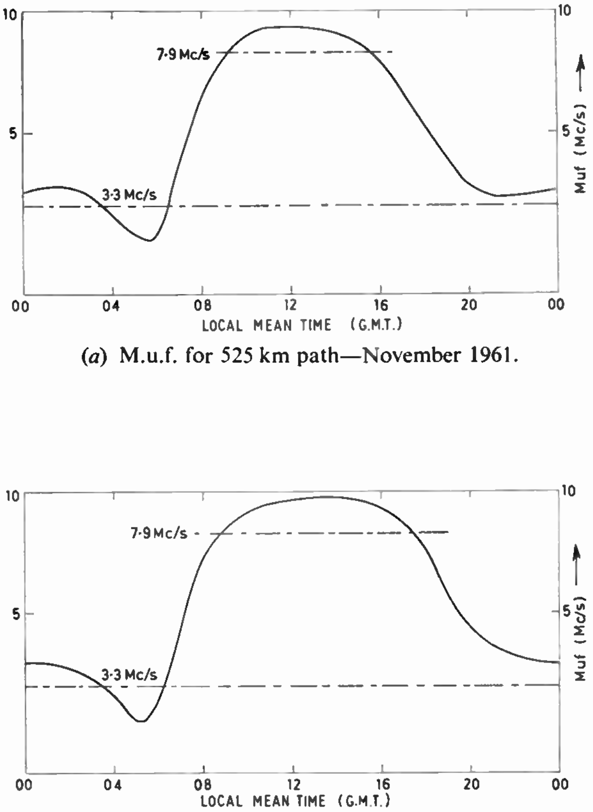


Fig. 7. Mean error rate vs. time of day. Frequency 3·3 Mc/s. 50 bits per second. Relative power − 6 dB.



(*a*) M.u.f. for 525 km path—November 1961.

(*b*) M.u.f. for 465 km path—January 1962.

Fig. 8. Maximum usable frequencies for paths under investigation.

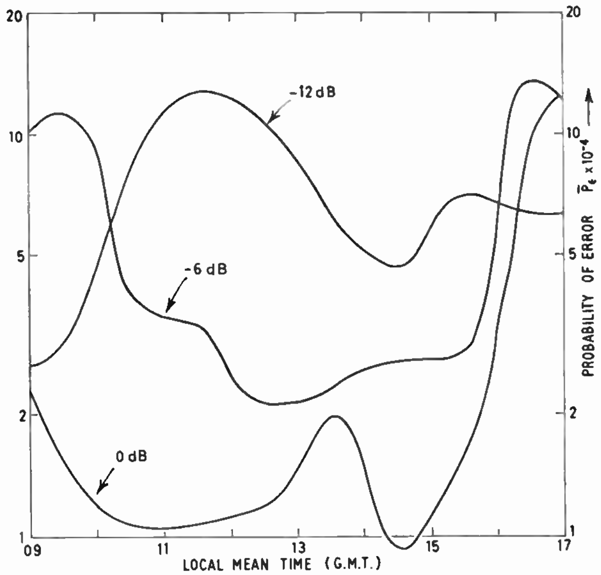


Fig. 9. Mean error rate vs. time of day. 0900 to 1700 G.m.t. Frequency 3·3 Mc/s. 50 bits per second for 0, −6, −12 dB relative powers.

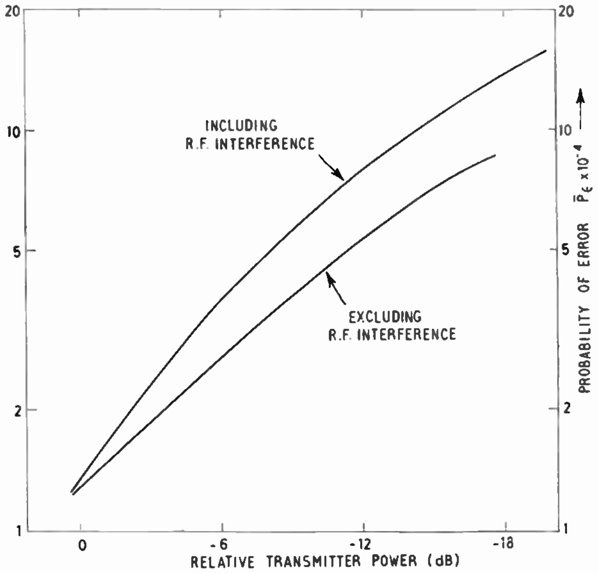


Fig. 10. Mean error rate vs. transmitter power. Frequency 3·3 Mc/s. 50 bits per second.

Table 1

Receiving station A. Transmitter frequency 3·3 Mc/s. Transmission rate 50 bits per second. Transmitter power -6 dB referred to 300 W.

Time G.m.t.	Total bits transmitted $\times 10^4$	Errors	Error rate $\times 10^{-4}$
00–01	20·5	251	12·2
01–02	32·8	97	2·96
02–03	19·5	62	3·18
03–04	16·4	567	34·6
04–05	17·4	643	37·0
05–06	23·5	728	31·0
06–07	23·5	2029	88·2
07–08	30·7	499	16·3
08–09	33·8	230	6·81
09–10	60·4	684	11·3
10–11	62·5	243	3·89
11–12	89·1	298	3·34
12–13	96·3	209	2·12
13–14	65·5	152	2·32
14–15	57·3	156	2·72
15–16	107·6	310	2·88
16–17	82·9	1141	13·7
17–18	45·1	536	11·9
18–19	29·7	304	10·2
19–20	28·7	1609	56·1
20–21	31·8	964	30·6
21–22	33·8	2951	87·3
22–23	32·8	123	3·75
23–00	36·9	78	2·11

Table 2

Receiving station A. Transmitter frequency 3·3 Mc/s. Transmission rate 50 bits per second. Transmitter power 0 dB referred to 300 W.

Time G.m.t.	Total bits transmitted $\times 10^4$	Errors	Error rate $\times 10^{-4}$
09–10	28·9	44	1·53
10–11	49·2	52	1·06
11–12	36·9	40	1·08
12–13	31·8	38	1·20
13–14	36·9	73†	1·98
14–15	32·8	30	0·92
15–16	32·8	55	1·68
16–17	33·8	336	9·94

† Includes 69 errors which were recorded in one hour while the station was unattended.

Table 3

Receiving station A. Transmitter frequency 3·3 Mc/s. Transmission rate 50 bits per second. Transmitter power -12 dB referred to 300 W.

Time G.m.t.	Total bits transmitted $\times 10^4$	Errors	Error rate $\times 10^{-4}$
09–10	25·6	77	3·01
10–11	53·2	462	8·68
11–12	42·0	536	12·8
12–13	45·1	479	10·6
13–14	37·9	235	6·20
14–15	43·1	200	4·64
15–16	49·2	351	7·14
16–17	28·7	183	6·38

made the period from 0900 to 1700 hours the most useful for transmission, and it was therefore decided to confine observations to this time.

Tables 2 and 3 show similar results to those of Table 1, with transmitter powers of 0 and -12 dB respectively for the period 0900 to 1700 hours. These results, together with those for corresponding times from Table 1, are plotted in Fig. 9.

Reducing the transmitter power by 6 dB had the effect of rather more than doubling the error rate throughout the greater part of the period with the exception of early morning and late afternoon. Surprisingly, a reduction in power from -6 to -12 dB during these times apparently lead to a reduction in error rate. This could be due to a combination of three factors. Firstly, increasing the transmitter power could obviously not reduce the error rate since the main source of errors was distortion due to multi-path propagation. Secondly, as can be seen from Table 2(*b*), only about 27×10^4 bits were transmitted at -12 dB during each of these times and these samples are probably too small to be representative. In fact a wide divergence of results was obtained on the two occasions on which this transmitter power was used at these times. Thirdly, a.g.c. was not used as it caused receiver paralysis following noise bursts, so receiver non-linearity might have contributed to the distortion during these low absorption periods at high transmitter powers.

The peak at 0 dB between 1300 and 1400 hours is due to 69 errors which were recorded in one hour while the receiving station was unattended.

At low transmitter powers, r.f. interference, namely interference from other stations, caused a large proportion of the errors. A log was kept in which the causes of errors were noted where possible and Fig. 10

shows the relationship between transmitter power and mean error rate for the period from 0900 to 1700 hours, both including and excluding errors due to r.f. interference. These results are also shown in Table 4.

Table 4

Transmitter power versus mean error rate for period 0900 to 1700 hours G.m.t. including and excluding errors due to r.f. interference.

Transmitter power, dB:	0	− 6	− 12	− 18
Total bits transmitted $\times 10^4$	249·3	538·7	296·1	114·6
Errors	332	2025	2340	1566
Mean error rate $\times 10^{-4}$	1·33	3·76	7·90	13·7
Total bits for which cause of errors was observed	249·3	513·1	276·6	—
Errors due to causes other than r.f. interference	322	1393	1479	—
Error rate computed in absence of r.f. interference	1·29	2·72	5·35	—

Tables 5 and 6 show the distribution of errors in blocks of 3×10^3 bits (one minute) and approximately 2×10^4 bits (seven minutes). Blocks of this length are of interest since the minimum useful message length will be of this order.

5.2. *Group II*

Table 7 summarizes the results obtained at Station B. Insufficient data were collected to give a realistic indication of the hourly variations in error rate and hence the mean values for the hours 0900 to 1700 have been tabulated for a transmitter frequency of 3·3 Mc/s, while for a frequency of 7·9 Mc/s the mean value is calculated for the hours 0900 to 1600, i.e. the period for which 7·9 Mc/s was below the m.u.f. as found from Fig. 8(*b*).

Table 7

Receiving site B. Mean values for the period 0900 to 1700 hours G.m.t. for a transmitter frequency of 3·3 Mc/s and for the period 0900 to 1600 G.m.t. for a transmitter frequency of 7·9 Mc/s.

Transmitter frequency Mc/s	Transmission rate bits/second	Transmitter power dB	Total bits transmitted $\times 10^4$	Errors	Error rate $\times 10^{-4}$
3·3	50	− 6	308·2	1181	3·83
3·3	50	− 6	263·1	797	3·03†
3·3	50	− 6	143·3	1734	12·1‡
3·3	50	0	110·6	200	1·81
3·3	100	− 6	49·2	268	5·45
7·9	50	− 6	105·6	645	6·11
7·9	50	− 6	105·6	212	2·02†

† Excluding errors due to r.f. interference.
‡ Using a short vertical aerial.

The results at 3·3 Mc/s and 50 bits per second indicate that station B was somewhat noisier than station A, the error rates for similar transmitter powers being increased by from 2 to 36%. Since at 3·3 Mc/s receiver noise is negligible compared to external noise, the results for the short vertical aerial indicate that much of the receiver noise input was from local sources.

Table 5

Receiving station A. Transmitter frequency 3·3 Mc/s. Transmission rate 50 bits per second. Transmitter power −6 dB referred to 300 W. Blocks of 3×10^3 bits—one minute transmissions.

Time G.m.t.	Error rate $\times 10^{-4}$	Percentage of error free blocks	Median value of errors per block	Upper decile value of errors per block
09–10	11·3	49	1	6
10–11	3·89	70	0	3
11–12	3·34	63	0	3
12–13	2·12	73	0	2
13–14	2·32		no data available	
14–15	2·72	79	0	2
15–16	2·88	76	0	2
16–17	13·7	40	1	11

Table 6

Receiving station A. Transmitter frequency 3·3 Mc/s. Transmission rate 50 bits per second. Transmitter power −6 dB referred to 300 W. Blocks of approximately 2×10^4 bits—seven minute transmissions.

Time G.m.t.	Error rate $\times 10^{-4}$	Percentage of error free blocks	Lower decile value of errors per block	Median value of errors per block	Upper decile value of errors per block
09–10	11·3	3	2	8	35
10–11	3·89	20	0	4	23
11–12	3·34	5	1	5	16
12–13	2·12	36	0	2	15
13–14	2·32		no data available		
14–15	2·72	33	0	2	14
15–16	2·88	32	0	2	11
16–17	13·7	6	2	12	55

The high error rate at 100 bits per second is mainly due to the larger numbers of errors that occurred during periods of severe distortion in the early morning and late afternoon. The compared messages lost synchronism at an average of approximately once every five minutes at this transmission speed, whereas at 50 bits per second the number of such losses was negligible.

The figures for 7·9 Mc/s are not very informative since about two-thirds of the errors were due to interference from other stations. In this connection it was noticed that the level of r.f. interference could be reduced in a matter of days by transmitting continuously on the same frequency.

6. Conclusion

These tests have demonstrated the feasibility of using h.f. radio for the transmission of low grade digital data over distances of 500 km with bit error rates of the order of 1 in 10^4.

Error rates of this order will probably be satisfactory for the transmission of raw seismic data, but should lower error rates be required further work will be necessary to determine the cheapest and most reliable method of achieving this. The two alternatives are either to improve the performance of the h.f. radio link by the addition of error correcting redundancy, or to use an inherently more reliable system. Since the effectiveness of any error correcting code depends critically on the error distribution, this would have to be analysed and a detailed knowledge of the difficulty of the terrain over which the link is required would be necessary before any choice could be made.

The experiment was limited in scope and the samples obtained rather small. Before any general conclusions can be drawn a far more comprehensive series of tests are necessary, possibly on the lines of the work being attempted in the United States by the U.S. Army Signal Research and Development Laboratory[6] on long distance h.f. radio links.

7. Acknowledgments

Acknowledgment is due to Standard Telephones and Cables for the loan of the equipment and the author would like to thank Mr. W. M. H. Mowat for his co-operation in the experimental work. Acknowledgments are also due to the Royal Aircraft Establishment for the facilities made available at West Freugh and to the United Kingdom Atomic Energy Authority for permission to publish this paper.

8. References

1. D. K. Bailey, "The effect of multipath distortion on the choice of operating frequencies for h.f. communication circuits", *Trans. Inst. Radio Engrs (Antennas and Propagation)*, **AP-7**, No. 4, pp. 397–404, October 1959.
2. E. V. Appleton and W. J. G. Beynon, "The application of ionospheric data to radio-communication problems, part I", *Proc. Phys. Soc. (London)*, **52**, pp. 518–33, July 1940.
3. H. Greenberg, S. Kıevsky and G. B. Bumiller, "Optimum h.f. prediction", *Trans. Inst. Radio Engrs (Antennas and Propagation)*, **AP-10**, No. 3, pp. 325–7, May 1962.
4. J. W. Allnatt, E. D. J. Jones and H. B. Law, "Frequency diversity in the reception of selectively fading binary frequency modulated signals", *Proc. Instn Elect. Engrs*, **104B**, No. 14, pp. 98–110, March 1957.
5. H. B. Law, "The detectability of fading radiotelegraph signals in noise", *Proc. Instn Elect. Engrs*, **104B**, No. 14, pp. 130–40, March 1957.
6. B. Goldberg, "H.f. radio data transmission", *Trans. Inst. Radio Engrs (Communications Systems)*, **CS-9**, No. 1, pp. 21–28, March 1961.

Manuscript first received by the Institution on 6th September, 1962, and in final form on 20th December, 1962 (Paper No. 828).

Nuclear Energy Research in Great Britain

All modern technologies, if they are to make steady progress, must be built on a broad base of scientific research. In many fields this research is undertaken mainly in the universities, but in the nuclear energy field nearly all the major problems need the use of expensive and specialized equipment such as 'hot' laboratories and high flux reactors. Thus much of the basic research in Great Britain in support of nuclear technology must be undertaken by the U.K. Atomic Energy Authority.

The Atomic Energy Research Establishment at Harwell near Oxford, is well known in this country and abroad as a leading centre of advanced work in this field of nuclear energy. Members will of course be aware of the immense amount of research and development into electronic measurement techniques which has been carried out at A.E.R.E. since it came into being in 1946 and the staff of the Electronics Division have contributed many papers to the Institution's *Journal*. The emphasis of present research at Harwell, several items of which were shown to the Press recently, is principally concerned with new materials for reactor technology, and instrumentation work is mainly in the consolidation and extension of established techniques. A number of interesting and novel approaches were noted, however, and, for example, the electrical and optical effects produced by individual particles as they pass through matter are being studied with a view to exploiting these effects for counting purposes. Similarly means for measuring time intervals shorter than a nanosecond are being investigated and new methods for storing and sorting electrical impulses are being examined in connection with collecting and analysing the data obtained from particle counters.

Sonic Spark Chambers

Spark chambers indicating the paths of charged particles are finding increasing uses in both medium and high energy nuclear physics experiments. In conventional spark chambers the position of the sparks are recorded photographically and in many cases the analysis of the photographs is both arduous and time consuming. Development work is being pursued at A.E.R.E. and elsewhere on alternative methods to determine the positions of the sparks and the 'sonic method' shows great promise. In this technique the time of arrival of the sound wave from a spark is determined for a number of sound detectors situated at the edges of the spark chamber. The measurement of these times allows the position of the spark to be calculated in a way similar, in principle, to that employed in radio navigational aids.

The time interval between the occurrence of a spark and the arrival of its sound wave at a particular detector is determined to the nearest microsecond (equivalent to 0·3 mm) by means of a scaler which is switched on by the spark and off when the sound arrives at the detector. In this way the time intervals for each sound detector are measured on separate scalers. The position of a spark can be reconstructed on an oscilloscope for direct viewing and visual monitoring of the events.

One problem associated with the technique is that a large amount of data, from which the spark locations can be deduced, is accumulated in a very short time and the processing of the data must therefore be done by automatic means. The number of microsecond pulses stored in each scaler, after the occurrence of a spark, is read out on to a punch or paper tape system ready for analysis on a computer. A second method in which the information is passed on-line to a computer is being developed and a further gain in speed of processing will be obtained.

Neutron Diffractometer

The atomic arrangement in solids is normally determined by the technique of x-ray diffraction. There are, however, particular problems for which only the neutron diffraction method can be used. One of these is the study of the crystal structures of substances containing both heavy and light atoms, e.g. UO_2.

The neutron method requires a nuclear reactor as a source of thermal neutrons. This is an extremely expensive item, which must be used to the best advantage by designing the neutron diffraction equipment to run automatically, and continuously, throughout the three-week cycle of the reactor.

A single crystal of the material under investigation contains a thousand or more atomic planes, and to determine the crystal structure requires measurement of the number of neutrons scattered in turn by each plane. A given plane is moved into the correct measuring position by rotating the crystal by pre-computed amounts about the three 'Eulerian axes'. The neutron detector is connected to the ω-axis of the centre table by a 2 : 1 gearing, which ensures that, with the atomic plane in the measuring position, the detector is automatically in the correct position for counting the neutrons diffracted by that plane. The positioning of the plane is achieved by a control console, similar to that used in the Ferranti machine tool control system, which receives instructions on punched paper-tape and records signals from a moiré positioning system on each of the Eulerian axes.

The sequence controller controls the counting procedure for each plane. This procedure comprises two peak counts and two background counts; all four counts are printed in tabulated form on the teleprinter in the format: 1. first background, 2. first peak, 3. second background, 4. second peak. A check on the reliability of the equipment is that counts 1 and 3, and 2 and 4 are the same within a few per cent. At the end of the counting procedure, which normally lasts about 15 minutes, the tape-reader reads in the instructions for the next atomic plane and the sequence for positioning, counting and printing-out is repeated.

The output from the teleprinter appears both on the printed sheet and the punched paper-tape. The paper tape is in a suitable form for feeding directly into the *Mercury* computer, which is programmed to give the atomic structure of a crystal from results representing the number of neutrons diffracted by each of its atomic planes.

U.D.C. 621-52: 681.3

The Development of ARCH—A Hybrid Analogue-Digital System of Computers for Industrial Control

By

G. B. COLE, B.Sc.(Eng.)†

AND

S. L. H. CLARKE, B.A.
(*Associate Member*)†

Presented at the Convention on "Electronics and Productivity" in Southampton on 18th April 1963.

Summary: ARCH is a hybrid analogue-digital system of computers for control applications. The need for both a design philosophy, and a range of modules with which a computer control system can be built without specialized circuit, or logical knowledge is discussed. The development of this philosophy is described with the economic and engineering implications which affect it. The basic structure of the resulting system is described in some detail, together with some novel features of individual modules. In particular the "link" units which connect together the various controllers and regulators in a system to form a hierarchy of control are discussed. Some typical applications are referred to by way of example.

1. Introduction

In early 1960, the need for a coherent plan for the eventual computer control of industrial processes was already apparent and the National Research Development Corporation decided to sponsor the development of a system capable of effecting progressive automation. The proposal for this system had grown out of many years experience in the design and development of computers on the one hand and of conventional industrial instrumentation and control on the other. This need and the development of the system, ARCH, are described in this paper, together with some of the initial applications and finally an assessment is made of the extent to which the resulting equipment is capable of meeting the need.

2. The Need for ARCH

During the few years before 1960 the use of digital computers in industrial processes was being talked of a great deal and a few, rather isolated, examples of practical applications had taken place. Perhaps the most significant facts which led to this state of affairs were the increasing general awareness of the possibilities of general purpose computers, and, even more important, the advent of transistor computers with their increased reliability. It became possible to think in terms of the reliabilities of better than 99% which would be desirable, if not necessary, for continuous on-line operation. In the vacuum tube era, computers had been used for data reduction in association with wind tunnels, and there were few examples of on-line use. One of these was at the National Gas Turbine Establishment where an Elliott 405 computer was installed in 1957 to present a certain amount of calibrated data during the course of a run in addition to recording data on magnetic film for subsequent analysis. In this case the computer was only required to run for short periods at a time, so that with extensive preventative maintenance it was possible to produce a sufficiently reliable system.

Whilst this system was a far cry from the integrated process control systems envisaged today, it contained all the essential features, and presented all the problems involved today. Although a general-purpose computer was used it was necessary to make certain modifications, and certain features of the machine were not used in practice. The need for different forms of analogue-to-digital conversion and for digital servo-control of external devices was encountered and satisfied, although *ad hoc* methods had to be used at times.

When the possible on-line application of computer continuous processes was reviewed in the light of such experience there were two main decisions which had to be made.

(*a*) Should process control computers be analogue or digital and, if digital, at what point should the conversion be made?

(*b*) Should control be vested in a single large computer or in a network of semi-autonomous machines?

The conclusions arrived at were to some extent interlinked, but it is perhaps easier to deal with the second first. It is theoretically possible for a single computer, if large enough, and fast enough, to carry out the whole of a control task. If one could achieve complete computer control, such a central computer might be more economical in capital costs. Experience shows that such "complete control" can never be specified as complex control is essentially evolutionary.

† E-A Data Processing Ltd., Borehamwood, Hertfordshire.

The reason for the economy is that equipment can be time-shared between different facets of the operation; better utilization of the individual components of the system can be obtained in this way. However, before embarking on such a course of action, one must consider the time when, for one reason or another, the plant is not under complete computer control. Initially one will not know how to control a whole process in an integrated fashion, and it will be necessary to build up, step by step, over considerable periods of time, particularly on existing plants. This will be uneconomical deployment of capital if a large system is not fully utilized; the scheme will not go ahead because the initial capital cost is too high compared with the savings which can be guaranteed. When the plant has been commissioned under complete computer control, one is then faced with the problem of computer failure. It is not feasible at present to produce computers which will never fail and probably never will be so. Consequently it is vital to establish a procedure whereby essential safeguards are present in the organization of control such that both life and property are adequately protected, and preferably such that the cost of failure is minimal. Unfortunately in any system where there is shared equipment the effect of failure must be more widespread than where individual components are used. Consequently one must either introduce redundancy on a large scale to reduce the probability of a system failure sufficiently, or adopt a system whereby a stand-by service can be put into operation by human intervention. The former action will further increase the initial capital cost of equipment, whereas manual stand-by, if it can be limited to the capacity of essential maintenance personnel, will not increase costs significantly.

It was therefore decided to devise a system where this reversionary mode was possible. This inevitably led to the choice of a network of semi-autonomous computers working in such a way that if any one machine fails then human intervention can provide rudimentary control to maintain plant operation, while the other computers can adjust their own control parameters to the non-optimum running condition of the affected area of the plant. This organization is analogous to a management hierarchy, and lends itself to the replacement of one defective element by a reserve, as well as to the re-adjustment of responsibilities to manage while such an element is out of action. A further advantage of the hierarchy system is that it enables the commissioning, and further investigation of different sections of a process, to be carried out in parallel rather than in series.

The question of the desirability of analogue or digital systems is, in itself, simply answered, since nearly all plant quantities are analogue by nature, and thus part of the system must be analogue. On the other hand, in any integrated scheme management and accounting figures must be processed, and thus the system must also be digital in part. However, one is then left with the supplementary question of where the transition should take place. One school of thought maintains that the digital computer should read the plant quantities directly and drive the valve actuators and other controls. On the other hand one

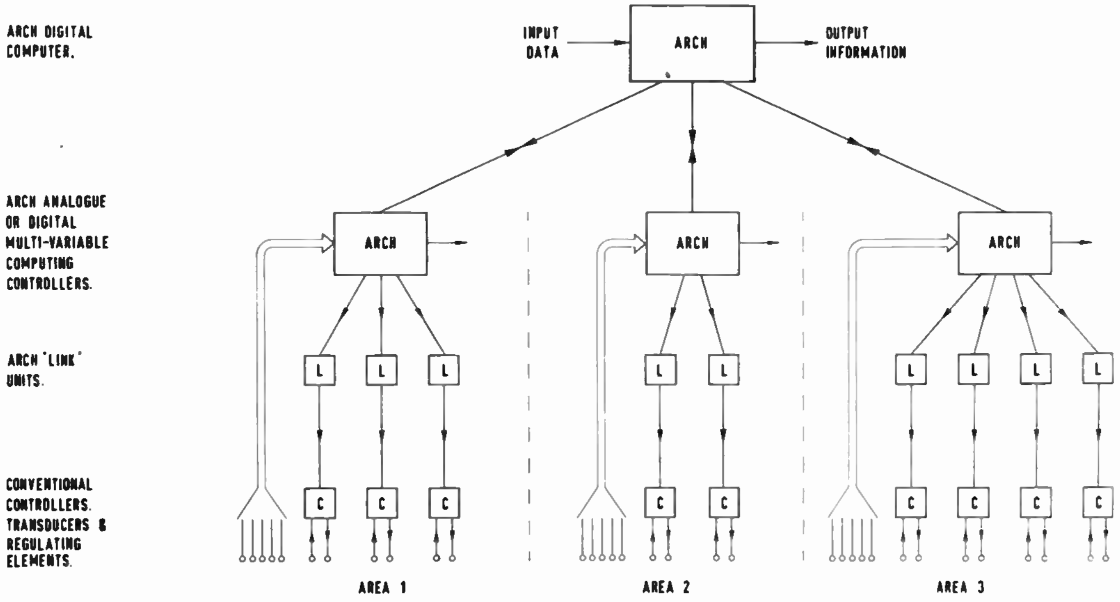

ARCH DIGITAL COMPUTER.
INPUT DATA
ARCH
OUTPUT INFORMATION
ARCH ANALOGUE OR DIGITAL MULTI-VARIABLE COMPUTING CONTROLLERS.
ARCH
ARCH
ARCH
ARCH 'LINK' UNITS.
L
L
L
L
L
L
L
L
L
CONVENTIONAL CONTROLLERS. TRANSDUCERS & REGULATING ELEMENTS.
C
C
C
C
C
C
C
C
C
AREA 1
AREA 2
AREA 3

Fig. 1. A typical control hierarchy.

could carry out many of the control computations by analogue computer, and only convert, to and from digital form, the data essential for management functions. However, it was felt that the first type of system was very similar to the large single computer in that in the event of computer failure the continuous control of all points in even a sub-system would be too great a task for the human operator, even assuming valve actuators which would fail safely in the last operating position. On the other hand the flexibility of the programmed digital computer leads one to make as much use of it as possible. This is particularly true where the role of the computer grows as experience builds up, from being an investigation tool to the controller over a long period of time.

From these considerations the requirements for a system may be evolved whereby hierarchies of computers can be built up above conventional controllers as the basic elements, control being exercised on the set points (Fig. 1). It was considered necessary for the system to be able to contain analogue computers for certain continuous computations. These would normally be at the lower levels of the hierarchy. It is necessary then to devise a system where a common language, or languages, exist between elements in the hierarchy and where it is possible to make elements of varying capacities.

3. The Required Functional Specification

The system can be broken down into the following groups of equipment:

(*a*) Digital elements (*c*) Conversion elements
(*b*) Analogue elements (*d*) Transducers.

The last group is, by definition, taken to mean conventional instruments, including controllers, and will not be discussed here. The conversion elements presented no special problems since voltage and shaft digitizers were, by this time, quite well known. However, routine development work was necessary to produce the range, described later in the paper, together with more fundamental work on noise problems arising in low-level plant signals.

The problems presented by the analogue elements again included much development work, since there did not exist at that time a fully transistorized range of analogue computing elements. There was little need for a specialized system for making computers of variable capacity since this is an essential feature of analogue computers as opposed to digital computers. Also it was possible to neglect integration over periods of more than an hour or so, since this is handled better by digital systems. Consequently the problem was reduced considerably from the production of a general-purpose analogue computing system to that of a system capable of co-ordinated control of a small number of points.

In the case of the digital machines a more substantial barrier was encountered. At the time there appeared to be a basic minimum price for a stored programme digital computer of £20,000 to £30,000 which was too high for the basic building brick. Also the capacity of such a machine is too great in many instances. However, if one was to design a special-purpose computer with more limited capacity for every element in a hierarchy the cost of development would rapidly become prohibitive using the existing methods of computer design. It was therefore necessary to avoid individual design for each element by producing much larger standard modules than the usual plug-in elements of commercially available computers. In this way the process system analyst could specify the capacity, and therefore the constitution, of the computers in a hierarchy without being experienced in logical design of pulse circuitry. To achieve this it was desirable that a typical machine should consist of from 6–18 modules. In order that this concept could work it was necessary to produce a standard framework for these modules in three senses—logical, mechanical and program.

In the logical sense the plan had to envisage the largest type of system without making the smallest system carry a lot of redundant equipment which could only be used in large systems. It should be made possible to add extra modules to the arithmetic unit as well as to the store of the machine without involving modifications to the internal arrangements of an existing system.

From the mechanical point of view, besides the obvious desirability of standard sizes in modules, it was imperative that the inter-connection of modules should be in standard form to avoid involved and intricate wiring.

Finally, from the point of view of programming, it was essential that order codes should be standardized throughout the whole range; time would be wasted through the use of different codes in a hierarchy, since the interconnection of the different machines necessitates similarity in modes of operation.

There is a distinct advantage in standardization when the time comes to design a special unit of the system. In all probability the result will not be the most elegant design possible, but a great deal of time will be saved. In fact if all the special purpose computers were designed from basic logic elements not only would the cost be much higher but the task would be impossible with the amount of skilled effort available.

4. Initial System Design

A logical design study was carried out to prove the feasibility of this type of system. Existing techniques of serial logic circuits and nickel delay lines registers were used and a typical small system was actually built from 405 type valve equipment. The choice of a serial system at this stage was governed by considerations of economy in the digital system together with the fact that a suitable form of fixed store was already available. The need for limited quantities of fixed storage in small digital systems, such as "interlock sequencers", as well as in larger quantities for big systems had been realized. A technique of using nickel delay lines,† giving a serial acoustic pulse train from magnetized areas as the result of an electrical pulse on the line, had been developed, and lines containing 500 digits of permanent storage were produced. This system of storage was compatible with that of 405 type logic and was capable of further development for a faster transistorized machine.

This prototype system is briefly described below and forms an interesting comparison with the parallel system which eventually emerged from the production system, and is described from Section 6 onwards.

4.1. *Logical Structure*

Each group of registers is capable of working independently although co-ordinating control may be exerted by a computer further up the hierarchy. The "register group" consists of a number of self-contained registers connected together by two busbars, and a number of common control lines.

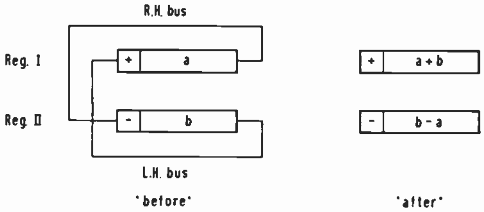


Fig. 2. A typical register pair.

Each operation of the device allows the contents of two registers to proceed along the two busbars respectively, while the inputs to the registers accept the information from the other busbar. These inputs pass through a function box to be combined with the stored contents according to the function combination of the particular register.

For example the two registers have "add" and "subtract" functions and original contents "a" and "b" (Fig. 2). The result of operating on Registers 1 and 2 is to place (a+b) in Register 1, and (b−a) in Register 2.

4.2. *Interconnection between Units*

Interconnection between units of the group is standard. Connections fall into three categories:

(*a*) Information channels,

(*b*) Control waveforms, and

(*c*) Power supplies and standard waveforms.

4.3. *Information Channels*

There are only two channels of information within a register group. These are the right- and left-hand busbars. These lines are joined to the inputs and outputs of all registers but it is only possible to have one register switched on to each line at a time.

4.4. *Control Waveforms*

The selection of registers and functions fall into the category of control waveforms. Four pairs of lines carry the necessary signals to select the two functions and their inverses for the left- and right-hand registers. Once inside the selected register these function waveforms set up the function appropriate to that register irrespective of its number in the group.

The register selection waveforms are carried on two sets of twelve wires. These wires go to all registers, and feed interchangeable units which perform the actual selection of the right- and left-hand registers. This means, for instance, that the number ascribed to a particular type of register need not be the same in all register groups, and also that the registers can be tested completely out of the context of the particular register group from which they came.

5. The Prototype System

After the experience gained on the valve equipment and preliminary experiments with a serial logical system working with a 1 Mc/s clock, it was decided that the desired system was feasible. It was decided, therefore, that a prototype system of both analogue and digital modules should be built, together with the necessary link units. By this time the asynchronous "Minilog" range of logical elements was well in to the production phase and was considered to be the best type of element for ARCH. Since the required word length of 18 bits was comparatively short, a parallel system would not be uneconomical. Such a system was chosen to match the core store systems and the input/output devices, which are usually parallel devices, of which many were already using "Minilog" elements. In addition a parallel machine could more readily use the fixed core store which had by this time been developed, and promised to be more economical than the static delay lines. This store, together with the other more important modules of the system are described in the following sections.

6. Digital Modules

The change from serial to parallel operations inevitably led to major changes in the logical structure of the system. The advantages to be obtained in delay line registers were no longer available and so the double busbar arrangement between registers was

† British Patent No. 862,364.

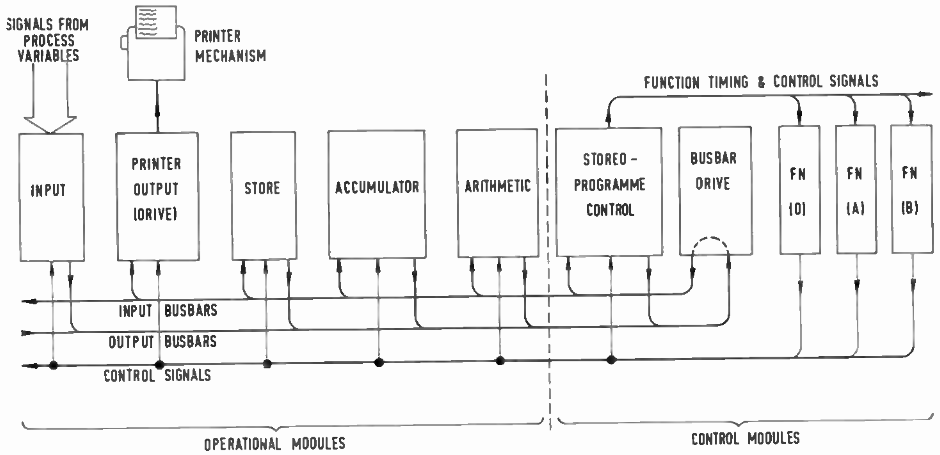


Fig. 3. An assembly of ARCH digital modules.

dropped in favour of a single set of parallel lines. In addition much of the control information is transferred on the same set of lines. There are two basic types of module: operational and control. The operational modules carry out specific input, output, storage or arithmetic operations whilst the control modules control the flow of data to and from the operational modules.

These units are described in the following section, when it will be clear that the three types of standardization laid down in Section 3 are still present in a slightly modified form.

6.1. *Operational Modules*

All operational modules are connected to a common system of data busbars via which the transfer of all data takes place (Fig. 3). The passage of data from one module to another is in 18-bit parallel digital form. The data pass from the outputs of the operational modules on to the output busbar, via the busbar drive unit on to the input busbar, and thence to the inputs of the modules. The busbar drive module is a set of power amplifiers which enable large numbers of modules to be connected to the busbars, without seriously loading their outputs. The only other connections to the operational modules are the control signals from the control modules and, in the case of input/output modules, the connections to the input/output devices.

All operational modules have the same basic structure typified by the module shown in Fig. 4. The parallel nature of both the module and busbars is clearly shown. The "device", whether it be for input, output, storage or arithmetic is always associated with one, and in some cases two registers for holding data whilst the module performs its operation. At the beginning of such an operation the register is reset to zero by the control signal R, the input gate is then opened by the signal G.I. and the register is set to the number currently on the busbars. The module then carries out its operation, at the end of which there may be new data to be fed on to the busbars by the opening of the output gate by signal G.O. It should be noted that the busbar drive module inverts the pulses so that a "one" which is a negative-going pulse on the output busbars becomes a positive-going pulse in the input busbars.

6.2. *Control Modules*

A digital computer constructed from ARCH modules is usually controlled by a program of instructions stored in one or more store modules. The stored program control module, contains the instruction sequence control register, the instruction register and the function timing and control units (Fig. 5). There are two beats in the control cycle. During the "read" beat an instruction is read out from a location in a store module into the instruction register, the address of the location being specified by the contents of the sequence control register. "One" is then added to the contents of the sequence control register. During the "obey" beat the instruction is decoded and obeyed.

A single address instruction code is used, each instruction consisting of a function (5 bits) and an address (13 bits). It can be seen from Fig. 3 that a

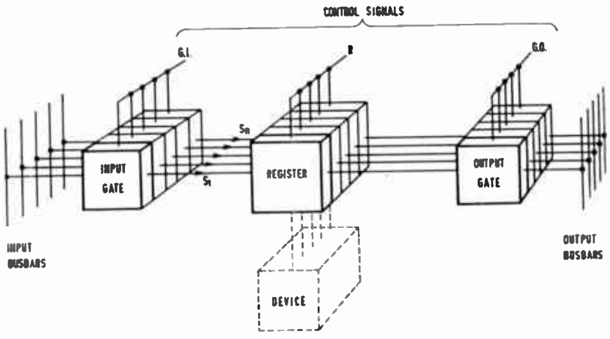


Fig. 4. A typical operational module.

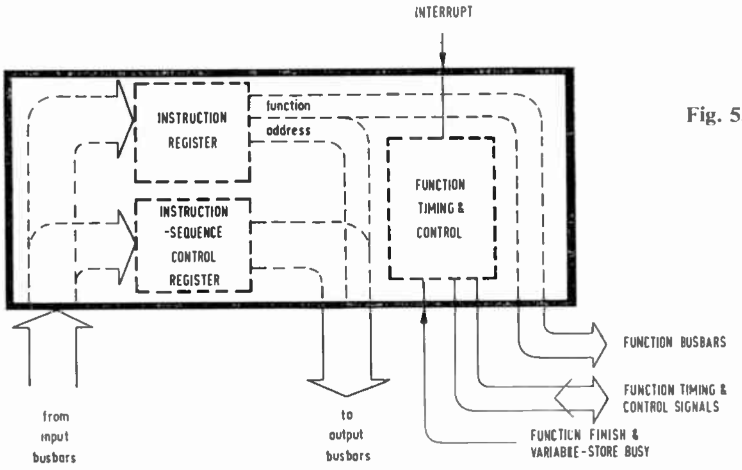

INTERRUPT
INSTRUCTION REGISTER
function
address
INSTRUCTION -SEQUENCE CONTROL REGISTER
FUNCTION TIMING & CONTROL
FUNCTION BUSBARS
FUNCTION TIMING & CONTROL SIGNALS
from input busbars
to output busbars
FUNCTION FINISH & VARIABLE-STORE BUSY

Fig. 5. The stored program control module.

number of function modules are connected to the stored program control via function timing and control lines. During the "obey" beat the function control lines feed the function part of the current instruction in the instruction register to the function modules. All function modules receive these signals and decode them, but only the one having the code corresponding to the current function is selected and energized. The function timing lines transmit the output of the timing unit in the stored program control to all the function modules. Under the control of these timing signals the selected function module feeds the appropriate operational modules with the various control signals, e.g. R, G.I., G.O., etc.

The address part of the current instruction is gated out of the instruction register on to the data busbars under the control of the selected function module. The appropriate operational module receives the address and prepares to act on the data which appears on the busbars a little later. In the case of store modules the address specifies the location required, and in the case of input/output modules it specifies the particular one required.

6.3. *Some Examples of Digital Modules*

The range of modules has been developed with particular reference to the construction of computers for on-line process control and monitoring.

6.3.1. Fixed store

The fixed store is a permanent (wired) core store in which the data or instructions are not destroyed by read-out or by switching off. It is normally used to contain fixed programs and fixed constants. This means that there is no need to feed in the program from a reel of paper tape every time the computer is switched on. Also there is no danger of accidentally destroying the program by incorrect operation—a very important feature in on-line machines.

The store is available in sub-modules of 256 words of 18 bits. There is one linear core for each word, each core being threaded with 18 wires either in one direction or the other, corresponding respectively to the "ones" and "zeros" of the stored instructions (Fig. 6). A particular word is read out by pulsing the primary winding of the required core via a diode selection matrix. The parallel output register is set by the outputs from the 18 secondary wires. The store has a good signal to noise ratio because only one core is energized at a time, and "ones" and "zeros" are represented by the presence of a positive or negative pulse respectively.

During manufacture the store is threaded semi-automatically by a weaving device fed from information on paper tape. The paper tape itself is prepared by a computer which is fed with its required data in direct decimal code.

6.3.2. Semi-fixed store

The semi-fixed store is a diode peg-board store used for storing data which may require to be changed at infrequent intervals, for example, alarm limit data for an alarm scanning program.

6.3.3. Variable store

The variable store is a conventional coincident current core store in which the data has to be re-generated after read out. It may be used for storing the results of calculations, histories of plant measurements or instructions during program development.

6.3.4. Arithmetic modules

There are a variety of arithmetic modules for building computers with more or less powerful instruction codes. The basic module provides for "add", "subtract" and "collate". Functions such as multiple shift, multiply, divide and square-root may be added by the use of auxiliary arithmetic modules.

6.3.5. Output modules

When a relatively small amount of output data is required for the guidance of process operators this may be achieved by the use of visual display modules. When large amounts of output data are required, printer modules may be used in conjunction with conventional teleprinters. A module feeding a wide carriage electric typewriter is available for printing periodic logs of measured variables on preprinted stationery. A module for driving a high speed line-at-a-time strip printer is also available for printing alarm information. It should be noted however, that the use of such relatively fast strip printers is normally confined to alarm scanning machines without a variable memory, the high speed being necessary to cope with the peak rate of print-out without slowing down the input scan rate. If the computer contains a store, a print-out queue can be formed, and a slower printer used to deal with the *average* rate of print-out. There is also a module for producing punched paper tape, which may be used for subsequent analysis in a more powerful off-line computer.

Finally there are modules for directly controlling the plant. One bit output modules for on/off operations such as starting and stopping motors, and "link" modules for driving the set points of other controllers in a "fail-safe" manner are discussed in Section 8.5.

6.3.6. Input modules

There are modules for input of data from manual decimal rotary switches, key boards, paper tape, on-off contacts, and from plant transducers, via analogue to digital converter and analogue selection modules described in Section 8.2. Manually operated rotary switches engraved in plant terms have been found to be the most convenient form of input device for use by process operators.

6.3.7. Power supply modules

The power supply modules produce the ± 10 V and − 24 V necessary to drive the various digital (and analogue) modules. The input to the power modules is 115/230 V ± 10%, 50 c/s ± 5% mains. A

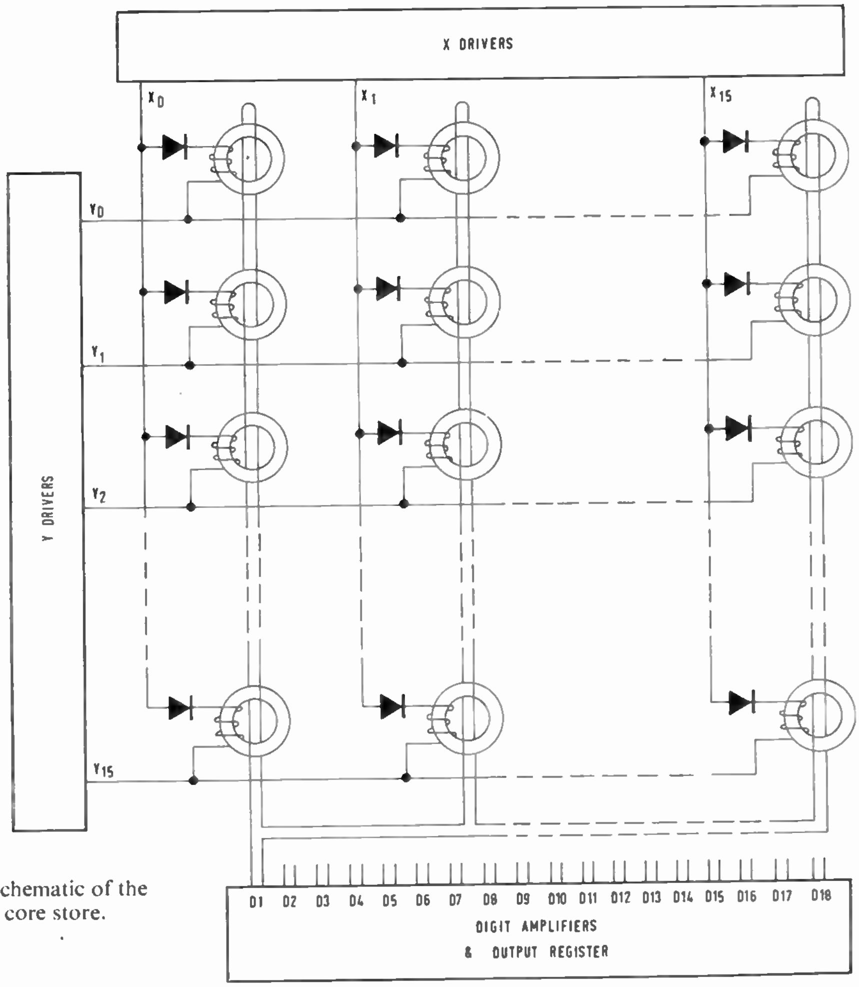


Fig. 6. Schematic of the fixed core store.

constant-voltage transformer may be used where the amplitude variations are in excess of 10%. A motor alternator or a special battery-driven converter module may be used when the interruption of the mains supply is expected.

6.4. *Features of Particular Interest for On-line Process Control Computers*

There are two features incorporated in the stored program control module of particular interest for on-line process control computers.

6.4.1. Priority interrupt

A "priority interrupt" line, which may be connected to a manual push-button, is provided to interrupt the instruction sequence (program). The interruption takes place after the completion of the current instruction, the sequence control register content is retained at the value set after the previously obeyed instructions, but the next instruction is read for location zero of a specified store. A short "red tape" sub-routine is normally stored in location zero onwards. This sub-routine stores away the contents of the sequence control register, accumulator etc., and transfers control to the branch of the program specified by the appropriate input channel. The priority "interrupt" may be locked out by the program and it is also possible to allocate different levels of priority to the various branches of the program.

6.4.2. Watch dog

A "watch dog" unit is provided which constantly attempts to sound an alarm unless it receives regular "all correct" instructions from the main computer. Once per scan of plant inputs, or at some other regular interval, the computer is programmed to perform a self-check routine, which usually includes a check of all arithmetic functions and sample measurements of known test input points. If the result of this self-check is satisfactory an "all correct" instruction is sent to the watch dog. The special "all-correct" instruction is synthesized by the routine and "destroyed" immediately after use, rather than being stored in a location of one of the stores. In this way the chances of accidentally giving the instruction during fault conditions is virtually eliminated.

6.5. *Module Construction and Interconnection*

The digital modules are constructed from sub-module boards which plug into specially wired module chassis, of various sizes. The plug-in boards carry up to 48 individual transistorized logical elements which are connected together by a printed circuit or conventional wiring.

The module chassis with their plug-in boards may be housed either in individual module cases or grouped together in larger cabinets holding between 5 and 20 modules (depending on the size of the modules to be housed). These cabinets themselves are modular and may be grouped together to form single, double or triple bays.

The connections to each module consist óf data inputs and outputs; the control signals and the power supplies. All these connections are taken to a set of pins at the rear of the module chassis. Each of these pins is given a number taken from a master list of signals. Interconnection of modules is achieved by connecting together all pins of the same number in all modules. For instance the data inputs are numbered 1 to 18, hence the input "busbars" are formed by connecting together pins with these numbers on all modules.

It should be noted that the method of inter-module connection is identical whether the modules are housed in individual cases or grouped in cabinets.

7. Analogue Modules

Analogue modules are constructed in the same basic way as digital modules, are driven from the same voltage power supplies and may be housed in similar individual cases or with digital modules in bays of cabinets.

7.1. *Operational Amplifier*

The basic sub-module used in all analogue modules is a transistor d.c. operational amplifier. It has an open-loop gain of over 100 000, a linear output of ± 5 volts up to 100 c/s and a drift, referred to the input, of less than ± 25 μV per deg C.

The d.c. gain may be increased by a factor of 1000 and the drift reduced by a factor of 5 by the addition of a chopper-correcter sub-module.

Analogue addition, subtraction and integration may be performed by the connection of appropriate networks of passive elements to the basic amplifier. The chopper-corrector is particularly useful for longer term integration, up to periods of one hour.

7.2. *Multiplication, Division and Square-root*

The majority of process control applications require single quadrant multiplication and division. Figure 7 shows the block schematic of the basic logarithmic multiplier used. This technique of

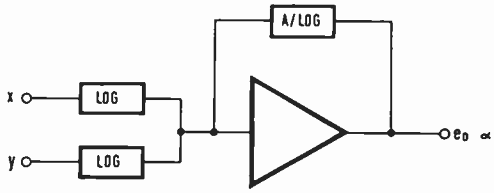


Fig. 7. Analogue multiplier module.

multiplication is very flexible since multiple products such as $x \times y \times z$ can be produced by the addition of an extra input network. Powers and roots can be obtained by scaling; division requires only a change of sign of the appropriate input voltage.

The logarithmic networks employed use a straight line approximation of the functions. Silicon diodes in a temperature controlled oven and precision wire wound resistors are used. The same technique is also employed to perform transducer linearization and other empirical functions.

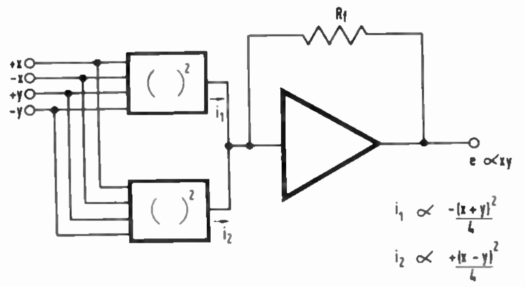


Fig. 8. Quarter squares multiplier module.

Figure 8 shows the "quarter squares multiplier" available for performing the occasional four quadrant multiplication which may be required. Note that the same techniques of using the basic operational amplifier with non-linear networks is used.

7.3. *Time-sharing and the Analogue store*

Process control applications often require the same calculation to be repeated for a number of different measurement points, e.g. the correction of instantaneous flows for temperature and pressure. An assembly of analogue modules may be "time shared" to perform such a calculation and the results stored in a number of analogue stores.

Figure 9 shows an analogue store which is essentially an integrator used only in the "sample" and "hold" modes. Errors of less than ± 1% can be obtained with a sampling period of 15 to 30 seconds. A chopper corrector sub-module may be used to extend the sampling period to several minutes.

8. Conversion and Link Modules

8.1. *The Conversion Problem*

Both analogue and digital computer modules must have common languages. In ARCH these are ±5 V d.c. and 18 bits binary respectively. The problem of conversion between measured signals and both the analogue and digital languages therefore arises, together with analogue-to-digital and digital-to-analogue conversion and the linking of the outputs of both analogue and digital computers to other controllers or directly to the plant.

Measured signals may be obtained directly from primary transducers such as thermocouples, resistance thermometers, or from existing plant instrumentation, often via a secondary transducer, e.g. pneumatic-to-electric converter. Even when electronic instrumentation is used some form of conversion will be necessary from one of the many "standards"—e.g. 0–10 mA d.c., 4–20 mA d.c., 0–50 mV d.c., 0–500 mV a.c. etc.

8.2. *Signal Amplifier*

The signal amplifier may be used to convert the outputs of d.c. primary transducers and most types of d.c. instrumentation into ± 5 V d.c. It is basically a feedback carrier amplifier in which the d.c. input is modulated by a transistor chopper at 500 c/s and subsequently demodulated and integrated to remove all traces of the carrier frequency. The amplifier has a floating input and has a very high common mode rejection (over 10^6) to both a.c. and d.c. noise signals up to several hundred volts amplitude.

The amplifier may be time-shared (multiplexed) between a number of points using an analogue selection module and may be switched at a rate up to about 20 points/second. The gain of the final stage of the amplifier may be selected remotely to allow for scaling to suit differing transducers.

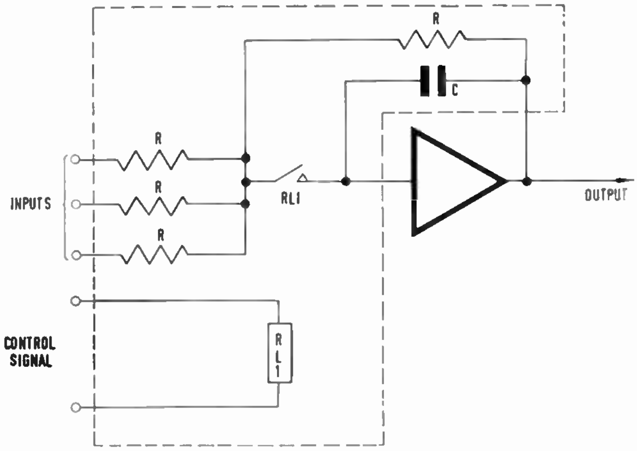


Fig. 9. Sample hold module.

It is often necessary to filter individually the outputs from d.c. instrumentation before feeding the selection module as these may contain up to 30% a.c. ripple. Further series mode noise rejection may be achieved at the expense of switching speed by the use of a common low-pass filter, between the selection module and the amplifier.

Where individual amplification of floating signals from relatively high level sources (over 1 V d.c.) is required, with a modest common mode rejection

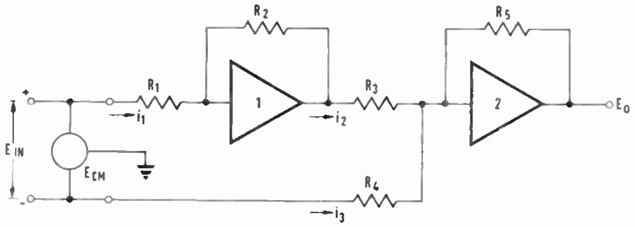


Fig. 10. Operational amplifiers with floating input.

(less than 10^4), then an economic solution may be obtained by using two operational amplifiers, as shown in Fig. 10.

8.3. *A.C.-D.C. Converter*

The a.c.-d.c. converter converts 0–500 mV 50 c/s a.c. signals to 0–5 V d.c. It uses an operational amplifier system with diode rectifier feedback, giving linear demodulation over the full range of input, without any dependance on the phase of the input signal. The converter may be time-shared using a selection module up to a rate of 5 points/second.

8.4. *A.D.C. and D.A.C.*

The analogue-to-digital converter (a.d.c.) works on the well-known potentiometric principle, in which the analogue input signal is compared with a binary sequence of currents derived from a reference power source (Fig. 11). The input is zero to ±5 V d.c. and the output consists of the sign plus 12 bits (resolution of one part in 8192). The module has a conversion time of 1 ms and an accuracy of ±0·05% of full scale. The comparator consists of a network of precision resistors and transistor switches, and a modified operational amplifier.

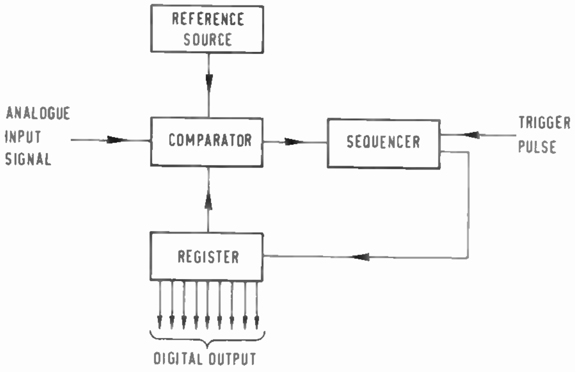


Fig. 11. Analogue-to-digital converter block diagram.

The same sub-modules of reference source, comparator (with a standard operational amplifier) and register may be used in reverse as a digital to analogue converter (d.a.c.) to give out ±5 V d.c. for an input of sign plus 12 bits.

8.5. *Link Module*

When constructing a hierarchy of controllers as described in Section 2, it is necessary to have fail-safe links between the various controllers in the hierarchy.

The link module converts the outputs of the computing modules into a form suitable for the remote set-value adjustment of proportional, two or three-term controllers. It allows the set-value to be raised or lowered in small discrete steps under the control of the computer output signal. The set value is indicated by a pointer and calibrated dial, and adjustable stops are provided to allow the permissible range of set-value variation to be restricted within given low and high limits. Link mechanisms are available to suit all the usual types of controllers, namely:

(*a*) pneumatic 3–15 lb/in^2

(*b*) electronic a.c. 0–500 mV a.c.

(*c*) electronic d.c. 0–10 mA d.c. etc.

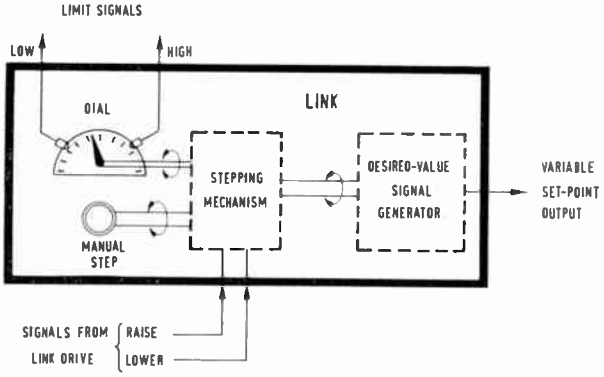


Fig. 12. A link mechanism.

Figure 12 shows the block diagram of an ARCH link mechanism. A stepping-mechanism is pulsed by a drive sub-module, which receives its input from the computer. Two lines are provided, one for "raise" and one for "lower". The low and high limit stops are both mechanical and electrical; when the indicated limit is reached, a micro-switch operates and provides a warning signal to the computer (an alarm signal can also be given to warn the operator). Should the warning signal be ignored, the mechanical stops prevent the variation of the set-value by more than one step beyond the limit.

In use, the link-drive is made inoperative in the event of a computer failure, and the link mechanism remains at its last setting. This arrangement provides the fail-safe characteristic, since the process controller continues to operate safely but not necessarily at optimum performance. The operator can, if necessary, intervene and adjust the set-value by manual operation of the link mechanism, using the control knob provided.

The output from the link mechanism is converted into the standard 5 V d.c. signal and fed back to the computer. This feedback signal is compared with the calculated desired set point, and the deviation is used to control the drive sub-module and hence the stepping mechanism. In order to avoid hunting a dead-band of just over plus and minus half a step is introduced. In digital computers the comparison and control is carried out by the program, whereas in analogue machines a special comparator and drive sub-module is used.

9. Applications

ARCH modules may be used to build a wide range of computers ranging in sophistication from little more than a three-term analogue controller or simple data logger, to full-scale combined analogue and digital systems having thousands of words of stored program. An outline only is given on two simple machines.

9.1. *Twenty-point Flow Integrator*

This is essentially a data logger using time-shared analogue computing modules. It will be installed in two phases. In the equipment making up phase one (Fig. 13) the outputs from twenty pairs of a.c. transducers for ΔP and P are selected at a rate of four pairs of points per second and fed, via a.c.-d.c. converters, to the analogue computing assembly, which perform the well-known flow correction:

$$F = K\sqrt{\frac{\Delta P.P}{T}}$$

The value of T is also selected from one of six manually set inputs. The output from the analogue computer assembly is fed in turn to twenty analogue memories which drive twenty low inertia integrating motors with counters having visual output registers.

In phase two a number of extra digital modules will be added to enable direct print-out of integrated flows and total flows on a typewriter. The ARCH modular approach was chosen in this case because, with the computing modules shared between twenty points, the initial cost compared favourably with conventional instrumentation; expansion to give a direct print-out is readily accomplished.

9.2. *Data Logger with Digital Arithmetic*

There is now an increasing demand for tin-plated steel strip to be supplied in coils, rather than in flat sheets. If efficient use is to be made of tin-plate in this form, an accurate "profile" of all defects must be supplied with each coil. The manufacturers also require a summary of total lengths of defective plate produced under the various defect headings, as well as other data such as average plating thickness per coil etc.

Some defects can be detected automatically whilst others require a human inspector, who signals the presence of a defect by pressing a button. The detectors and inspectors are positioned at a number of points between the plating baths and the coiler and shear. The collection of data is complicated by the fact that it has to be stored for the time taken for the steel strip to travel from a given defect detector (or inspector) to the shear, so that it may be associated with the correct coil after shearing.

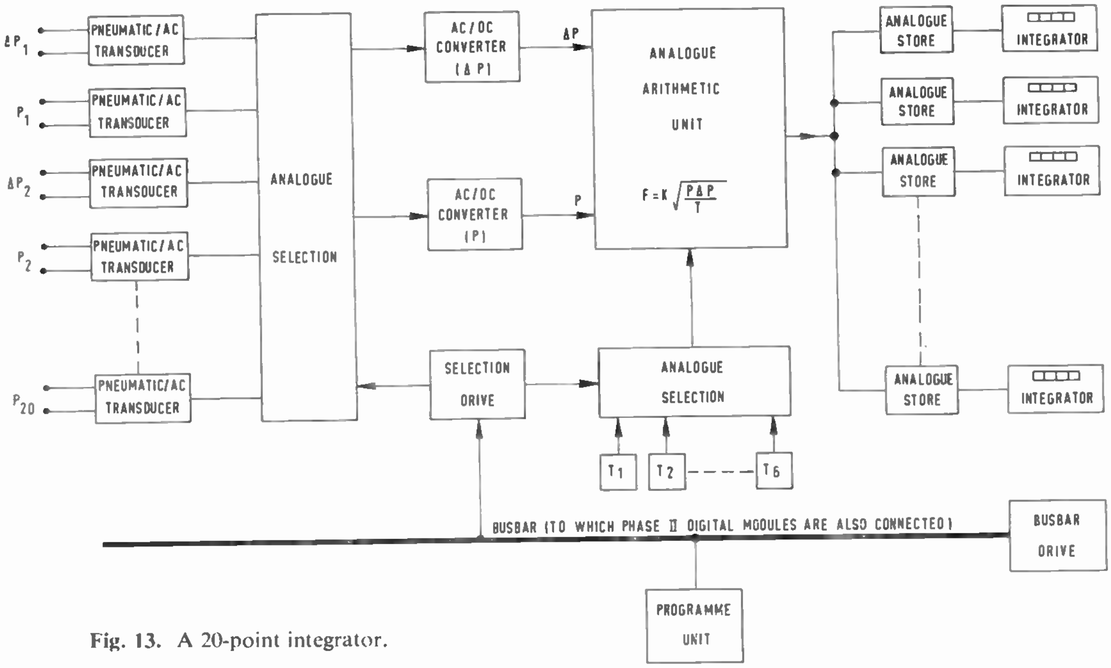


Fig. 13. A 20-point integrator.

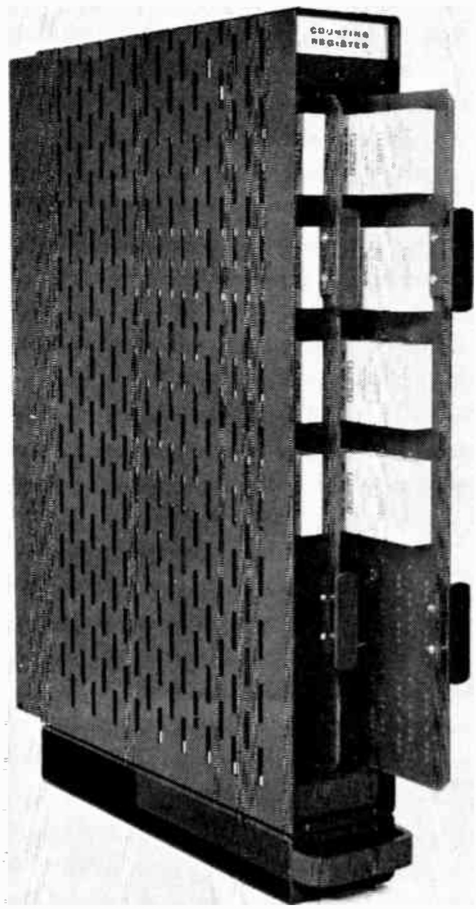


Fig. 15. A typical ARCH digital module showing sub-module boards.

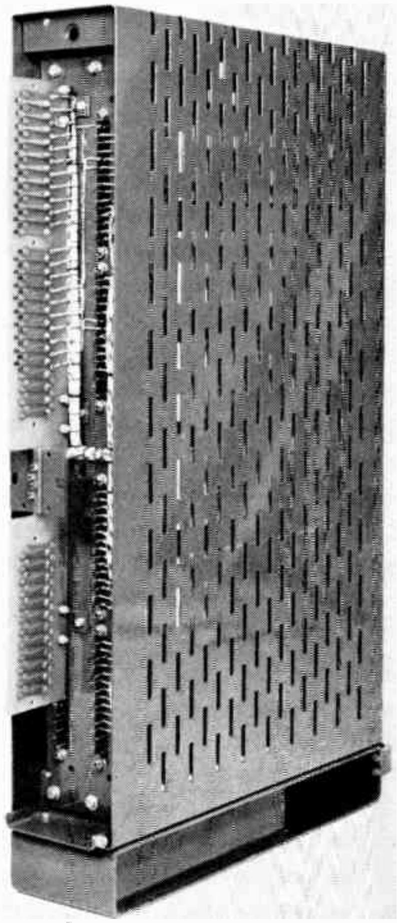
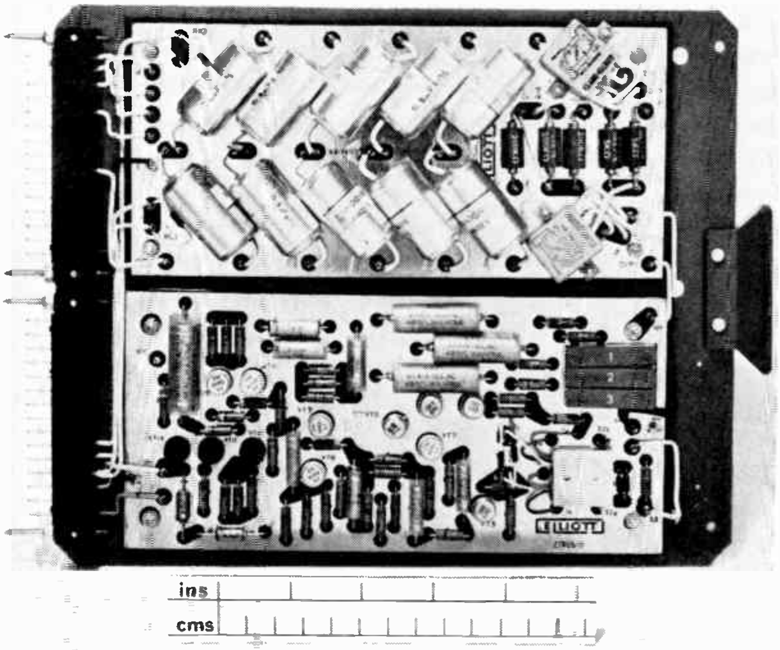


Fig. 16. Rear view of an ARCH digital module showing bus-bar connections.

Fig. 17. A typical ARCH analogue sub-module.



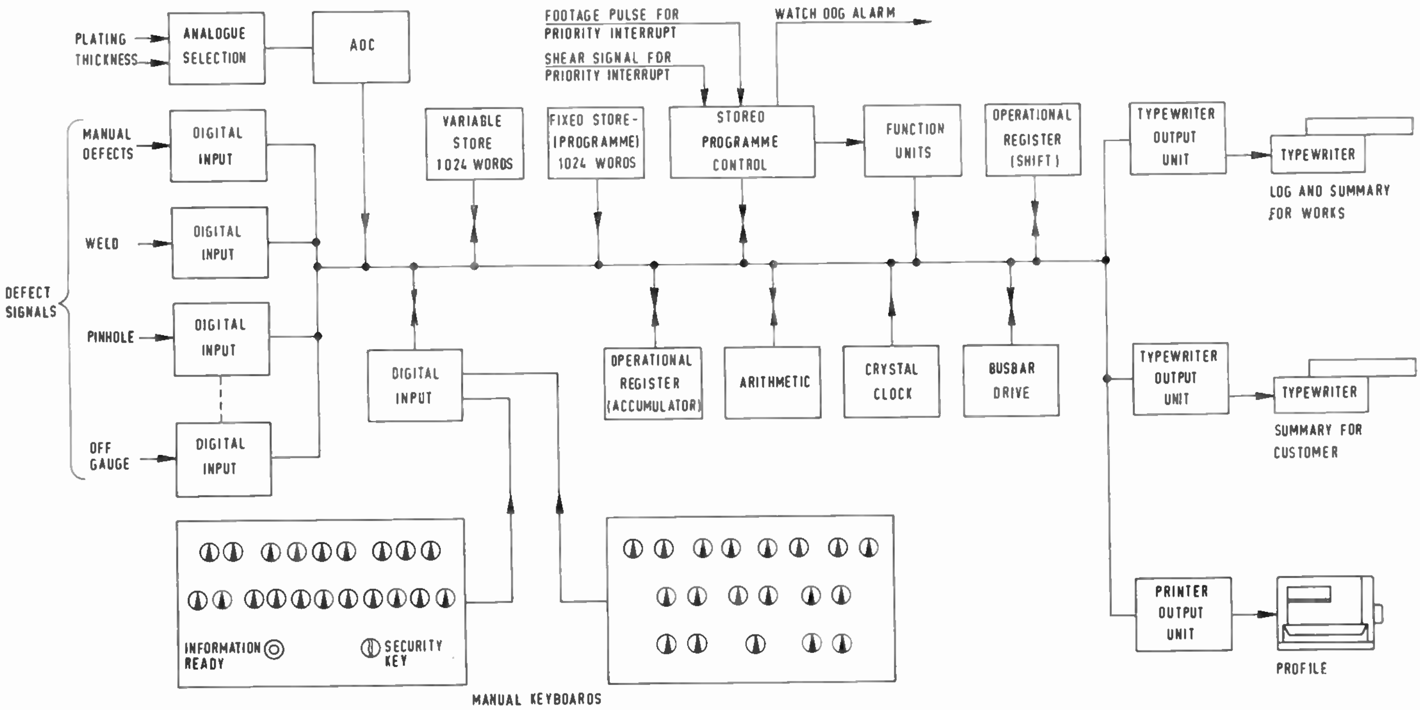


Fig. 14. ARCH data logger with arithmetic and memory.

The need for data storage and a certain amount of calculation, as well as quite a sophisticated program, means that a conventional data logger is not suitable; on the other hand a general purpose digital computer is hardly justified. An assembly of ARCH digital modules, as shown in Fig. 14 provides the correct amount of storage and arithmetic power required to solve the problem.

The program is stored in 1024 words (18 bits) of fixed store, and 1024 words of variable store provide working space for intermediate results, temporary data storage and space for a "defect queue". The presence of the various defects is signalled to the computer by the manually and automatically operated contacts shown. The plating thickness on both sides of the strip is continuously measured and fed as two analogue signals via the analogue selection module to the a.d.c. A basic arithmetic module for "add", "subtract" and "collate" is provided together with a register for "shift". The function units provide a simple instruction code having some 10 functions. The profile of defects (showing the exact location and duration for each type of defect in a coil), the summary and log for the manufactures and the defect summary for the customer are produced on the three printers.

A priority interrupt signal is obtained from a pulse produced at intervals of one foot of tin plate. This causes the program to scan all inputs and note any changes from the previous foot, e.g. the beginning or end of a given defect etc. Items requiring print out on the profile printer are placed in a queue in the store and printed out as fast as the printer can go (10 characters per second). A priority "interrupt" signal from the shear causes the program to print out both summaries of total defects and average plating thickness.

10. Conclusion

The ability to start the automation of a plant in a limited sphere, and then expand a hierarchy of control is now possible by the plant specialist without the need for logical and circuit designers. This means that the plant manager can gain both experience and confidence in advanced electronic computer control, without the large capital outlay which would be involved in automating the whole plant. This is especially important since in most cases the information on which to base the control is at best imperfectly known.

Finally, the reliability of the hierarchy system will be better due to the limited effect of breakdown of a single module and the easier fault diagnosis in a simple system.

Manuscript received by the Institution on 6th February 1963. (Paper No. 829/C58).

(*For Discussion on this paper turn to next page*)

DISCUSSION

(Under the chairmanship of Mr. W. Renwick)

Mr. D. W. Thomasson (*Associate Member*): In many of the control and/or surveillance applications for which ARCH is likely to be used, a free availability of numeric data in decimal form would be useful, apart from the normal read-out provisions. Both these output processes would be simpler if a decimal (or b.c.d.) form of numeric representation were used throughout, at the cost of some 20 per cent increase in storage cost, and minor problems in addition, etc. Was this possibility considered to determine whether the usual preference for pure binary working definitely applies in this equipment?

The Authors (*in reply*): The question of binary or binary coded decimal form of numeric representation is an old problem, and we would remind Mr. Thomasson that a 20 per cent increase is not only in storage cost and addition, but also such matters as conversion. We do find it desirable in some circumstances to use binary code decimal, but in general we feel that the straight binary given is more flexible.

Mr. J. M. Denvir: I wish to question the relative cabling costs between a hierarchical system and a centralized system.

The Authors (*in reply*): There is no fundamental difference in the cabling costs between hierarchical system and centralized system, although such costs will vary from application to application. In many cases it is likely that all machines of the network will be situated in a central control room. Where this is not the case, the system will tend to this form as far as cabling is concerned since better use can be made of central channels by multiplexing.

Mr. W. T. Lee: I would suggest that the reliability of a distributed system is less than that of a unified computer system because there are more components.

Secondly, do the authors agree that the best way to evolve a control of processes is to have a central digital computer program and not extra staff?

Thirdly, by doing a bit at a time, one is constrained to repeat the previous un-automated technique. Cannot so much more be done when considering the whole scheme?

The Authors (*in reply*): In reply to Mr. Lee's first point, it should be said that although one may indeed increase the probability of individual faults with a distributed system, one puts up the availability and operability of the system overall. One also decreases the mean time to repair faults by having simpler modules.

In his second point he inevitably implies a far greater capital expenditure than is initially called for, and although we naturally agree that program development is the way to evolve a control of processes, this must begin on the small scale.

Finally, we believe that Mr. Lee must be referring to new plants, since in an existing plant one does not repeat un-automated techniques as much as continue for the time being. In a new plant if one knows how to control the processes, we would naturally have no objection to providing control equipment.

Mr. J. A. Sargrove (*Member*): In British management thinking the bit-by-bit approach to investment in on-line computer control is bound to succeed more easily than the concept of Mr. Lee of a central giant computer. As purists however we might agree with Mr. Lee that if given the whole task at once his method might conceivably cost less.

U.D.C. 621.396.677.3

Synthesis of Multi-Element Directional Patterns using a Two-Element Single-Frequency Receiving Array

By

E. D. R. SHEARMAN,

B.Sc.(Eng.)†

Presented at the Symposium on "Sonar Systems" in Birmingham on 9th–11th July 1962.

Summary: The principle of space-frequency equivalence has been shown elsewhere to permit synthesis of multi-element directional patterns using only two physical elements by employing a multi-frequency carrier in the transmission medium. Disadvantages of the system are the use of more spectrum space and the need for a special transmission.

The present paper shows that even if a single-frequency carrier only is used, multi-element receiving patterns can be synthesized from the outputs of two physical elements. By the use of frequency-multipliers at the element outputs, multi-frequency carriers are produced synthetically at the receiver. This makes possible the production by electronic methods of a wide variety of directional patterns at the receiver without the co-operation of the distant transmitter. Penalties are paid in signal/noise ratio and multi-source resolution relative to a physical multi-element array.

1. Introduction

This paper presents what appears to be a novel technique for obtaining with two radio or sonar receiving elements alone, the directional response normally associated with a linear array of many elements.

The principle of the technique can best be appreciated by considering the way in which the directional response is built up in a conventional array. Such an array is shown in Fig. 1. The outputs from an even number of elements spaced d apart along a straight line are added together so that the contributions from all elements are in phase for a plane wave incident normally on the array.

If a plane wave of frequency $\omega/2\pi$ is incident at an angle θ to the normal, there is a progressive phase-shift along the array. The phase-shift relative to the array centre at the elements ± 1, ± 2, ± 3 etc. can easily be shown to be $\pm\phi/2$, $\pm 3\phi/2$, $\pm 5\phi/2$ etc., where $\phi = (\omega/c)\, d \sin\theta$, the electrical or acoustic phase angle between two adjacent elements. For a normally incident wave $\theta = 0$, so that $\phi = 0$ and all the element outputs add in phase. As the direction of the incident wave is changed, so the progressive outphasing of the various element outputs reduces their vector sum and produces the familiar directional response.

In the proposed technique, only the centre two elements of the array are needed, giving outputs phase-shifted $+\phi/2$ and $-\phi/2$ relative to the array centre. The output of the right-hand element is split and applied to a number of phase multipliers which give outputs of frequency ω and phase-shift $+3\phi/2$, $+5\phi/2$, $+7\phi/2$ etc. Similar phase multipliers are connected to the left-hand element giving outputs phase-shifted $-3\phi/2$, $-5\phi/2$, $-7\phi/2$ etc. The vector sum of all these outputs (together with the original element outputs) is thus of the same form as that from the original complete multi-element array, and a similar directional response will be obtained.

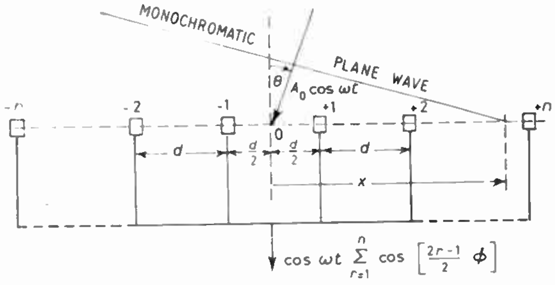


Fig. 1. Spatially-distributed multi-element additive array.

The essential feature of the proposed system is clearly the phase-multiplier. This can conveniently take the form of a harmonic generator giving an output $n\omega$ with phase $n\phi$, followed by a modulator to translate the frequency back to ω while retaining the phase $n\phi$. The switching signal for the modulator can have frequency $(n-1)\omega$ or $(n+1)\omega$ but must be in phase with the incident plane wave at the array centre. (If no suitable signal is available it can be derived from an additional receiving element at the array centre followed by a harmonic generator of order $(n-1)$ or $(n+1)$.

The properties of a two-element phase-multiplication array of this 'linear-additive' kind are discussed in the paper, together with suggested practical arrange-

[†] Electrical Engineering Department, University of Birmingham.

ments. However the necessity for a reference carrier having the phase of the wave at the array centre is an undesirable complication. It will be shown later that an alternative arrangement employing multiplication avoids this necessity.

It thus appears possible to synthesize electronically and, if desired, to scan a wide range of directional response patterns using the outputs of two physical receiving elements alone. It will be shown later that with orthogonal pairs, two-dimensional beams may also be synthesized and scanned.

The purpose of this paper is to study the feasibility and limitations of receivers using these techniques. In studying the limitations, the effects of noise and multiple sources are important since the above arguments have assumed a single source in the absence of noise or interference

Although the system using harmonic generation at the receiver appears to be new, it is very closely related to the directional systems using multi-frequency carriers in the transmission medium proposed by Kock and Stone,[1] Welsby[2] and Tucker.[3] The present proposals were in fact evolved following a study of these systems. In this paper the theory is developed by first considering the directive properties of the normal multi-element array, then showing how the same directivity can be obtained with two receiving elements by transmitting a multi-frequency carrier in the medium, and finally how with a single frequency in the medium, a multi-frequency carrier can be generated in the receiver by the harmonic generation techniques already mentioned. A study of the properties of various practical harmonic generators and frequency multipliers then leads to a comparison of the signal/noise performance of the phase multiplication array with that of the multi-element and multi-frequency types.

2. Directivity of a Multi-Element Additive Array

Consider a linear array having an even number, $2n$, of elements spaced d apart as in Fig. 1. We use x to measure distance along the array from its centre O which will lie half-way between two elements.

If a monochromatic plane wave incident at an angle θ to the normal to the array gives rise to a field $A = A_0 \cos \omega t$ at O, the field at any point x along the array is

$$A(x) = A_0 \cos\left[\omega t + \frac{\omega}{c} x \sin\theta\right] \qquad \text{......(1)}$$

(The phase constant $2\pi/\lambda$ has been put in the form ω/c for reasons which will become apparent later.)

At the location of the individual elements, x will take the values $\frac{1}{2}d$, $\frac{3}{2}d \ldots (2n-1)d/2$, for elements $+1$ to $+n$, and $-\frac{1}{2}d$, $-\frac{3}{2}d \ldots -(2n-1)d/2$ for elements -1 to $-n$ respectively. The field at a typical element r, distant $+(2r-1)d/2$ from O is thus

$$A_r = A_0 \cos\left[\omega t + \frac{\omega}{c} \,.\, (2r-1)\frac{d}{2}\sin\theta\right] \qquad \text{......(2)}$$

For convenience put ϕ for the electrical phase angle between two points d apart,

$$\phi = \frac{\omega}{c} d \sin\theta$$

Thus (2) becomes

$$A_r = A_0 \cos\left[\omega t + \frac{(2r-1)}{2}\phi\right]$$

We suppose that the field A_0 gives rise to an e.m.f. E_0 in an element, this e.m.f. appearing in series with the radiation resistance R of the element.

If the outputs of all the elements are connected together, then, assuming that there is no mutual coupling between the elements, the source impedance of the combined array will be $R/2n$. If a matched load $R/2n$ is connected to the array, the voltage appearing across it is

$$V = \tfrac{1}{2}E_0 \left\{ \sum_{r=1}^{n} \cos\left[\omega t + \frac{(2r-1)}{2}\phi\right] + \right.$$
$$\left. + \sum_{r=1}^{n} \cos\left[\omega t - \frac{(2r-1)}{2}\phi\right] \right\} \qquad \text{......(3)}$$

The first and second summations referring to the right-hand and left-hand halves of the array respectively.

Expression (3) can clearly be written

$$V = E_0 \cos\omega t \sum_{r=1}^{n} \cos\left[\frac{(2r-1)}{2}\phi\right] \qquad \text{......(4)}$$

All the terms in this series are co-phasal sinusoids of frequency ω and amplitude proportional to the cosine of the electrical phase difference between the pair of elements $+r$ and $-r$.

Carrying out the summation gives the familiar expression for the directivity of an array of $2n$ elements,

$$V = E_0 \cos\omega t \,.\, \frac{\sin[2n\phi/2]}{\sin[\phi/2]} \qquad \text{......(5)}$$

Although this is a standard result, it has been derived here in a slightly unusual way to stress the role of the symmetrical pairs of elements in building up the pattern.

3. A Multi-frequency Analogue of a Spatial Additive Array

To introduce the idea of space-frequency equivalence, consider the output voltage due to the pair of elements $+r$ and $-r$ when irradiated with the monochromatic plane wave incident at θ. The field at the element $+r$ is given by eqn. (2), and this may be written in a different way as

$$A_r = A_0 \cos\left[\omega t + \frac{(2r-1)\omega}{c} \cdot \frac{d}{2} \sin\theta\right] \quad \ldots\ldots(2b)$$

In this form the phase-shift appears as that produced by propagation at a frequency $(2r-1)\omega$ with the element spaced $d/2$ away from O. This suggests the possibility of using two elements only for reception and achieving the directivity of an array by using a multi-frequency signal in the medium. However the phase shifts under such conditions would be obtained at different frequencies $(2r-1)\omega$ and not with ω as in eqn. (2b).

One method of achieving a phase shift corresponding to a transmission frequency of $(2r-1)\omega$ at an output frequency of ω is to use $(2r-1)\omega$ in the medium and include a modulator in the receiver to translate the frequency, with phase-shift unaltered, to ω. We can repeat this process for each pair of elements by using a multi-frequency carrier in the medium consisting of frequencies ω, $3\omega \ldots (2r-1)\omega \ldots 2(n-1)\omega$, translating the element outputs at each frequency to ω in a separate modulator and then adding all the common-frequency modulator outputs.

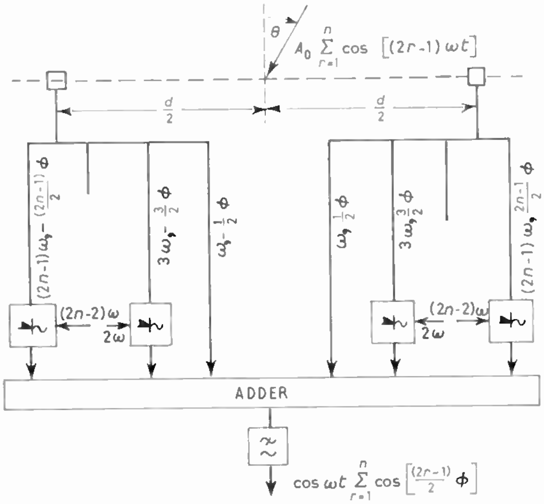

θ
$A_0 \sum_{r=1}^{n} \cos\,[(2r-1)\,\omega t]$
$\frac{d}{2}$
$\frac{d}{2}$
$(2n-1)\omega, -\frac{(2n-1)}{2}\phi$
$3\omega, -\frac{3}{2}\phi$
$\omega, -\frac{1}{2}\phi$
$\omega, \frac{1}{2}\phi$
$3\omega, \frac{3}{2}\phi$
$(2n-1)\,\omega, \frac{2n-1}{2}\phi$
$(2n-2)\omega$
2ω
$(2n-2)\omega$
2ω
ADDER
$\cos \omega t \sum_{r=1}^{n} \cos\left[\frac{(2r-1)}{2}\phi\right]$

Fig. 2. Two-element multi-frequency additive array (after Tucker). (Multi-frequency analogue of Fig. 1.)

This technique has been proposed by Tucker[6] and, the arrangement is illustrated in Fig. 2. A complication is the necessity at the receiver for coherent local signals of frequency 2ω, $4\omega \ldots (2n-2)\omega$ to switch the modulators. The switching signals might be derived from the multi-frequency carrier as received at a special element at O, suitable filtering and limiting being included.

The system proposed by Kock and Stone and by Welsby, which preceded Tucker's proposal, provides a simpler solution. Its performance is not, however, as closely analogous to the original multi-element spatially distributed array as the above

4. A Multi-frequency Multiplicative Array

Welsby's two-element multiplicative array is illustrated in Fig. 3. The multi-frequency plane wave is received as before by two receiving elements, but the outputs from these are applied to a multiplier.

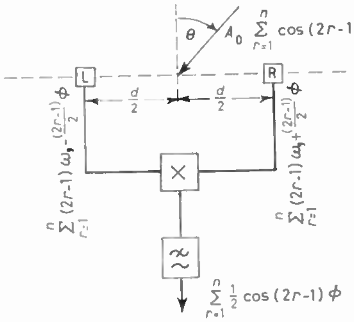

θ
$A_0 \sum_{r=1}^{n} \cos(2r-1)\,\omega t$
L
R
$\frac{d}{2}$
$\frac{d}{2}$
$\sum_{r=1}^{n} (2r-1)\,\omega, -\frac{(2r-1)}{2}\phi$
$\sum_{r=1}^{n} (2r-1)\,\omega, +\frac{(2r-1)}{2}\phi$
×
$\sum_{r=1}^{n} \frac{1}{2} \cos(2r-1)\,\phi$

Fig. 3. Two-element multi-frequency multiplicative array. (After Welsby).

We assume that the multiplier is a high impedance device and that the elements are separately terminated in resistive loads equal to their radiation resistance before multiplication. It is also assumed that this radiation resistance is the same for each frequency. The voltages applied to the multiplier from the right-hand and left-hand elements are then V_{RH} and V_{LH}, where

$$V_{RH} = \tfrac{1}{2}E_0 \sum_{r=1}^{n} \cos\left[(2r-1)\omega t + \frac{(2r-1)\omega d}{c} \frac{}{2}\sin\theta\right] \quad \ldots\ldots(6a)$$

$$V_{LH} = \tfrac{1}{2}E_0 \sum_{r=1}^{n} \cos\left[(2r-1)\omega t - (2r-1)\omega \frac{d}{2}\sin\theta\right] \quad \ldots\ldots(6b)$$

When these are multiplied, only those components of the two voltages which have the same frequency yield a d.c. output. These d.c. components are selected by the low-pass filter. The sum of the components is given by

$$V_{d.c} = \tfrac{1}{4}E_0^2 \sum_{r=1}^{n} \tfrac{1}{2}\cos\left[2 \cdot \frac{(2r-1)\omega}{c} \cdot \frac{d}{2}\sin\theta\right]$$

or in terms of ϕ,

$$V_{d.c} = \tfrac{1}{4}E_0^2 \sum_{r=1}^{n} \tfrac{1}{2}\cos\left[2 \cdot \frac{(2r-1)}{2}\phi\right] \quad \ldots\ldots(7)$$

Carrying out the summation we obtain,

$$V_{d.c} = \tfrac{1}{8}E_0^2 \cdot \frac{\sin 2n\phi}{\sin\phi} \quad \ldots\ldots(8)$$

This is seen to be rather similar to expression (5) for the spatially distributed array, but with three important differences.

(1) The amplitude is proportional to the square of the incident field.

(2) The argument of the sine terms in (8) is twice that of the corresponding term in expression (5), which has the effect of halving the lobe width as a function of ϕ.

(3) The multiplier output is bipolar d.c. and reverses sign between successive lobes, whereas the additive array output is at r.f. and merely reverses phase. Envelope detection of the additive array output will give a unipolar d.c. which will be proportional to the modulus of the r.f. voltage and will not change sign between lobes.

These differences are introduced by the process of multiplication, not by the use of multiple frequencies. They are found if a two-element single-frequency multiplicative array is used.[4]

5. A Spatial Analogue of the Multi-frequency Multiplicative Array

It is interesting to notice that although the spatially-distributed additive array resembles the multi-frequency multiplicative array only very superficially, a spatially-distributed array can be devised which is more closely analogous.

In this system, shown in Fig. 4, the outputs of the innermost pair of elements are taken to one multiplier, those from the next pair out to another multiplier and so on. The sum of all the d.c. outputs of the multipliers is then the same as that of the multi-frequency array of Fig. 3, and is given by expression (8). This arrangement has much in common with Ryle's "aperture synthesis" system, and has some useful possibilities which are discussed elsewhere.[9]

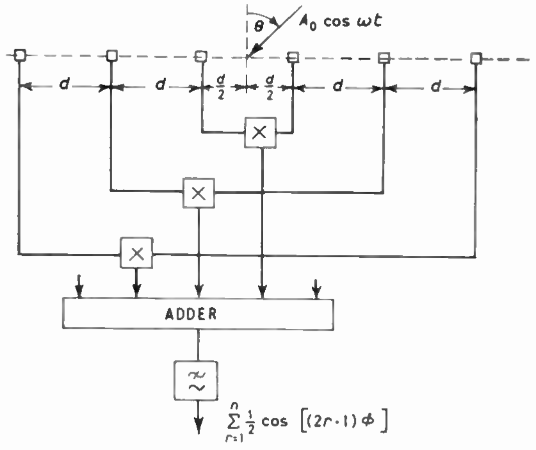


Fig. 4. Spatially-distributed multi-element array with multiplicative pairs. (Single-frequency analogue of Fig. 3.)

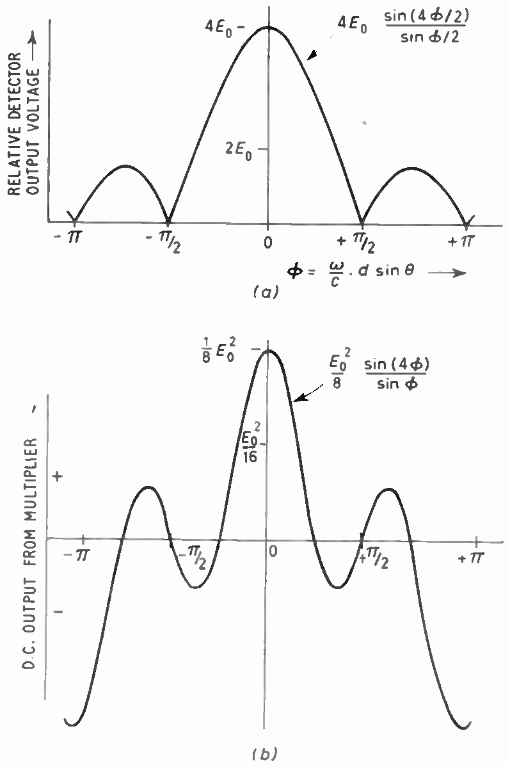


Fig. 5. (*a*) Amplitude of detected signal from 4-element additive array (Fig. 1). (The same pattern will apply to Fig. 2 for 2 frequencies of transmission.)

(*b*) Amplitude of d.c. output of multiplier for 4-element array of Fig. 4. (The same pattern will apply to Fig. 3 with two frequencies of transmission.)

To illustrate the difference between the directional patterns for additive and multiplicative systems with the same number of elements, Fig. 5 shows these patterns for a four-element array plotted against ϕ. Fig. 5(*a*) applies to the additive system of Fig. 1 and Fig. 2, while Fig. 5(*b*) applies to the multiplicative systems of Fig. 3 and Fig. 4.

The squared amplitude, halved angular scale and bipolar characteristic of the multiplicative systems can be clearly seen.

6. Synthesis of Harmonics of a Single-frequency Carrier at the Receiver

In the two types of multi-frequency array described in Sections 3 and 4, a multi-frequency carrier, ω, $3\omega \ldots (2n-1)\omega$, was transmitted in the medium and gave rise at the load resistance of the right-hand element of the multifrequency voltage given by eqn. (6a). This expression may be rewritten as

$$V_{RH} = \tfrac{1}{2}E_0 \sum_{r=1}^{n} \cos\left[(2r-1)\left\{\omega t + \frac{\omega}{c}\cdot\frac{d}{2}\sin\theta\right\}\right] \qquad \ldots\ldots(9)$$

This is the form of a fundamental frequency sinusoid,

$$\cos\left(\omega t + \frac{\omega}{c}\cdot\frac{d}{2}\sin\theta\right)$$

and a train of odd harmonics. The phase of each harmonic bears the same ratio to the phase of the fundamental as does its frequency. Apparently, the only extra directional information carried by a harmonic is its harmonic number.

If a single frequency of transmission, ω, is used in the medium and the resulting output of the right-hand element is applied to a frequency multiplier arranged to generate harmonics of order $2r-1$, the output will be of the form

$$\cos\left[(2r-1)\left\{\omega t + \frac{\omega}{c}\cdot\frac{d}{2}\sin\theta\right\}\right]$$

This is identical to the rth term of eqn. (9). For the present the factor giving the amplitude of the harmonic as a function of E_0 is omitted; this is dependent on the type of multiplier and will be considered later. We assume here that the output of the multiplier is proportional to the input, which will be shown to be true for certain types of multiplier.

The outputs of two sets of frequency multipliers, one for the right-hand element and one for the left, yield two correctly phased multi-frequency signals of the form of eqns. (6a) and (6b). They may be processed either by the additive scheme of Fig. 6, in which case local co-phasal carriers are required for frequency translation, or by the multiplicative scheme of Fig. 7.

Before studying these schemes in more detail, it is desirable to study the performance of the principal types of frequency multiplier which might be used in such systems.

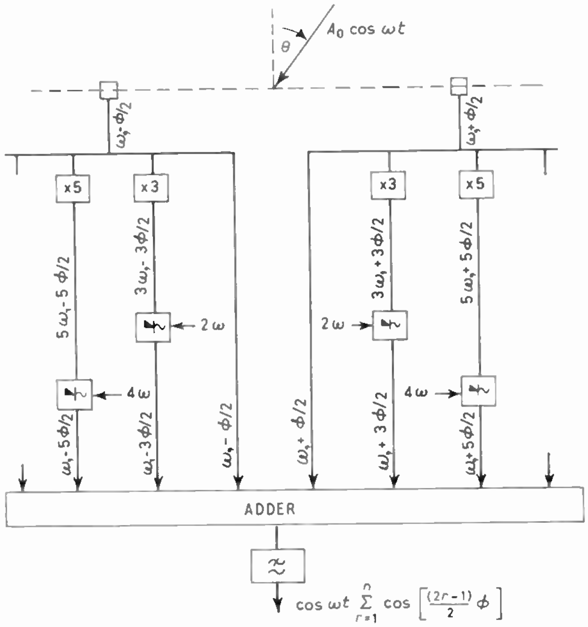


Fig. 6. Linear additive receiver with harmonic generation.

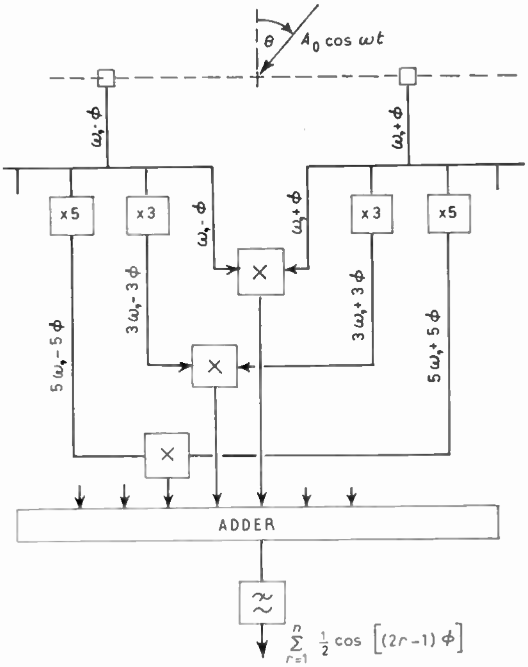


Fig. 7. Multiplicative receiver with harmonic generation.

7. The Characteristics of Harmonic Generators

There are two main classes of harmonic generators, the two having very different properties. The first makes use of recording techniques and is typified by a multi-speed tape recorder. If a signal is recorded at one speed, and then played back at two, three, etc. times this speed, the frequency of the signal will be multiplied by a corresponding factor. Such frequency multipliers are linear in operation and cause no degradation in signal/noise ratio.

The second type uses a non-linear device and an output selective filter. These devices produce harmonics of an input signal in real time, but also inevitably produce cross products when presented with a complex input signal or a signal plus noise.

These two classes of harmonic generator will be discussed separately.

7.1. *Recording Devices*

A multi-speed recording device produces frequency multiplication by time contraction. A signal of angular frequency ω which is recorded for an interval T may be played back at n times the recording speed to yield a frequency $n\omega$ for time T/n.

In the present application it is wished to generate a train of harmonics $3\omega, 5\omega \ldots (2n-1)\omega$ from the fundamental frequency ω. Clearly a single sample of the frequency ω of duration T will only yield the harmonic $(2n-1)\omega$ for a time $T/(2n-1)$. The indicating device of the receiver will therefore only be operative for $1/(2n-1)$ of the time of observation.

A suitable magnetic recording system might use a rigid rotating magnetic drum with a single recording head and a large number of replay heads distributed around the drum circumference. On replay the frequency ω would be obtained from one stationary head, while the frequencies 3ω, 5ω, etc. would be derived by scanning the head outputs round the drum in the opposite direction to the direction of rotation at twice, four times etc. the drum speed.

Analogous arrangements using acoustic delay lines are also possible.

Such schemes are seen to involve considerable complication. They also have one major disadvantage which has been concealed by restricting consideration to a constant-frequency sinusoidal signal.

If the signal is modulated, for example by pulses or on/off keying, the modulation speed will be increased by the frequency multiplying process. When the various components which have been speeded up to different extents are added, the resulting modulation will be a meaningless jumble of different speed versions of the same signal.

A similar argument will apply to a sinusoidal signal if its frequency or amplitude changes during the period of recording.

Clearly such a system can only be used for determining the direction of near-stationary c.w. transmitters.

7.2. *Non-Linear Devices*

The simplest frequency multipliers analytically are rth law devices, where r is the order of multiplication. For a sinusoidal input $E \cos \omega t$, these yield an output at the rth harmonic proportional to $E^r/2^{r-1}$ together with outputs of lower order harmonics which can be removed by filtering.

To make use of such devices in a multi-frequency receiver, some kind of circuit is needed which makes the outputs of the various multipliers remain constant or vary in the same manner as each other as the input voltage is varied. If this is not done, the relative amplitude of the different components changes and the directional pattern alters with field strength.

For this reason, and also to avoid distortion of any amplitude modulation present on the signal, it is necessary to make the multiplier output proportional to the input. A number of systems using a.g.c. or r.f. envelope feedback are available which can achieve this. Perhaps a more promising system is one in which a harmonic generator is used to give $(n-1)\omega$, and this is used as the switching signal in a modulator used to translate the original frequency ω to $(n-1)\omega + \omega = n\omega$. If the switching signal is arranged to be large compared with the input at ω, the output will vary linearly with the input as required.

The other important property of the non-linear device in the present application is that it produces cross-modulation and spurious frequencies when the input consists of more than one frequency. There are two situations in which this is important; when signals from more than one transmission source are arriving at the aerial, and when noise (either external or internal) is present with the signal. The multi-source situation needs detailed analysis for particular arrays, but some useful information about the response of frequency multipliers to noisy signals is already available.

Blachman[6] gives the signal/noise power ratio R for the mth harmonic output from an nth power law device, ($n = 0$ for an ideal limiter), as

$$R' = \frac{2R}{m^2 + n^2} \qquad \ldots\ldots(10)$$

where the input signal/noise ratio is R and is large. The ideal limiter is seen to cause the least degradation in signal/noise, giving $R' = 2R/m^2$. In the multi-frequency receiver, the effect of this signal/noise degradation will thus be progressively more serious for higher orders of multiplication. Consequently the "synthetic elements" produced by frequency multiplication in the receiver will have a progressively worse signal/noise ratio the further out they are from the two physical elements.

8. Examples of Pattern Synthesis using Harmonic Generation at the Receiver

In Section 6 it was shown that the principle of directional pattern synthesis using harmonic generation at the receiver should be applicable equally to the linear additive system of Tucker and to the multiplicative system.

With a single frequency transmission and a two-element array, the whole range of directional pattern synthesis techniques using addition and multiplication so extensively discussed in the literature is available.

If two further elements are introduced in a line at right angles to the first two, the "synthetic elements" may be arranged to cover an area instead of a line, and fully steerable beams can be produced. With multiplicative systems the pattern synthesis techniques discussed by Tucker in another paper[8] may be employed with a resulting increased flexibility.

In communication applications the linear additive system is attractive as it introduces no modulation distortion, whereas the square-law characteristic of the multiplicative system renders amplitude modulation by a complex signal (frequency division multiplex or telephony) unusable. It is also free from undesirable effects in the presence of multiple sources.

If such a linear additive system were to be adopted in a receiver employing harmonic generation, a linear harmonic generator would be essential. The considerations of Section 7.1 show that the use of recording devices for this purpose inevitably destroys the modulation. The scheme outlined in Section 7.2 for a linearized frequency multiplier seems to offer the only practical solution.

Attention is here restricted to multiplicative systems which necessarily employ non-linear devices. Although complex amplitude modulation is destroyed in such devices, pulse or on/off modulation is unaffected, and they are thus suitable for pulse echo-ranging or binary a.m. communication systems.

8.1. *Multiplicative System*

The block schematic of a multiplicative two-element receiver harmonic generator is shown in Fig. 7. The output of the harmonic generators is seen to be of exactly the same form as the element output of the multiplicative multi-frequency array of Fig. 3.

Provided that the harmonic generators are rendered linear as discussed in Section 7.2, the directional pattern will be independent of signal amplitude. By inserting phase shifters and attenuators, however, the pattern may be varied as wished, and in particular may be scanned. The system is convenient for this since only two phase shifters are needed, one in the output of each element, and these operate at a fixed frequency.

The output of the adder will be of the bipolar form of Fig. 5(*b*) and a rectifier may therefore be used to remove the negative lobes.[4]

8.1.1. Signal/noise ratio

The signal/noise power ratio for a simple multiplicative pair system with equal elements is

$$\frac{P_s}{P_n'} = \frac{1}{\sqrt{2}}\frac{P_s}{P_n}$$

where $P_s/P_n = R_2$ is the signal/noise ratio for a linear additive array with the same two elements,[4] (P_n, P_n' are the noise powers from the additive and multiplicative arrays respectively for the same signal power P_s). When harmonic generation is added using an ideal limiter, eqn. (10) shows that the signal/noise ratio of the *m*th harmonic is $2/m^2$ times that of the fundamental. For a constant signal amplitude (and therefore P_s) from each harmonic generator, (uniform array), the noise power contributed by the fundamental is $\sqrt{2}\,.\,P_n$ and that by the $(2r-1)$th harmonic is $\sqrt{2}\,.\,P_n\,.(2r-1)^2/2$. Thus the total noise power when all the harmonic generator outputs are added, is

$$\sqrt{2}P_n\left[1+\sum_{r=2}^{n}\tfrac{1}{2}(2r-1)^2\right] \qquad \ldots\ldots(11)$$

The signal voltages add in phase, so that the signal power is n^2P_s, when P_s is the signal power from the additive array. The signal/noise ratio of the multiplicative array with harmonic generation expressed in terms of that for the two-element additive array is therefore,

$$R_{2n} = R_2\,.\,\frac{n^2}{\sqrt{2}\left[1+\sum_{r=2}^{n}\tfrac{1}{2}(2r-1)^2\right]} \qquad \ldots\ldots(12)$$

The signal/noise ratio is seen to decrease as the number of elements is increased. For comparison the signal/noise ratio of a multi-frequency two-element array is independent of the number of elements, if account is taken of the reduction in transmitter power per carrier necessary when multiple carriers are radiated. The signal/noise ratio for a linear additive array of $2n$ elements, on the other hand, increases in proportion to n.

One assumption made in this analysis requires further investigation. It has been assumed that the noise powers from the various multipliers can be added, implying that they are in random phase. In fact they are derived by a complex process from the same waveform, and their statistical relationship thus require further study. The signal/noise ratio calculated above, already lower than that of the other arrays, might thus be further reduced.

8.2. *Multi-source Resolution*

The response of a multiplicative receiver to multiple sources has been discussed elsewhere.[4, 7] The response is shown to be dependent on the directivity of the two parts of the array whose output is multiplied. In the present system these two parts have been assumed to have no directivity. No resolution of multiple sources would therefore be expected. This is similar to the behaviour of another small-aperture directional system, the Adcock direction-finder, which gives high directional accuracy, but only gives the multi-source resolution of the individual elements.

9. Conclusions

It has been shown that a two-element array can provide the directional pattern of a multi-element array for a single frequency signal if harmonic generation is employed at the receiver. The technique makes it possible with two elements to synthesize the full range of steerable patterns possible by the use of additive and multiplicative array techniques without the necessity for modifying the transmitted spectrum.

By using two orthogonal pairs of elements, two dimensional beams can be produced.

The directional patterns are however only applicable for single sources. The multi-source resolution is dependent on the directivity of the two individual elements.

The signal/noise ratio of the system decreases as the number of elements increases if non-linear devices are used as harmonic generators.

10. References

1. W. E. Kock and J. L. Stone, "Space-frequency equivalence", *Proc. Inst. Radio Engrs*, **46**, p. 499, 1958.
2. V. G. Welsby, "Two-element aerial array", *Electronic Technology*, **38**, p. 160, 1961.
3. D. G. Tucker, "Space-frequency equivalence in directional arrays", *Proc. Instn Elect. Engrs*, **109C**, No. 15, p. 191, March 1962. (I.E.E. Monograph No. 479E, November 1961.)
4. V. G. Welsby and D. G. Tucker, "Multiplicative receiving arrays", *J. Brit.I.R.E.*, **19**, p. 369, 1959.
5. F. E. Terman, "Radio Engineers Handbook", p. 622, 1st edn. (McGraw-Hill, New York, 1943).
6. N. M. Blachman, "The output signal-to-noise ratio of power law devices", *J. Appl. Phys.*, **24**, p. 783, June 1953.
7. V. G. Welsby, "Multiplicative receiving arrays", *J. Brit.I.R.E.*, **22**, p. 5, 1961.
8. D. G. Tucker, "Multiplicative arrays in radio-astronomy and sonar systems", *J. Brit.I.R.E.*, **25**, No. 2, p. 113, 1963.
9. E. D. R. Shearman, "Non-collinear and cylindrical multiplicative arrays", *J. Brit.I.R.E.*, 1963. (To be published.)

Manuscript first received by the Institution on 22nd June 1962 and in final form on 9th April 1963. (Paper No. 830/SS24).

The Third British Satellite

The development contract for UK.3, the first all-British satellite, has been awarded to the British Aircraft Corporation's Guided Weapons Division. UK.3, which will be based on specifications issued by the Space Department of the Royal Aircraft Establishment, will be the third in a series of joint Anglo-American scientific research satellites. The first two in the series, UK.1 and UK.2, are American-built and fitted with British instruments. UK.1 (*Ariel*) is in orbit,† while UK.2 is due to be launched later this year.

The proposed date for launching UK.3 is 1966, when the satellite will carry five scientific experiments into a circular orbit, 400 miles above the earth. It will transmit experimental data for a year, after which its transmitter will be switched off so that its radio frequencies can be allocated to a new satellite.

The five experiments that UK.3 will be conducting have been chosen by the Royal Society, and will include the following:

University of Birmingham: Electron density near the satellite.

University of Cambridge: Mapping of galactic noise sources.

University of Sheffield: Study of radio signals below 20 kc/s.

Meteorological Office: Atmospheric distribution of molecular oxygen.

Radio Research Station: Intensity and distribution of natural terrestrial noise and anomalous radio propagation.

With an overall height of 5 ft and a span of 7 ft, the satellite will weigh 1¼ cwt. Its shape has been determined partly by the American *Scout* rocket, which will inject it into orbit, and partly by the requirements of experiments, power supplies, thermal control, aerials, structural strength and access. The 30-in diameter of the main body of the satellite is the largest which can be carried by the *Scout* rocket.

To control the experiments, and to store data for subsequent re-transmission, the satellite will carry several other electronic units, including timers, coding devices, battery chargers, a transmitter, a receiver to accept commands from earth, and a tape recorder. Electrical power for these units and the experiments will be provided by 6000 solar cells. These will also re-charge the storage batteries which take over when the satellite passes into the earth's shadow.

† "Space research experiments in the world's first international satellite", *J. Brit.I.R.E.*, **23**, No. 5, p. 398, May 1962.

U.D.C. 621.395.92

Modern Aids to Hearing

By

M. C. MARTIN†

Presented at the Symposium on "Practical Electronic Aids for the Handicapped" in London on 28th March 1962

Summary: A brief history of hearing aids is presented together with a general survey of current types. Their technical performances and the problems associated with their manufacture and operation are discussed.

1. Introduction

According to Goldstein[7] the first electric amplifying device was produced in 1900 by Ferdinand Alt in Austria, being "a loudspeaker microphone with telephone and dry cell battery". This aid was the forerunner of the carbon microphone aids which became widely used by the 1920's, and used an external earpiece (Fig. 1).

By 1920 valve aids were also being produced. One of this type was the Marconi Otophone having a two stage amplifier with a desk standing microphone and an external earphone; this aid had to be built into a small suitcase because of its bulk and weight. The low tension batteries used in these aids were of the lead acid type and were consequently very bulky.

As a result of improved design, which made possible the production of small valves, hearing aids became smaller and by 1933 were reduced to the size of a small handbag (Fig. 2). They were, however, still relatively heavy and cumbersome. The first self-contained body-worn aid was not produced until 1943 when improved low current consumption valves enabled smaller and lighter types of batteries to be used. Modern aids are very much smaller because the introduction of the transistor eliminated the need for the bulky high tension battery; they have been reduced in size to such a degree that they can easily be worn behind the ear.

A profusely illustrated and detailed account of the historical development of hearing aids up to 1930 has been prepared by Goldstein,[7] and Watson and Tolan[13] have carried the history up to 1943. In 1951 the British Institution of Radio Engineers held a symposium in which the medical and engineering aspects were discussed.[1, 2, 11]

2. Categories of Aids Available

The number of different models of electrical hearing aids available today is very large but three basic divisions can be made, namely:

(*a*) body-worn aids,

(*b*) head-worn aids,

(*c*) special aids used for educational purposes such as speech trainers and group aids.

2.1. *Body-worn Aids*

Nowadays this type of aid is often regarded as being rather old fashioned by many people but is in fact, more efficient as a speech reproducer and is capable of giving much greater amplification than a head-worn aid. Body-worn aids are, therefore, still the only portable means of amplification for the severely deaf person.

Body-worn aids are composed of four main components:

(*a*) a microphone,

(*b*) an amplifier,

(*c*) a source of power,

(*d*) a separate earpiece or receiver.

The first three components are usually housed in one container.

2.2. *Head-worn Aids*

Head-worn aids can be sub-divided into those fitted

(*a*) behind the ear,

(*b*) in spectacles,

(*c*) in the ear,

(*d*) as a hair slide, or on a head band.

The most popular type among head-worn hearing aids at present is the one worn behind the ear.

2.3. *Educational Aids*

School aids are used for normal teaching purposes and for teaching speech and vocabulary. They can be used either for individual or group use.

3. Performance of Hearing Aids

While considerable progress has been achieved in producing more compact and readily wearable hearing aids, their performance has necessarily been affected with this reduction in size. Until the publication of the Harvard University Report[4] on the design of hearing aids in 1945, little scientific information was available to enable manufacturers to select a

† Royal National Institute for the Deaf, London, W.C.1.

suitable frequency response for a hearing aid. One of the objects of the Harvard report was to determine, if possible, a response curve for a general purpose aid by using a master hearing aid[2] or speech transmission system whose characteristics could be freely adjusted. At the same time work was also undertaken independently in Great Britain by the Medical Research Council on similar lines and its results, published in 1947,[8] were found to be in close agreement with those of Harvard. The frequency response selected by the M.R.C. as giving the most suitable response for a general purpose hearing aid is shown in Fig. 3† and still is the basis of most hearing aid designs.

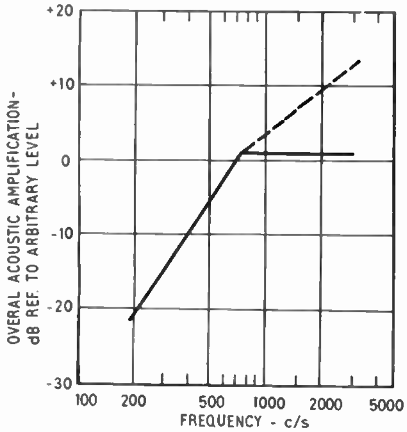


Fig. 3. Overall acoustic frequency response desirable in a general purpose hearing aid. (From M.R.C. Report No. 261, page 9).

The aids produced in 1945 often had large peaks in the frequency response which limited the degree of amplification because of the problems of acoustic feedback, and the intelligibility of speech received. These peaks were mainly caused by the crystal earpieces in use at that time.[10] The subsequent introduction of magnetic earpieces has produced smoother frequency response characteristics in modern hearing aids, and has also enabled higher maximum acoustic outputs to be obtained. The frequency range of amplification has also been increased particularly for the lower frequencies so that modern aids approach the M.R.C.'s recommended characteristics.

The problems of increasing the low frequency response of hearing aids, however, reappeared to a certain extent with the introduction, in 1955, of head worn aids which used sub-miniature earpieces and microphones. These transducers were insensitive at low audio frequencies.

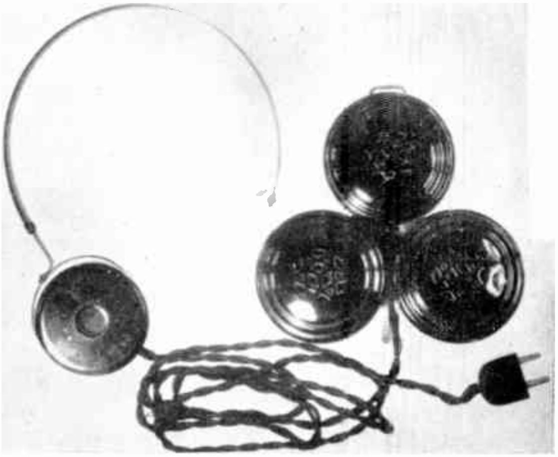


Fig. 1. Carbon granule microphone, battery, and external earpiece as used in the 1920's. Each microphone is about 2½ in diameter; the battery (*not shown*) is approximately 3¾ in high by 2¾ in wide by ¾ in thick. The connector is a standard 5A 2-pin plug.

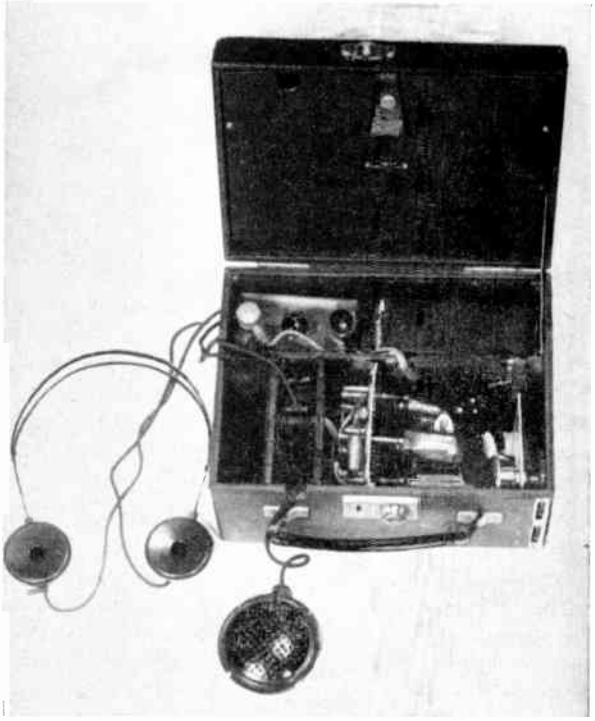
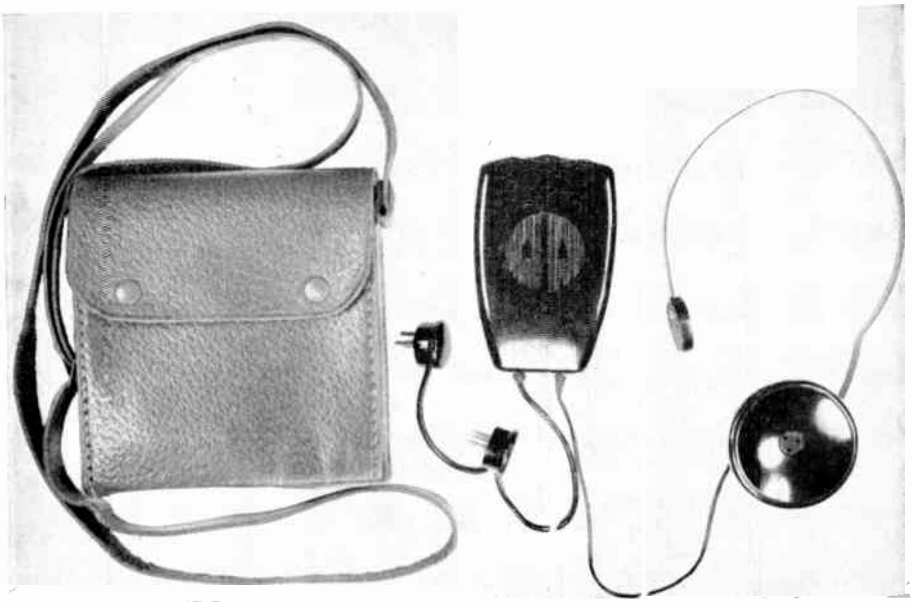


Fig. 2. Valve hearing aid made about 1934. It had a carbon granule microphone and a pair of headphones but between them came a three-valve amplifier. Low tension for the valves was provided by a 2 V accumulator and high tension by a 60 V dry battery. Milled hand-wheels on the right controlled volume and tone. For carrying, the headphones were packed in the lid and the microphone slipped into the pocket on the left.

3.1. *Medresco Hearing Aids*

The characteristics of a hearing aid for general use as recommended by both the Medical Research Council

† All acoustic measurements in this paper are free field measurements unless otherwise stated. Acoustic outputs are measured in a 2 cm³ acoustic coupler.



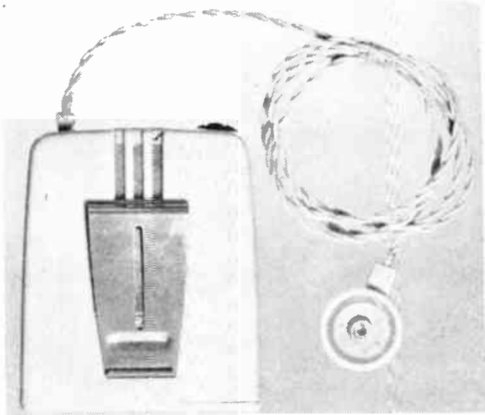


Fig. 4. (*Top*) first Medresco hearing aid and (*below*) a modern Medresco transistor aid with insert receiver.

and Harvard Reports have been closely followed in practice by the National Health Medresco aids[3, 12] (Fig. 4). While the Medresco aid suits many people (860 000 aids have been issued since 1948), others cannot or will not use this aid. Three main reasons for not using a Medresco aid are: (*a*) the tone of the aid does not suit the user, (*b*) the aid is not powerful enough for more profound forms of deafness, and (*c*) the Medresco aids are only of the body-worn type and, therefore, have a cord and button receiver which many individuals find socially unacceptable.

The introduction of the National Health transistor aid to adults (May 1960) provoked many complaints from users who had changed over from valve to transistor models of the Medresco type; the complaint was that due to a great difference in tonal quality, the transistor model was claimed to be unacceptable to the user. The difference in performance between the Medresco transistor and valve models is shown in Fig. 5. Many people considered that this apparent difference was due to the use of a magnetic microphone, in place of the crystal microphone that was used in the valve model. In actual fact, the responses of both types are very close to the Medical Research Council specified response, except for a significant difference in the 300 to 1000 c/s range.

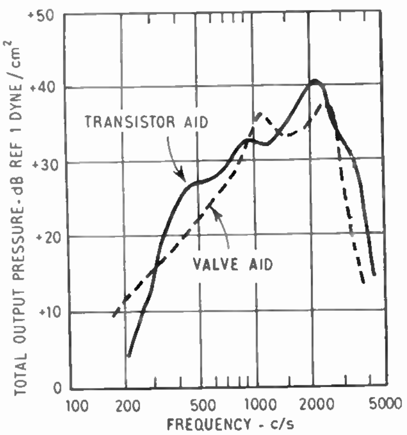


Fig. 5. Medresco hearing aid frequency response at maximum gain. Acoustic input — 20 dB ref. 1 dyne/cm². The output is measured in a 2 cm³ acoustic coupler.

Considerable discussions have taken place to find out whether crystal microphones or magnetic microphones give the better response.[14] In a recent experiment the transistor model. in which the response was modified so as to be identical to that of the valve aid, was tried by former users of the valve aid and the results revealed that these users could not find any difference in performance between the two types of aid. This experiment would, therefore, appear to discount the view that the use of a crystal microphone, rather than a magnetic one, would give a more satisfactory overall performance in a hearing aid. Observations on both the valve and transistor Medresco hearing aids show that a small difference in frequency response will give rise to a large subjective difference.

Although one can decide on the approximate magnitude of amplification and maximum acoustic output which is thought desirable for any one individual, the apparent effect on the patient of small changes in the frequency response of the ear makes it difficult to prescribe a hearing aid accurately. Although the majority of aids have rising frequency response characteristics , some have a wide flat characteristic to suit many types and degrees of deafness.

Head-worn aids tend to have an average frequency response which is more level than that of body-worn aids. Despite the limited frequency response, particularly at lower frequencies, head-worn aids amplify well up to 3500 c/s or more; this is a slight improvement on the performance of many body-worn aids, although the amount of amplification available is limited.

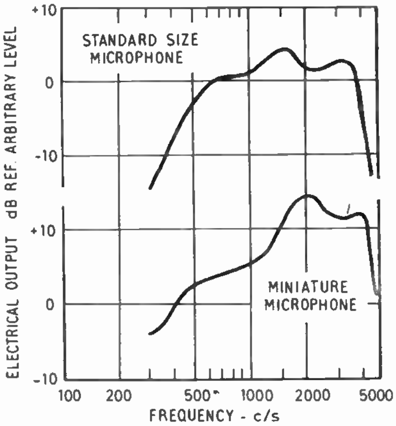


Fig. 6. Frequency response of microphones used in hearing aids at constant acoustic input. Electrical output is measured across the microphone terminals.

3.1.1. Body-worn aids

As stated earlier, the overall frequency responses of most body-worn aids conform to the recommendations of the Medical Research Council's report. Using a modern transistor aid as an example, the recommended response is obtained by a suitable combination of microphone and earphone. The small standard type of hearing aid microphone has a response as shown in Fig. 6, although this depends to some extent on the size and quality of the microphone, together with the manner in which it is housed. It is coupled to a transistor amplifier, having three or more stages and producing a flat frequency response characteristic with perhaps a slight rise at the higher frequencies. Consequently, the overall frequency response of the complete amplifying unit is determined largely by that of the microphone. Owing to the low sensitivity of the microphone and earpiece the electrical gain of the amplifier has to be in the order of 85 dB for an acoustic gain of 50 dB.[9]

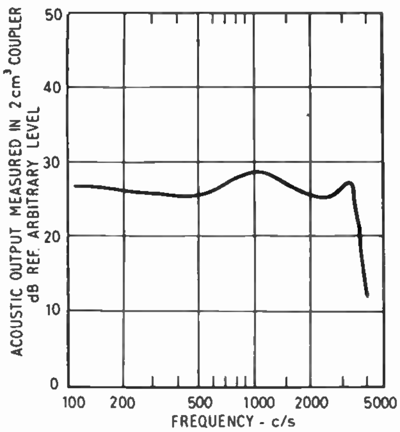


Fig. 7. Typical hearing aid receiver frequency response with constant current electrical input.

The earpiece or receiver is a moving iron device which controls to a very large extent the high frequency response of the aid, and is connected to the main unit by means of a stranded cord. The overall response of the microphone and amplifier (Fig. 6) when added to the response of a typical receiver

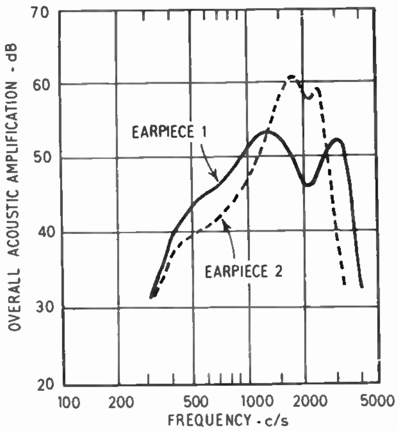
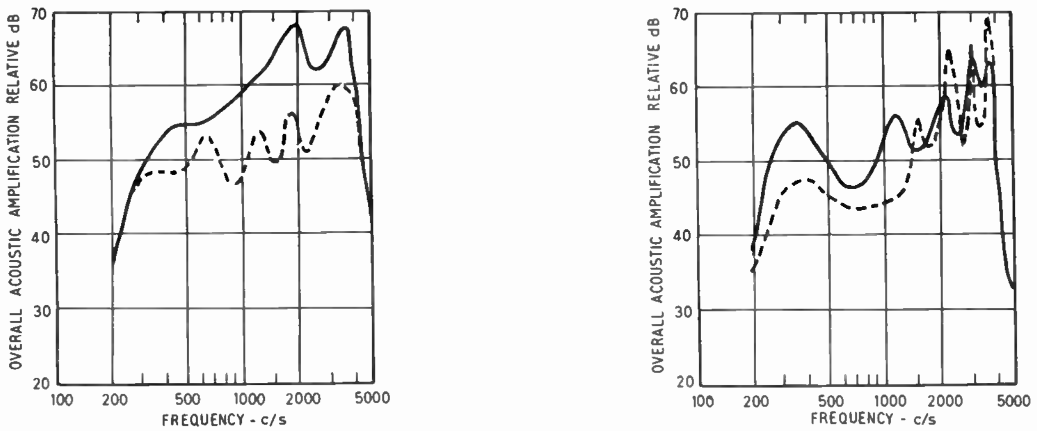


Fig. 8. Body-worn hearing aid frequency response using two receivers of differing characteristics measured in the free field condition.



(*a*) Top curve normal response of aid—lower curve 10 in length of 1·5 mm dia. plastic tube between receiver and coupler.

(*b*) Top curve $5\frac{1}{4}$ in length plastic tube—between receiver and coupler. Lower curve $2\frac{1}{8}$ in length plastic tube.

Fig. 9. Effect of thin plastic tubing connected between a button earpiece and a 2 cm^3 acoustic coupler.

(Fig. 7) gives the air-to-air response of a body-worn aid shown in Fig. 8. The effect upon the overall frequency response of changing one receiver for another of different characteristics is also shown in Fig. 8. In this manner, the response can be readily altered by having a range of receivers for any particular aid, but an increase in amplification is accompanied by a subsequent decrease in the frequency range of the aid, particularly at high frequencies.

In addition to the range of receivers available many aids have some form of switched tone control. An average aid may have three tone positions so as to obtain one of the following frequency responses:

(*a*) the maximum bandwidth of the aid,

(*b*) a low-frequency cut,

(*c*) a high-frequency cut.

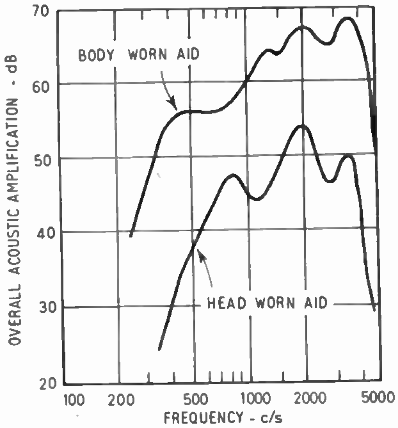


Fig. 10. Frequency response of head-worn and body-worn hearing aids.

These three characteristics are achieved by the use of a series of shunt capacitors in the amplifier circuit, thus enabling the same type of aid to be used by many people. The low-frequency cut is particularly useful in reducing background noises.

The overall response can be further modified by means of acoustic filters fitted into the bore of the receiver. Many hearing aid receivers are fitted with long plastic tubes so that the earpieces may be hidden under the clothes; the less conspicuous tube leads to the mould in the ear. The effect of this tube on the frequency response of the aid is very pronounced (Fig. 9). Such tubes generally produce undesirable resonances and anti-resonances but, in some cases tubes can be used to advantage. For example, a certain hearing aid having a peak in its characteristic at about 3000 c/s will be rather harsh in tone to some users, and those with perceptive forms of deafness would find this type of sound uncomfortable. Unless a range of receivers with varying responses, or some form of tone control is available, the acoustic output of the aid cannot be varied, but the addition of a $10\frac{1}{2}$-in length of tube would have a considerable effect in reducing the high frequencies (see Fig. 9). Thus, many people would be able to use this otherwise unsuitable aid.

3.1.2. Head-worn aids

In the head-worn type of aid, the microphone, amplifier, battery and earpiece are all housed in the same container which is worn behind the external ear, and the sound is conducted to the ear by means of a plastic tube. Owing to the reduction in size of microphone and earpiece used in head-worn aids, the frequency response is restricted at the lower frequencies.

The frequency response of a typical aid worn behind the ear compared with that of a body-worn aid, is shown in Fig. 10.

The amplification of the head-worn aid is restricted not only by component size but also by acoustic feedback both through the microphone and mechanically through the case. The elimination of feedback through the case is a major criterion and has to be overcome before an aid can be put into production. Acoustic feedback is a practical problem which every hearing aid user faces, but head-worn aids present a much greater problem because the effective distance between the transmitter and the receiver is so short. Also, the proximity of the head both to the microphone and earpiece provides another route for acoustic feedback. The art of taking impressions and making well-fitting earmoulds becomes very important if the full available amplification of the aid is to be used effectively.

3.1.3. Behind-the-ear aids

Behind-the-ear hearing aids (Fig. 11) have acoustic amplification up to 45 dB and maximum acoustic outputs of up to 120 dB ref. 0·0002 dyne/cm^2 which is as great as that available from many body-worn aids. If the maximum amplification is to be used, an efficient acoustic seal must be obtained between the receiver and the ear, and the frequency response characteristic of the aid must not have excessive peaks.

A receiver which is not built into the case but uses a conventional button earpiece may increase the amplification by about 10 dB to 55 or even 60 dB, and the maximum output to 125 dB ref. 0·0002 dyne/cm^2.† The frequency response characteristic is more even, thus giving greater intelligibility of speech. Feedback is reduced because no direct connection exists between receiver and aid. The resultant increase in amplification frequently enables a head-worn aid to be used, instead of the larger body-worn type. The head-worn aid, when mounted on a head band, is an advantage for small children, where a body-worn aid is difficult to fit or wear (see Fig. 12).

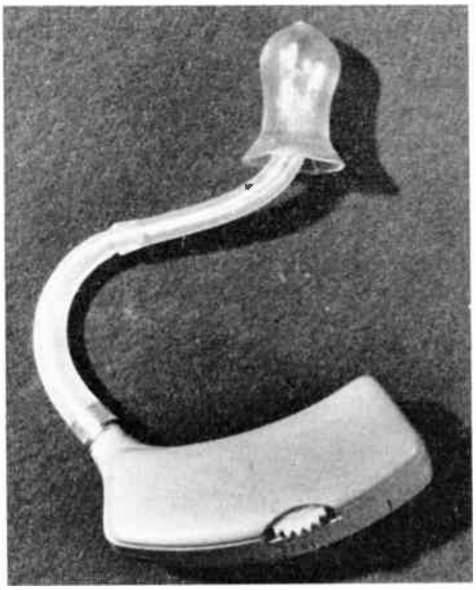


Fig. 11. Hearing aid designed to be worn behind the ear. Size 1·4 in × 0·5 in × 0·5 in.

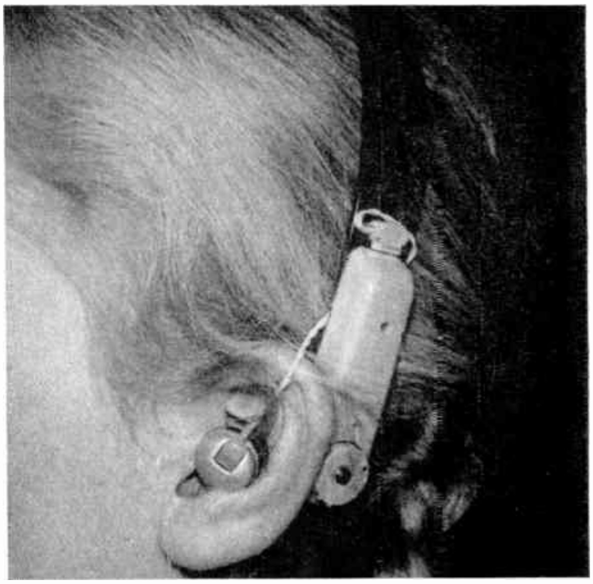


Fig. 12. Hair slide type of hearing aid fitted to a head-band. (*Courtesy of Lectron Hearing Aid Co.*)

3.1.4. Spectacle aids.

Spectacle hearing aids possess a great advantage in providing the most convenient and inconspicuous means of wearing a binaural system of aids; each arm may contain a complete aid, or a common microphone mounted in the 'bridge' may feed separate amplifiers —one in each arm of the spectacles.

The placing of the microphone in the 'bridge' of the spectacles is claimed to give better directionality and discrimination of speech in noisy conditions. Behind-the-ear aids possess a directional effect which in practice causes some difficulty to users in noisy surroundings, particularly if the source of the noise is behind the user.

Spectacle aids also make use of bone conduction which is advantageous to the patient who has a

† Two pressure scales are commonly used in acoustics; one has its zero at 1 dyne/cm^2 and the other at 0·0002 dynes/cm^2. A pressure of 0·0002 dynes/cm^2 at 1000 c/s is considered to be the threshold of hearing at this frequency and this scale is called the Sound Pressure Level Scale (SPL). The difference between the SPL scale and the 1 dyne/cm^2 scale is 74 dB.

bone conduction threshold appreciably higher than that of air conduction, or where the use of an ear plug would be undesirable. Bone conduction is normally confined to body-worn aids because the amount of power required to operate the vibrator is relatively large. One arm of the spectacles contains the bone-conduction pad, whilst the other carries the microphone and amplifier, thus providing an 'invisible hearing aid' with 'nothing in the ear'. Unfortunately the number of persons with useful bone conduction is limited, being about 2% of all cases of individuals with significant hearing defects. Moreover the amplification available with this type of aid is small although a body-worn aid with a bone conduction vibrator fitted to the arm of a pair of modified spectacles can give more amplification.

3.1.5. In-the-ear aids

A hearing aid made to fit entirely within the ear canal is obviously the ideal solution, but so far, the various considerable technical problems in providing such an aid, particularly in preventing acoustic feedback, have not been fully solved. 'In-the-ear' aids are marketed, but the degree of amplification available is only about 30 dB. As the aid is fitted in the external ear, which can be considered as a shallow dish reflector, the problems of feedback become extremely acute and in many cases prevent the use of the full amplification of the aid. The 'in-the-ear' aid itself is in many cases more conspicuous than the 'behind-the-ear' model. The oldest form of hearing aid, the cupped hand held behind the ear, gives a gain of about 9 dB over the useful part of the speech spectrum.[5]

3.2. *Aids used for Educational Purposes*

Individual hearing aids provide an adequate means of amplification for most purposes. If the aid has to be used for school purposes, and the teacher requires the whole of the pupil's attention, the distractions caused by room noise may become excessive; also, the limited frequency response of the individual aid is not suitable for speech training purposes. The group hearing aid has a much wider frequency response than that of the individual aid, but, contrary to popular opinion, the amount of amplification obtained from a group hearing aid, is not necessarily greater than that obtainable from individual aids, although the higher amplification from the individual aid is over a more limited frequency range. The teacher, therefore, prefers to use a group hearing aid so that both the excessive noise factor and the restricted frequency response are eliminated. The group aid consists of a single teacher's microphone connected to an audio frequency power amplifier with from 1 to 10 pairs of headphones, one pair for each student, linked to the amplifier output. The head phones are fitted with separate volume controls for each ear and can be readily adjusted by the wearer. In many group aids, each pupil has his own microphone directly in front of him. By means of a suitable mixing unit the pupil can hear the speech of the teacher and/or the rest of the class in addition to his own speech; if his speech is bad, his own voice can be eliminated at the teacher's discretion. Such equipment requires to be operated from the mains. Some smaller battery operated training units having transistor amplifiers may be used with only one or two pairs of headphones.

The real advantage of the group aid system is in the use of headphones in place of the normal insert receiver. The response of a high quality pair of external earphones (as shown in Fig. 13) enables a much wider frequency response to be maintained up to relatively high output sound levels. A good quality microphone has a frequency response better than that of the earphones and an overall flat frequency response characteristic can be obtained extending up to 6 kc/s; the power amplifier itself has a level characteristic.

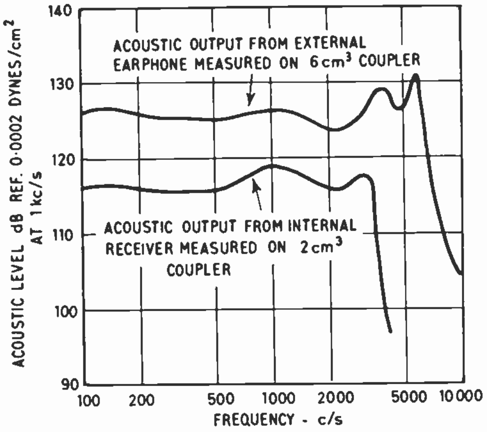


Fig. 13. Frequency response of an external and an insert earphone measured at a constant voltage electrical input.

Despite the high quality of the equipment and the great skill of the teachers many profoundly deaf children do not acquire speech which is intelligible to everyone.

4. Choice of Aids for Various Types of Deafness

Forms of deafness can be broadly divided into two classes—perceptive and conductive. The perceptive type of deafness is due to defects in the inner ear, which cause distortion and a serious loss in the ability to understand the sounds heard. Conductive deafness is caused by defects in the middle ear and normally there is attenuation of sound without the distortion resulting from the perceptive type. These two types of deafness present quite different problems. The perceptive type is always seeking clarity of speech and in many cases a hearing aid is of little use. Elderly people normally have a form of perceptive deafness and this is one reason why they do not get on well with hearing aids.

The conductively deaf patient is one who benefits most from the use of an aid, as all that is required is an increase in the volume of sound coming into his ear.

The various degrees of hearing loss are not easily put into categories, but the following is a rough guide:

(*a*) A person who is profoundly deaf has a hearing loss in excess of 90 dB.

(*b*) A person who is severely deaf has a loss of 60–90 dB.

(*c*) A partially deaf person has a loss of up to 60 dB.

Hearing aids for the profoundly deaf person are limited to body-worn types because of their greater amplification. The number of manufacturers producing very high powered instruments is small since the demand is limited and specialized; hence the choice of suitable aids is at present extremely limited.

The severely deaf person is also confined to body-worn aids but the choice is much larger as more manufacturers produce models powerful enough for him. The individual buying an aid, therefore, has to decide not only which aid is likely to give him the most satisfaction but also what extra facilities, such as an inductive pick-up coil, a tone control, etc. are suitable for him.

The partially deaf person has a very wide range of models to choose from and almost any type of head-worn or body-worn aid would probably suit him. He may prefer a head-worn aid because this type is not so conspicuous; such aids are not affected by the noise of rustling clothes, but wind noises can be a serious problem when they are used out of doors. The elderly person who has difficulty in moving his arms and fingers, however, may find head-worn aids difficult to operate because of the very small controls and may prefer to use a body-worn type with larger and easily operated controls.

5. The R.N.I.D. Hearing Aid Test Service

One problem facing the user is that posed by a faulty hearing aid; he never knows whether it is the performance of the aid or his hearing which has changed.

The R.N.I.D. Hearing Aid Test Service was started in May 1961 by the Technical Department of the R.N.I.D. to give guidance in these cases. This Department was originally set up 15 years ago to inform the deaf public of the comparative performances of various types of hearing aid. For this specific service, special automatic equipment was installed in the laboratory (Fig. 15) so that frequency response curves from the hearing aid being tested could be taken quickly and recorded on pre-printed paper. This service is completely free to anyone, both in Great Britain and from abroad, and is believed to be the only one of its kind in the world.

In the first six months of the Hearing Aid Test Service 210 aids were tested, while another 400 people were seen individually for advice regarding their aids. The surprising point concerning the 210 aids tested was that no less than 31 different manufacturers were represented. The majority of the aids tested were of the body-worn type although in the latter 3 months more head-worn aids were being handled.

After servicing or repair, no check on the overall acoustic response is generally available. It is common practice, therefore, for the supplier or repairer to listen to the aid by ear as a check but a reliable and accurate judgment is not possible with this subjective procedure. Some servicing agents are now installing equipment capable of making the necessary objective measurements. More important still, certain manufacturers are developing the practice of taking frequency response curves for each new aid. These may be issued with the aid or filed for future reference in the event of a change in the performance.

R.N.I.D. experience confirms that there are three main types of complaint, which are already fairly well known to otologists and hearing aid designers:

(*a*) Lack of clarity in speech heard by the user,

(*b*) Acoustic feedback,

(*c*) Noise produced by friction between the case and clothing.

5.1. *Lack of Clarity*

One of the most common complaints was the lack of clarity in speech heard by the user; this defect occurred in approximately 30% of the aids tested.

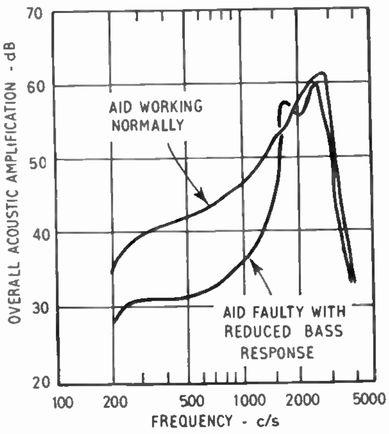


Fig. 14. Frequency response of a hearing aid with reduced bass response.

In addition there were complaints of lack of clarity which were in fact not due to the aid, but to the user having a perceptive form of deafness. In nearly every case, the cause of the complaint was found to be due to a restricted low frequency response. The frequency response of a typical defective aid, compared with that of an aid working normally, is shown in Fig. 14. In some cases, the defect was caused by faulty earpieces or microphones. A poor low-frequency response in an aid can easily pass undetected, especially by a deaf child whose development of speech and vocabulary depend on the sounds coming through his aid.

5.2. *Acoustic Feedback*

Acoustic feedback was also responsible for a large number of complaints and in practically every case this feedback was due to poor fitting earmoulds. The majority of these moulds were individually made in a hard acrylic base which did not produce an efficient seal. Ears vary a great deal in shape and in many cases a good fit is very difficult to obtain, particularly with elderly people and young children. New soft plastic materials have now been introduced; these are resilient and tend to grip the ear better than the hard acrylic.

5.3. *Extraneous Noise*

Noise from rustling clothes presents the biggest problem to all users of body-worn aids. Hearing aid containers are polished to allow the clothing to move freely with a minimum amount of friction. The microphone aperture, controls and clothes clip, however, present surface irregularities which can give rise to case noise. The conventional position for the microphone is on the front face of the aid, but several manufacturers are now mounting the microphone on the top edge. In this way, the largest area of the aid can be made perfectly smooth, so reducing case noise.

The mounting of the microphone is, of course, a very important factor in reducing mechanical noise. Microphone mountings vary from rubber jackets to foam rubber or strip rubber suspension. In many cases the mountings are not resilient enough to be effective; the microphone is often wedged too tightly between the chassis and case.

6. Conclusions

The ideal aid apparently would be one fitted entirely within the ear and which would give sufficient amplification over a wide frequency range with no peaks. In-the-ear aids now available have a relatively small amount of amplification and only over a restricted frequency range. The reduction in the size of aids achieved by the use of transistors enables the amplifier circuit of many head-worn aids to occupy less space than any other single component and is, therefore, no longer the limiting factor in size. Limits in size reduction are now imposed by the microphone, earpiece, volume control and battery. Although very small batteries are available, their capacity is smaller than that of the larger ones, hence the aid becomes uneconomical to run. Any further reduction in size of earpieces and microphones would impair their sensitivity and frequency response.

Hearing aids have now reached a point where any decrease in size would be of little value because of manipulation difficulties. Perhaps an automatically controlled aid might be the ideal method to overcome this problem.

The present emphasis has been on reducing size but much more thought will have to be given to the development of aids for those persons who do not benefit from modern instruments, particularly the profoundly deaf and those suffering from perceptive forms of deafness, and to the study of the problems of using hearing aids in environments that are other than ideal. Despite considerable technical refinements in the production of smaller, more powerful and easily worn aids, the benefit that a deaf person can obtain from a modern hearing aid is not vastly superior to that obtainable from aids produced 15 years ago.

7. Acknowledgments

The author wishes to acknowledge the assistance given by Dr. T. Littler of the Wernher Research Unit on Deafness and Dr. P. Gorman, Librarian, R.N.I.D., in preparing the paper.

8. References

1. J. P. Ashton, "The design of commercial hearing aids" *J. Brit.I.R.E.*, **11**, No. 2, pp. 51–9, 1951.
2. E. Aspinall, "A master hearing aid", *J. Brit.I.R.E.*. **11**, No. 2, pp. 45–50, 1951.
3. C. J. Cameron and E. W. Ayres, "The Medresco hearing aid", *P.O. Elect. Engrs J.*, **44**, pp. 153–8, 1952.
4. H. D. Davis, S. S. Stephens and R. H. Nichols, "Hearing Aids: an experimental study of design objectives", p. 197. (Harvard University Press, 1947.)
5. H. Davis and R. S. Silverman, "Hearing and Deafness", p. 573. (Holt, Rinehart and Winston, New York, 1960.)
6. D. Fry and P. Denes, "The Testing of Hearing Aids", National Institute for the Deaf, No. 490, p. 39. (London, 1951.)
7. M. A. Goldstein, "Problems of the Deaf", p. 580. (Laryngoscope Press, St. Louis, 1933.)
8. "Hearing Aids and Audiometers", Medical Research Council report, special series, No. 261, p. 71. (H.M.S.O., London, 1947.)
9. "Reference Manual of Transistor Circuits", Mullard Ltd., chapter 15, pp. 161–4. (London, 1960.)
10. "Evaluation of Hearing Aids", Office of Scientific Research and Development, Report No. 4666. (Harvard University, May 1945.)
11. E. R. G. Passe, "Deafness; its clinical aspect and the possibilities of its alleviation medically and by deaf aids" *J. Brit.I.R.E.*, **11**, No. 2, pp. 40–4, 1951.

12. H. Thwaite and E. Aspinall, "A transistor hearing aid", *P.O. Elect. Engrs J.*, **52**, pp. 163–9, 1959.

13. L. A. Watson and T. Tolan, "Hearing Tests and Hearing Instruments", page 597. (Williams and Wilkins, Baltimore, 1949.)

14. J. Jessop, "Some psycho-acoustic and engineering aspects of hearing aid design". Paper read at a meeting of the Institution held in London on 14th February 1961. (Unpublished.)

9. Appendix 1

R.N.I.D. Service Testing Procedure

In order to obtain a quick overall evaluation of the user's hearing aid, tests are limited to the following procedures:

(1) Making a check of the battery voltage and inspecting the aid for electrical and mechanical defects.

(2) Taking a frequency response curve for an input sound field below the limiting point of the aid, usually -20 dB ref. 1 dyne/cm^2, to obtain the amplification of the aid.

(3) Taking a frequency response curve for a high input sound field, i.e. $+ 20$ dB ref. 1 dyne/cm^2, to derive the maximum acoustic output of the aid.

(4) Taking the transfer characteristic (input to output) at one or two fixed frequencies and examining the acoustic output for distortion.

(5) Assessing the transient performance of the aid by interrupting the acoustic input and examining the output wave form either on a long persistence oscilloscope or on a high-speed pen-recorder.

The frequency response curves are taken with the maximum gain of the aid under free field conditions.

All aids are tested in the R.N.I.D.'s anechoic chamber (which has a working space of 9 ft × 7 ft × 6 ft with 31 in wedges and a 4½ in air gap behind them) (Fig. 15). The test signals, generated by a beat frequency oscillator, are fed into a 20 watt power amplifier which drives a 15 in loudspeaker in the chamber. A condenser microphone close to the hearing

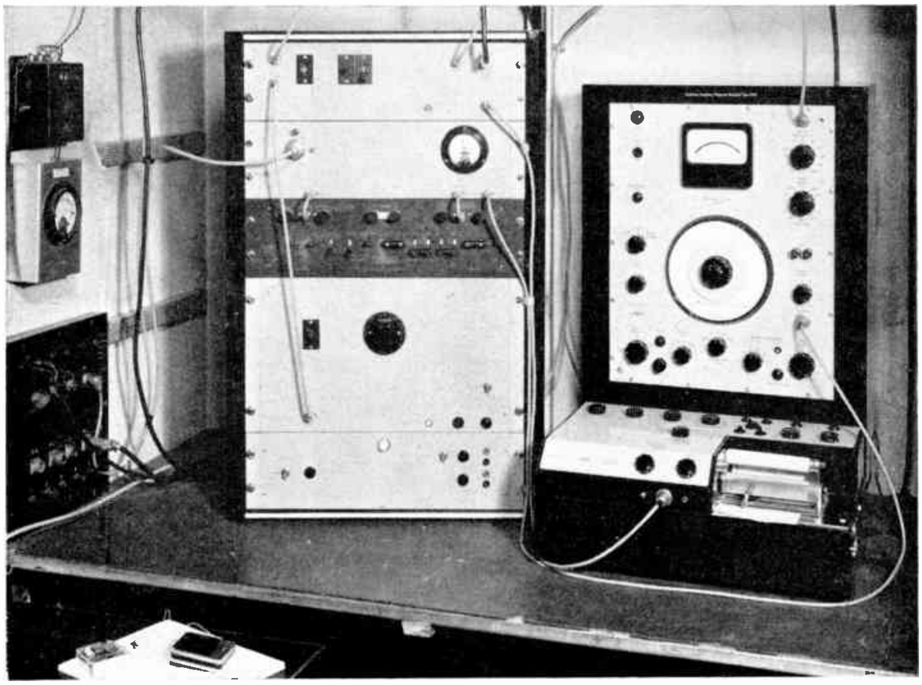


Fig. 15. Anechoic chamber and test equipment set up for a hearing aid test in the R.N.I.D. technical laboratory. (*Above*) The anechoic chamber with a hearing aid is being set up for test. (*Right*) The hearing aid test equipment. The aid is placed inside the anechoic chamber and connected to this equipment which draws a response curve, automatically, in a few seconds on a moving chart.

aid monitors the sound field and signals from this microphone are amplified and fed to a monitoring amplifier. The gain of the monitoring amplifier can be changed by means of an adjustable attenuator pad. The output from the monitoring amplifier operates a compression circuit in the oscillator; the attenuator on the monitoring amplifier controls the output of the oscillator and, therefore, the sound field.

The acoustic output from the aid is fed into a $2\ \text{cm}^3$ standard hardwall acoustic coupler[†] which has a condenser microphone in the base.

The signals from the microphone are amplified and fed to a pen-recorder. Synchronization is achieved between the pre-printed paper on the pen-recorder and the beat frequency oscillator, enabling the sweep frequency response to be obtained. A slow sweep speed is used to ensure that no sharp peaks in the response of the aid are missed.

10. Appendix 2

Batteries for Hearing Aids

Body-worn aids usually use Leclanché dry cells having an e.m.f. of 1·5 V and capable of providing a working life of up to 300 hours. Consequently they involve a very low running cost. Two standard sizes frequently used are

(*a*) $\frac{9}{16}$ in diameter × 2 in and

(*b*) $\frac{7}{16}$ in diameter × $1\frac{3}{4}$ in.

The reduction in the size of some of the aids (e.g. head-worn types) has, however, necessitated the use of smaller batteries which are available only at a higher price and possess a significantly lower capacity. Mercury cells having an e.m.f. of 1·34 V are often used for this purpose; although relatively expensive they have a greater capacity, per unit size, than the dry cell. The mercury cell also has the advantage of almost constant working voltage as compared with the steadily decreasing working voltage of the dry cell.

Mercury cells are produced in the form of a thick button and have sufficient capacity to operate an average head-worn aid for approximately 80 hours. Thus a person using an aid all day long has to change batteries every three or four days. Owing to the expense of these mercury cells, a rechargeable cell is desirable and nickel cadmium cells have been produced for this purpose. It has a relatively small capacity if produced in the size required for head-worn aids, (e.g. a cell of 0·45 in diameter × 0·2 in gives about 20 mAH) so its use in the smaller type of aid is restricted because its life is not sufficient to last a full day without recharging. The larger rechargeable cell, (0·61 in diameter × 0·23 in) with a capacity of 50 mAH, would be ideal but its size precludes its use in small head-worn aids.

[†] American Standard method for the coupler calibration of earphone No. A.S.A. Z24.9—1949.

Manuscript first received by the Institution on 5th March 1962 and in final form on 17th January 1963. (Paper No. 831/EA8).

Standard Frequency Transmissions

By

J. McA. STEELE, B.Sc.(Eng.)

(*Communication from the National Physical Laboratory, Teddington, Middlesex.*)

In the United Kingdom there are three stations transmitting standard frequency signals in the l.f. and v.l.f. bands. Two are at Rugby: MSF 60 kc/s operated by the Post Office on behalf of the National Physical Laboratory and GBR 16 kc/s, which is also controlled in frequency by the MSF oscillator. The third is the Droitwich 200 kc/s transmitter of the British Broadcasting Corporation which is regulated independently of the other two stations. The deviations of all three frequencies from their assigned values are determined at the N.P.L. with respect to the caesium atomic frequency standard and until July 1962 the results were published in *Electronic Technology*. When that journal ceased to exist a monthly bulletin continued to be circulated to users requiring the highest accuracy of calibration. It has now become clear, however, that the increasing applications of precise frequency control justify a wider distribution and accordingly the Laboratory has been pleased to accept the suggestion of the Institution that the results should appear each month in this *Journal*. The values for June 1963 are given on page 34 and there follows a brief account of the methods of measurement together with an indication of the present limits on the accuracy of the several transmissions.

Two systems of comparison are in use. In the first, and simplest, the received transmission and a signal of the same frequency obtained from the Laboratory frequency standard are combined in a phase-detector and the slowly varying output recorded on a continuously moving chart. The quasi-sinusoidal detector output falls to zero at relative phase angles of $\pi/2$ and $3\pi/2$ and a mean frequency value is obtained from the number of excursions of π or 2π occurring in a known time interval. Recorders of this type are in continuous operation for GBR and Droitwich. At 16 kc/s the phase changes between received and locally-generated signals are integrated over a 24-hour period centred on 0300 U.T., this interval being chosen to facilitate international frequency comparisons at v.l.f. The phase discrimination at 16 kc/s is about 6 degrees or 1 microsecond and the mean frequency over 24 hours is thus known to 1 part in 10^{11}. The higher frequency of Droitwich enables this same precision to be obtained in a shorter averaging period of 2–3 hours but in practice it is found that variations in frequency of several parts in 10^{10} may occur during the day. The accuracy of the tabulated values is therefore restricted, for the present, to 1 part in 10^{10}. In using the Droitwich carrier it has been found necessary to introduce a crystal oscillator, locked in frequency to the received signal, before the phase detector. This serves the double purpose of removing the broadcast modulation and at the same time preserving an output during the short interruptions in the carrier produced by bursts of 100% modulation.

On weekdays Droitwich is also measured by the second method, using the equipment developed primarily for the calibration of MSF 60 kc/s. This makes use of a servo-driven phase-shifter, operated by an error signal from the phase detector, to maintain a constant phase difference, of approximately $\pi/2$, between received and locally-generated frequencies. The phase shifter is linear to ± 0·01 μs and both analogue and digital read-out of relative phase to this accuracy are provided. The MSF values in the table are obtained from the accumulated phase change during the hourly period of transmission. The results are given to an accuracy of 1 part in 10^{10} but in future months this will be increased to 1 part in 10^{11}. When Droitwich is measured by this method the results are an average extending over an hour centred on 1030 U.T.

The B.B.C. has announced that the deviations of the Droitwich carrier will be maintained within the limits of ± 50 parts in 10^{10} of nominal frequency, 'nominal' being defined on the basis of the second of Ephemeris Time (E.T.) and realized in practice as 9 192 631 770 cycles of the caesium resonance frequency at zero field.† In this respect Droitwich differs from MSF and GBR which are offset from nominal by, at present, 130 parts in 10^{10} so that the time signals, which are locked to the carrier frequencies, will remain as nearly as possible in agreement with Universal Time (U.T.2). This offset is subject to annual review and may be changed if necessary on the first day of January each year.

Since the suspension of publication last year the deviations of MSF/GBR from the offset value have not exceeded ± 3 parts in 10^{10}. The frequency of Droitwich in the period since closer control was introduced has varied over a range of + 27 to − 60 parts in 10^{10}, the frequency increasing at an average rate of rather less than 1 part in 10^{10} per day. It is not proposed to publish these results in detail but they are available on application to the N.P.L.

This summary has been prepared as part of the research programme of the National Physical Laboratory and is published by permission of the Director.

† L. Essen, "Atomic time and standard frequency transmissions", *J. Instn Elect. Engrs*, 9, pp. 247–50, June 1963.

U.D.C. 519.272:534.88.

Space-Time Sampling and Likelihood Ratio Processing in Acoustic Pressure Fields[†]

By

PHILIP L. STOCKLIN,
A.B., M.S., Ph.D.[‡]

Presented at the Symposium on "Sonar Systems" in Birmingham on 9th–11th July 1962.

Summary: Space–time sampling plans based on a three-dimensional Whitaker sampling function are reviewed for single frequency and series band-limited time limited acoustic pressure fields. Following a brief discussion of probabilistic wave fields, likelihood ratio measurement procedure is outlined. Two examples are treated: the one-dimensional resolution problem for Gaussian signal in Gaussian noise, and the space-time likelihood ratio processor for signal known exactly in Gauss and Markov-Gauss noise.

1. Introduction

Basic problems in present-day acoustics, especially underwater acoustics, possess many highly complex elements and consequently bear little relation to the basic problems in this field of even one or two decades ago. To aid in their solution, the powerful analytic techniques of related fields, especially those of mathematical statistics, offer great potential. To assist in bringing these techniques to bear upon basic acoustic problems, this paper develops a conceptual and analytic approach to acoustic pressure measurement from the point of view of statistical decision theory.

Normally, measurement errors are associated with run-to-run fluctuations in the properties of the source, the propagating medium and the receiver. While under well-controlled laboratory conditions fluctuations are often the major source of measurement error, the problem of the underwater acoustics experimenter in a state of partial knowledge and control is not only trial-to-trial fluctuation, but additionally the uncertainty as to which of a number of gross experimental conditions holds true. For example, lack of precise knowledge of the deep velocity gradient or source location may lead him to regard the deep refracted propagation mode as predominant, while in fact surface channel propagation may dominate.

Clearly, such lack of knowledge can cause far greater error than fluctuation. Both fluctuation and lack of precise knowledge of experimental conditions are taken into account in the probabilistic description of wave fields used in space–time decision theory. In this description the number of probability density functions describing the field is the number of possible gross experimental conditions, identified in the following sections as boundary conditions, and the detailed form of the density functions is largely determined by the physical statistics of the fluctuation.

The basic distinction between the repeated well-controlled experiment and the partially-controlled experiment is the possibility for a number of systematic errors or, more generally, correlated errors to occur in the latter. Actually, both random and correlated errors may occur in either the time or space extension of measurement. It is logical, therefore, to seek a theoretical framework for measurement which incorporates these two extensions within a single quantitative description. Space–time decision theory provides such a description. Since the state of knowledge (of the experimenter) of the acoustic field to be measured is incomplete prior to experiment, and in fact may vary from an almost completely determined field to an almost completely undetermined field, the theoretical framework for measurement should allow for this range of prior knowledge. The act of measurement—that is, the act of processing—is defined as an operation upon the acoustic field in space and time which changes the state of knowledge of the experimenter with respect to one or more properties of the field. This change, of course, is usually an improvement in the state of knowledge for the properties of interest, or, in statistical language, obtaining the most accurate estimate of the statistics having to do with these properties of the probabilistic wave fields on which the experiment is performed. Finally, the purpose of a theoretical framework for measurement or processing is to find the limits to acoustic field measurement by finding the minimum average total error associated with the measurement operation.

With these requirements in mind, the following two postulates are made:

[†] Condensed from "Limits of Measurement of Acoustic Wave Fields", a thesis submitted in partial fulfilment of the requirements for the Ph.D. degree, University of Connecticut, June, 1962.

[‡] Department of the Navy, Office of Naval Research, Washington 25, D.C.

Postulate 1: The most general description of an acoustic pressure field over the volume V, time T and the band of frequencies $(0, W)$ available for measurement is the probability density function describing the joint probability of occurrence of the acoustic pressure x at every point in V over T in band $(0, W)$.

Postulate 2: The most efficient measurement, in the sense of minimizing the average total error of property estimation, is the measurement described by the likelihood ratios formed from the set of probability density functions describing the acoustic field. One such probability density function is associated with each possible value of the property to be measured or, more generally, each possible set of values for the set of properties to be measured.

The first postulate is the basic one of the present development. Normally, acoustic fields are treated as deterministic—that is, as explicit solutions to the wave equation. Given the boundary conditions, then, the most convenient co-ordinate system can be found and the general series form of the solution written down. If the boundary conditions completely specify the field, a unique particular solution can be found—that is, a unique set of coefficients in the general series—and the field properties of interest specified uniquely.

The problem of measurement, on the other hand, is concerned with the case of incomplete boundary conditions—under-specification of the field. This may be due to stochastic elements of the physical situation—fluctuations from point to point in space or time, for example—or to uncertainty as to which of a set of possible boundary conditions is true. The unique solution of the completely specified case is then replaced by a probability density function over the field, according to the first postulate. This density function specifies the probability of occurrence of each possible unique solution for each possible set of property values. In terms of probability density functions, knowledge of boundary conditions which completely specifies the field, and thus uniquely specifies the field properties of interest, corresponds to a delta-function probability density function and removes the need for measurement.

Since the acoustic field is continuous in V over T, an infinite order joint probability density function is implied in the first postulate. Specification of the frequency band $(0, W)$, however, implies that small but finite separation in time or space should result in a largely predictable acoustic pressure—that is, that the density function should be of finite order for finite V and T. Because of its central importance in specifying the degrees of freedom of space–time likelihood ratio measurement, this point is developed in quantitative detail in Section 2 for both single frequency and band-limited acoustic fields. Having arrived at finite order descriptions of acoustic fields, the set of probability density functions describing the probabilistic acoustic field is taken up in Section 3. Likelihood ratio measurement is then described and followed by two examples: one dimensional resolution of one of three Gauss signals in Gauss noise, and the space–time likelihood ratio processor for a signal known exactly in Gauss and Markov–Gauss noise.

2. Space–Time Sampling Plans

As indicated in the previous section, the basic reason for space–time sampling plans is to permit replacing the infinite order joint probability density functions describing the wave field by finite order functions, with little or no loss in ability to describe the field at every point. Having done this, results may be established in two areas:

(a) First, physical interpretation of likelihood ratio measurement becomes possible in terms of both finite and discrete spatial and time measurements. This fact is important since in practice measurements are almost always made at discrete points in space and for finite periods of time.

(b) By associating the number of space–time sampling points with the number of degrees of freedom used in establishing the statistics of the wave fields, limits in measurement accuracy are readily found, and relations established between these limits and the volume, time and frequency band of the wave field.

A space–time sampling plan may be considered an interpolation formula in which, by expressing the field at any point $(x, y, z; t)$ in terms of its values at specific points, the entire field may be reconstructed from measurements made only at these points. The spatial sampling function used in developing these theorems is the Whitaker interpolation function,[1] also called the Nyquist–Shannon sampling function.[2] Its form in cartesian co-ordinates for a single frequency field is

$$S(l, m, n; x, y, z) = S(l; x) \, . \, S(m; y) \, . \, S(n; z)$$

$$= \frac{\sin \pi \left(\frac{2x}{\lambda} - l\right)}{\pi \left(\frac{2x}{\lambda} - l\right)} \cdot \frac{\sin \pi \left(\frac{2y}{\lambda} - m\right)}{\pi \left(\frac{2y}{\lambda} - m\right)} \cdot \frac{\sin \pi \left(\frac{2z}{\lambda} - n\right)}{\pi \left(\frac{2z}{\lambda} - n\right)} \qquad \text{......(1)}$$

where l, m and n are integers and λ is the wavelength

This set of functions is selected for two reasons. They are orthogonal over the infinite (space) interval, which is necessary for simple development of the integral equation giving the sampling coefficients. Secondly, they have the less familiar property of *point* orthogonality. By point orthogonality is meant that if the point in space

$$x = b\lambda/2, \quad y = d\lambda/2, \quad z = e\lambda/2$$

is considered, where b, d and e are integers, then at that point all the sampling functions of the set (1) are zero except the one function $S(b, d, e;\ x, y, z)$ which is unity. Point orthogonality results in any coefficient being just the value of the field itself at the sampling point, rather than a function of all the sampled values with this function changing from point to point.

A drawback of this set of sampling functions is that they are individually not solutions of the wave equation, expressed in cartesian (x, y, z) co-ordinates. The latter have been selected because of their symmetry; in spherical co-ordinates, while $\sin kr/kr$ is a solution to the wave equation, similar point orthogonal functions cannot be found for the θ and ψ co-ordinates. Mathematically, the failure of the individual sampling functions to satisfy the wave equation means that without some further restriction on the set of sample coefficients or, alternatively, on the field itself, use of these sampling functions may not be valid. Analysis reveals, however,[3] that the conditions under which they may be used are no more restrictive than those necessary for the more familiar Fourier series expansion, or, indeed, these implied in the application of the wave equation itself.

2.1. *Space and Space–Time Sampling Theorems*

Two spatial and space–time sampling theorems fundamental to the development of likelihood ratio measurement will now be stated. Proofs and discussion are given in Reference 3.

Theorem I (Uniform space sampling of a monochromatic field): If an acoustic pressure field $p(x, y, z;\ t)$ consists entirely of acoustic radiation of a single frequency f, formed from the superposition of fields from an arbitrary number of single frequency sources of arbitrary phase and amplitude, then at any instant of time t, $p(x, y, z;\ t)$ is given within the volume not containing these sources as a function of x, y and z by its complex values at points spaced a half-wavelength apart in x, y and z times a three-dimensional sampling function, such sampling extending throughout space:

$$p(x, y, z;\ t) = \mathrm{e}^{\mathrm{j}\omega t} \sum_{l} \sum_{m}^{\infty} \sum_{n} p\left(\frac{l\lambda}{2}, \frac{m\lambda}{2}, \frac{n\lambda}{2}\right) \times$$
$$\underset{= -\infty}{} \times S(l, m, n;\ x, y, z) \quad \ldots\ldots(2)$$

where $p(l\lambda/2, m\lambda/2, n\lambda/2)$ = complex amplitude of the acoustic pressure at $x = l\lambda/2$, $y = m\lambda/2$, $z = n\lambda/2$.

and $S(l, m, n;\ x, y, z)$ is given in expression (1).

Theorem II (Continuous frequency spectrum, finite time): If an acoustic field $p(x, y, z;\ t)$ of interest for finite time T consists almost completely of radiation within the band of frequencies $(0, W)$, then $p(x, y, z;\ t)$ is given everywhere in the volume not containing acoustic sources by its amplitude at discrete points in time and space times a space–time sampling function:

$$p(x, y, z;\ t) = \frac{1}{2WT} \sum_{q=-WT}^{WT} \sum_{s=-WT}^{WT} \underset{= -\infty}{\sum_{l} \sum_{m}^{\infty} \sum_{n}} \left\{ p\left(\frac{l\lambda_q}{2}, \frac{m\lambda_q}{2}, \frac{n\lambda_q}{2};\ \frac{s}{2W}\right) \mathrm{e}^{-\mathrm{j}\pi qs/WT} K(s, q) \times \right.$$
$$\left. \times\ S_q(l, m, n;\ x, y, z)\, \mathrm{e}^{\mathrm{j}\omega_q t} \right\} \quad \ldots\ldots(3)$$

where $\omega_q = 2\pi q/T$, $\lambda_q = 2\pi c/\omega_q$, $K(s,q) \simeq 1$ for almost all q and s, and c = speed of sound.

2.2. *Sampling in Finite Volume V*

For a single frequency field and for a finite volume V, if the number of space-sampling points M is large compared to unity, and if they are a half-wavelength apart as specified in Theorem I, then

$$M \simeq V/(\lambda/2)^3 \quad \ldots\ldots(4)$$

For a field of frequency band $(0, W)$, time T and volume V, if both $WT \gg 1$ and the number of half wavelengths (cubed) in V is large for every ω_q in $(0, W), q \neq 0$, then the number of space–time sampling points, n_{II}, is given by:

$$n_{\mathrm{II}} \simeq (2WT)^2\, V/(\lambda_W)^3 \quad \ldots\ldots(5)$$

where $\lambda_W = c/W$ and $2WT$ time samples are taken at each space sampling point.

Discussion of the reconstruction errors implied by the approximations of eqns. (4) and (5) is given in Reference 4.

3. Probabilistic Wave Fields

A probabilistic acoustic wave field within volume V during time T is a wave field describable by the joint probability of occurrence of acoustic pressure at every point within V and T. The complete description of a probabilistic wave field consists of a set of probability density functions, each density function giving the joint probability of occurrence (per unit volume in M or n_{II} space) of the acoustic pressure over V and T for one possible set of boundary conditions. The need for a probabilistic description arises from uncertainty on the part of the experimenter regarding the complete set of experimental conditions or, in terms of field analysis, uncertainty regarding the set of boundary conditions which completely specify the field. This uncertainty arises from two principal sources: noise or thermal fluctuation in the acoustic field, and lack of knowledge of macroscopic conditions such as source geometry or propagation mode.

3.1. *Formulation in Terms of Space–Time Sampling*

The immediate use of Theorems I and II of Section 2.1 is to permit the use of finite order joint probability density functions to describe the probabilistic wave field. For the single frequency field, the joint probability density function $\mathrm{f}_A({}_1X)$ is written as:

$$\mathrm{f}_A({}_1X) = \mathrm{f}_A({}_1x_1,\ {}_1x_2,\ \ldots\ {}_1x_M) \qquad \ldots\ldots(6)$$

where A represents an assumed set of boundary conditions, ${}_1x_j$ is the sampled acoustic pressure phase and amplitude at time t_1, at the jth space sampling point in V, and M, the total number of spatial sampling points, is given by eqn. (4). Expression (6) is read: "the probability that if A is true, then ${}_1x_1$, and ${}_1x_2$ and ... and ${}_1x_M$ occur at time t_1."

For a field of frequency band (0, W), finite time T and volume V, the joint probability density function is, from Theorem II,

$$\mathrm{f}_B(X) = \mathrm{f}_B({}_1x_1,\ {}_1x_2,\ \ldots\ ,\ {}_1x_M,\ {}_2x_1,\ \ldots\ {}_2x_M,\ \ldots\ {}_{2WT}x_1\ \ldots\ {}_{2WT}x_M) \qquad \ldots\ldots(7)$$

where B represents an assumed set of boundary conditions, ${}_ix_j$ is the acoustic pressure *amplitude* sampled at the ith time instant at the jth sampling point in space, and M, the total number of space sampling points, is given by $n_{\mathrm{II}}/(2WT)$. The expression in (7) is read: "the probability that if B is true, then acoustic pressure amplitude ${}_1x_1$ occurs at the first sampling instant at spatial sampling point number one, ${}_1x_2$ occurs at the first sampling instant at the second spatial sampling point, and ... and ${}_{2WT}x_M$ occurs at the $2WT$th sampling instant at the Mth spatial sampling point.[11]

The probability density function of a wave field may be connected with deterministic fields usually considered in connection with the wave equation in the following manner. Consider a single frequency field in a lossless medium. If it is known that this field is of angular frequency ω and is a single plane wave coming from direction (θ, ϕ), and this state of knowledge is all the experimenter knows about the field, then only two quantities remain to be determined: the phase of the field at any point with respect to that at a reference point, and the (maximum) amplitude of the field at any point. These two quantities can be determined by measurement (sampling) at just one point in the field. If the number of possible single frequency plane waves increases, however, and their directions or phases or amplitudes are unknown, the number of sampling points in space required completely to specify the field increases. The gist of Theorem I is that no matter how many plane waves, phase angles, and maximum amplitudes are unknown, the field can be determined almost completely within a finite volume V by using only a finite number of spatial sampling points.

When considering a time-limited, series band-limited field in a lossless medium, the requirement on a number of sampling points is similar to that for a single frequency field for identical states of knowledge. For the case of a single plane wave, instead of phase relative to a reference point (in time) the waveform being propagated as a plane wave must be found—that is, the phase and amplitude of each of the WT frequency components—must be found with respect to a reference point. This may be done by a Fourier transform of the field measured at one spatial sampling point. For the case of unknown direction, the delay rather than the phase shift between the field measured at two spatial sampling points must be found. A second distinction to be made in the case of a band-limited, time-limited case is the role of time. Here, discrete spatial sampling is achieved by replacing the acoustic field in V outside the time interval of length T by replicas of the field within this time interval. This is no more than an expression of the fact that all the experimenter knows is what he measures in the time available to him. If this time is changed in length, not only do the discrete frequencies change which are used to synthesize the waveform at any one (non-sampled) spatial point, but the locations of the spatial sampling points themselves change. Within a given V, if T is increased then the number of spatial sampling points increases linearly as T according to eqn. (5), corresponding to the acquisition of further information on the field.

Summarizing, a probability density function is constructed for each possible set of boundary conditions (e.g. a number of possible plane wave directions). Fluctuation, the second source of uncertainty, determines the actual form of the density functions; that is, what is meant by the term 'fluctuation' is an acoustic pressure which is a stochastic function of time and/or space, and so is describable only in terms of its probability of occurrence.

3.2. *Stationarity of Space and Time Components*

Consider the probability density function $\mathrm{f}_B(X)$ for a state of knowledge B for a series band-limited, time limited acoustic field in volume V. From Theorem II, such a density function is of order $n_{\mathrm{II}} = 2WTM$ where M is the number of spatial sampling points in V. If X_j is the set of $2WT$ time samples obtained at the jth space sampling point:

$$X_j = ({}_{1}x_j,\ {}_{2}x_j,\ \ldots\ {}_{2WT}x_j) \qquad \ldots\ldots(8)$$

then $\mathrm{f}_B(X)$ given by eqn. (7) may be written

$$\mathrm{f}_B(X) = \mathrm{f}_B(X_1,\ \ldots\ X_M)$$

$$= {}_{1}\mathrm{f}(X_1) \prod_{j=2}^{M} {}_{j}\mathrm{f}_{X_1,\ \ldots\ X_{j-1}}(X_j) \qquad \ldots\ldots(9)$$

where the subscript j on f indicates explicitly that the form of f may depend on j, and where B is omitted on the right side for brevity.

If ${}_{i}Y$ is taken to be the set of M spatial samples of acoustic pressure for the ith sampling point in time:

$${}_{i}Y = ({}_{i}x_1,\ {}_{i}x_2,\ \ldots\ ,\ {}_{i}x_M) \qquad \ldots\ldots(10)$$

then $\mathrm{f}_B(X)$ may also be expressed as:

$$\mathrm{f}_B(X) = \mathrm{f}({}_{1}Y,\ {}_{2}Y,\ \ldots\ ;\ {}_{2WT}Y)$$

$$= {}_{1}\mathrm{f}({}_{1}Y) \prod_{i=2}^{2WT} {}_{i}\mathrm{f}_{{}_{1}Y,\ \ldots\ {}_{i-1}Y}({}_{i}Y) \qquad \ldots\ldots(11)$$

where the subscript i on f indicates that the form of f may depend on i, and again B has been omitted for brevity.

Considering (9) if, except possibly for end effects at or near $j = 1$ and $j = M$,

$${}_{j}\mathrm{f}_{X_1,\ \ldots\ X_{j-1}}(X_j) = \mathrm{f}_{X_1,\ \ldots\ ,\ X_{j-1}}(X_j) \qquad \ldots\ldots(12)$$

then ${}_{j}\mathrm{f}_{X_1,\ \ldots\ X_{j-1}}(X_j)$ is said to be locally statistically stationary in space[5]—that is, the probability density *function* of order $2WT$ describing the acoustic pressure field at any jth point in space (in V) is independent of j.

Similarly, considering (11), if, except for possible end effects at or near $i = 1$ and $i = 2WT$,

$${}_{i}\mathrm{f}_{{}_{1}Y,\ \ldots\ {}_{i-1}Y}({}_{i}Y) = \mathrm{f}_{{}_{1}Y,\ \ldots\ ,\ {}_{i-1}Y}({}_{i}Y) \qquad \ldots\ldots(13)$$

then ${}_{i}\mathrm{f}_{{}_{1}Y,\ \ldots\ ,\ {}_{i-1}Y}({}_{i}Y)$ is said to be locally statistically stationary in time—that is, the probability density *function* of order M describing the acoustic pressure field over V at the ith instant of time is independent of i.

If local space stationarity or local time stationarity holds, then eqns. (9) or (11) may be written, respectively, as :

$$\mathrm{f}_B(X) = \mathrm{f}(X_1) \prod_{j=2}^{M} \mathrm{f}_{X_1,\ \ldots\ ,\ X_{j-1}}(X_j) \qquad \ldots\ldots(14)$$

for local space stationarity, and

$$\mathrm{f}_B(X) = \mathrm{f}({}_{1}Y) \prod_{i=2}^{2WT} \mathrm{f}_{{}_{1}Y,\ \ldots\ {}_{i-1}Y}({}_{i}Y) \qquad \ldots\ldots(15)$$

for local time stationarity. It should be noted that invariance of the functions themselves is the quantity specified in eqns. (14) and (15), and not the values of the acoustic pressure.

Local stationarity is extremely useful in furnishing a guide for the separation of space and time measurement operations. General stationarity, while useful to establish local stationarity, does not appear to be an essential part of the present development, although desired properties of the sampling coefficients may in some cases depend upon the existence of general stationarity.[6]

4. Likelihood Ratio Measurement

As indicated briefly in Section 1, signal processing and measurement both consist in operating upon the acoustic field in order to make a decision concerning the properties of that field. The distinction between them may be made on the basis of typical decisions required. Decisions required in measurement are values of the properties of the field; signal processing decisions are, typically, the presence or absence of a source, the speed of the source, the classification of the source, or other target characteristics. In fact, signal processing appears at most to be one degree removed from measurement, i.e. in measuring the field, then ascribing target presence or absence, etc. to the measurement result. From this point of view there is no essential difference between processing and measurement except that the former can give a rationale for selecting field properties to be measured. In this Section we are concerned, therefore, with the development of efficient measurement procedure and will restrict attention to the binary case, since, conceptually, this is all that is necessary to develop more complex measurement or processing operations.

Measurement, being an operation performed upon the acoustic field, implies an interaction of the measurement apparatus with the pressure field itself. While a good deal of attention has been devoted to this point,[7, 8] such interaction is not treated explicitly in the present development, but instead the assumption is made that the acoustic pressure at a point in space may be exactly reproduced as an electrical voltage or the like in the measurement apparatus. Some justification for this lies in the fact that attention here is being focused upon the wave field itself, and the macroscopic limits of finite time, volume and frequency band upon the measurement accuracy are the primary topic. Additionally, interaction effects may often be expressed in terms of a probabilistic state of knowledge and, if so, fall naturally into the present development.

In keeping with the point of view of signal processing as a measurement operation, the criterion of optimum processing is that the average total error of measurement be minimized. For probabilistic wave fields, this criterion is met with a likelihood ratio processor, that is, a processor whose design is the likelihood ratio itself. Beginning with the work of

R. A. Fisher[9] and more recently Wald,[10] Grenander,[11] Neyman and Pearson,[12] Middleton and VanMeter[13] and Peterson, Birdsall and Fox,[14] measurement has been recognized as a procedure of statistical hypothesis testing or, more generally, of statistical decision theory. Their work is here extended to probabilistic wave fields, in which the best measurement operation to perform upon them in space and time is given by the space–time likelihood ratio(s), formed from the space–time probability density functions describing the wave field.

4.1. *Binary Likelihood Ratio*

The binary likelihood ratio is defined as follows: Given one of two possible sets of boundary conditions, C or D, with corresponding probability density functions $f_C(X)$ and $f_D(X)$, both in either space (6) or space–time (7), and given data set X, then the space or space–time likelihood ratio $l_{CD}(X)$ is given by:

$$l_{CD}(X) = \frac{f_C(X)}{f_D(X)} \qquad \text{......(16)}$$

$l_{CD}(X)$ is the description of the optimum measurement or processing operation on X to decide C or D—that is, the ratio of the probability density functions describing the wave field determines the optimum processor. $l_{CD}(X)$ has other properties closely related to the measurement error characteristic (m.e.c.) (Fig. 1), the most direct being that if X_0 is a particular set of data used as decision thresholds for the processor output, then $l_{CD}(X_0)$ is the negative of the m.e.c. slope at the point on the m.e.c. fixed by the error probabilities determined using X_0 as the threshold set.

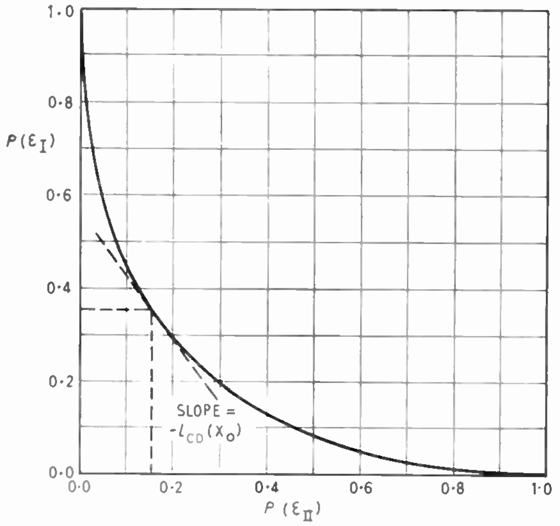


Fig. 1. Typical measurement error characteristic, showing relation between error probabilities and value of likelihood ratio at threshold giving those probabilities.

Likelihood ratio analysis is most readily achieved for a binary decision—that is, the experimenter knows *a priori* that one of two situations is true and measures to find which one is true. Many measurement problems, however, are concerned with higher-order decisions—that is, the problem is to find which of n possible situations is true, where n may be a very large number. It has been shown by Middleton[15] that if the n situations are mutually exclusive—for example, if the set of properties being measured can have only one set of values during the experiment—then higher-order decisions can be analysed as a set of binary decisions, and the likelihood ratio procedure constructed as a set of binary likelihood ratio procedures. This approach has the advantage of confining the basic analysis to a tractable level and of synthesizing the higher order procedure in relatively simple steps.

4.2. *Average Total Probability of Error*

Two types of binary hypotheses can be considered, the *simple*, in which there are only two alternatives to select from, and the *composite*, in which one alternative is to be decided against a number of other possible alternatives. In making a decision on a binary hypothesis of either type, two kinds of errors can be made. If B_0 is the alternative to be tested, and B_i, $i \neq 0$, are all the other possible alternatives, then the experimenter can decide B_i when in fact B_0 is true (Type I error) or he can decide B_0 when in fact B_i is true (Type II error). Or he can make one of two correct decisions. If $P(\epsilon_I)$ denotes the probability of committing a Type I error, and $P(\epsilon_{II})$ denotes the probability of committing a Type II error, then the average total error, $P(\epsilon)$, is given by

$$\left.\begin{aligned} P(\epsilon) &= P(B_0)\,P(\epsilon_I) + \sum_i P(B_i)\,P_i(\epsilon_{II}) \\ P(\epsilon_{II}) &= \Big[\sum_i P(B_i)\,P_i(\epsilon_{II})\Big]\Big/\big[1 - P(B_0)\big] \end{aligned}\right\} \qquad \text{......(17)}$$

where $P(B_0)$ = *a priori* (before the experiment) probability of occurrence of B_0

$P(B_i)$ = *a priori* probability of occurrence of B_i $(i \neq 0)$

and

$$P(\epsilon_I) = \int_R f_{B0}(X)\,\mathrm{d}X \qquad \text{......(18)}$$

$$P_i(\epsilon_{II}) = \int_A f_{Bi}(X)\,\mathrm{d}X \qquad \text{......(19)}$$

where A = acceptance region of X (decide B_0 when X falls in A)

R = rejection region of X (decide 'not-B_0' when X falls in R).

Specification of A and R means dividing up the space of all X into two regions. Since A can be made larger only by making R smaller, and vice versa, then generally, as $P(\epsilon_{\mathrm{I}})$ is decreased, $P(\epsilon_{\mathrm{II}})$ is increased for fixed experimental conditions. A plot of $P(\epsilon_{\mathrm{I}})$ versus $P(\epsilon_{\mathrm{II}})$ is called a measurement error characteristic, following the receiver operating characteristic of Ref. 14. A typical m.e.c. is shown in Fig. 1. As the boundary conditions are changed—for example, the noise energy in the field—a family of m.e.c.'s can be generated.

The average total error criterion, originally suggested by Siegart[16], is not the only criterion for which the likelihood ratio processor is optimum. It generally minimizes average risk,[13, 14] and includes the Neyman–Pearson criterion and Woodward's *a posteriori* observer criterion.[14]

4.3. *Separation of Time and Space Measurements*

From eqn. (16), the space and space–time likelihood ratios may be written down explicitly as follows:

$$l_{CD}({}_1X) = \frac{\mathrm{f}_C({}_1x_1,\ {}_1x_2,\ \ldots,\ {}_1x_M)}{\mathrm{f}_D({}_1x_1,\ {}_1x_2,\ \ldots,\ {}_1x_M)} \qquad \ldots\ldots(20)$$

where $l_{CD}({}_1X)$ is the spatial likelihood ratio, and

$$l_{CD}(X) = \frac{\begin{matrix}\mathrm{f}_C({}_1x_1,\ {}_2x_1,\ \ldots, \\ {}_{2WT}x_1,\ {}_1x_2,\ \ldots\ {}_{2WT}x_2, \\ \ldots\ {}_1x_M,\ \ldots\ {}_{2WT}x_M)\end{matrix}}{\begin{matrix}\mathrm{f}_D({}_1x_1,\ {}_2x_1,\ \ldots, \\ {}_{2WT}x_1,\ {}_1x_2,\ \ldots,\ {}_{2WT}x_2, \\ \ldots\ {}_1x_M,\ \ldots\ {}_{2WT}x_M)\end{matrix}} \qquad \ldots\ldots(21)$$

where $l_{CD}(X)$ is the space–time likelihood ratio.

From Theorem I, measurement or processing of a single frequency field is primarily a matter of space measurement and so (20), as a complete description of the measurement operation, applies chiefly to such fields. Equation (21) applies primarily to series band-limited, time-limited acoustic fields, in which, from Theorem II, measurement must occur in both space and time. From the discussion of Section 3.2, $\mathrm{f}_C(X)$ and $\mathrm{f}_D(X)$ may be decomposed into a product of lower order density functions, either by sets of time samples at one spatial sampling point (eqn. (9)) or by sets of spatial samples at one time sampling point (eqn. (11)). From this discussion, eqn. (21) may be rewritten in either of two ways:

$$l_{CD}(X) = \frac{{}_1\mathrm{f}_C(X_1) \prod_{j=2}^{M} {}_j\mathrm{f}_{C,\ X_1,\ \ldots\ X_{j-1}}(X_j)}{{}_1\mathrm{f}_D(X_1) \prod_{j=2}^{M} {}_j\mathrm{f}_{D,\ X_1,\ \ldots\ X_{j-1}}(X_j)} \qquad \ldots\ldots(22)$$

$$l_{CD}(X) = \frac{{}_1\mathrm{f}_C({}_1Y) \prod_{i=2}^{2WT} {}_i\mathrm{f}_{C,\ {}_1Y,\ \ldots,\ {}_{i-1}Y}\,({}_iY)}{{}_1\mathrm{f}_D({}_1Y) \prod_{i=2}^{2WT} {}_i\mathrm{f}_{D,\ {}_1Y,\ \ldots,\ {}_{i-1}Y}\,({}_iY)} \qquad \ldots(23)$$

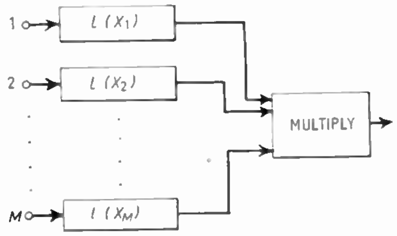


Fig. 2. Likelihood ratio for space stationarity. (Spatial independence illustrated)

Considering (22), if the lower order density functions making up $\mathrm{f}_C(X)$ *and* $\mathrm{f}_D(X)$ according to (9) are locally statistically stationary in space as defined in (12), then $l_{CD}(X)$ can be expressed as the product of locally statistically stationary (in space) likelihood ratios of lower order, each such ratio describing a time measurement. This is indicated in Fig. 2 in block diagram form, and in the following equation:

$$l_{CD}(X) = l_{CD}(X_1) \prod_{j=2}^{M} l_{CD,\ X_1,\ \ldots,\ X_{j-1}}\,(X_j) \quad \ldots\ldots(24)$$

If $\mathrm{f}_C(X)$ and $\mathrm{f}_D(X)$ are both made up of lower order density functions which are locally statistically stationary in space according to (12), then spatial and temporal measurement procedures may be separated, with the temporal measurement operation being performed first. It should be noted that this is a *sufficient*, but not a *necessary* condition for space and time operation separation; for example, $\mathrm{f}_C(X)$ and $\mathrm{f}_D(X)$ could both be made up of component density functions which are non-stationary in space in exactly the same way, in which case the same result would follow.

Similarly, considering (13) and (23), if $\mathrm{f}_C(X)$ and $\mathrm{f}_D(X)$ are both made up of lower order density functions which are locally statistically stationary in time according to (13), then spatial and temporal measurement operations may also be separated, but with the spatial operation being performed first. This is illustrated in Fig. 3, and the consequent space–time likelihood ratio is given in the following equation:

$$l_{CD}(X) = l_{CD}({}_1Y) \prod_{j=2}^{2WT} l_{CD,\ {}_1Y,\ \ldots,\ {}_{i-1}Y}\,({}_iY) \quad \ldots(25)$$

As in (24), it should be noted that local stationarity is a sufficient, but not a necessary condition.

Separation of space and time measurement operations, then, can be specified on the basis of local

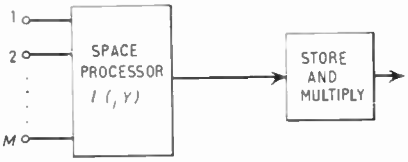


Fig. 3. Likelihood ratio for time stationarity. (Time independence illustrated)

stationarity in either time or space. This should be contrasted with the basis of sample independence for time and space separation. If each space and time sample is statistically independent of all others, but no requirement for local stationarity is made, then, from (9) or (11), the space–time likelihood ratio becomes simply the product of $2WTM$ likelihood ratios, each of first order and each possibly different from all the rest. It may or may not be possible to simplify the space–time measurement operation by some sort of separation. If, however, both space and time local stationarity hold *and* the samples are statistically independent in both space and time, a particularly simple form of the space–time likelihood ratio results:

$$l_{CD}(X) = \prod_{j=1}^{M} l_{CD}(X_j) = \prod_{i=1}^{2WT} l_{CD}({}_iY) \qquad \ldots\ldots(26)$$

This last equation is the basis for the space–time example worked in the following section.

4.4. *Likelihood Ratio Measurement Procedure*

The likelihood ratio measurement procedure may be written down as a sequence of six steps:

Step 1: State each possible set of experimental conditions, that is, each possible set of values of the boundary conditions. Each possible set constitutes a hypothesis to be tested against all the other hypotheses. If the number of possible sets is finite—say, N—then N statements are made, each sufficient to define the resultant probabilistic wave field. If N is infinite—for example, if the amplitude of the signal field can be anything within two limits, or if the signal wavefront can come from any direction—then an iteration procedure may be employed. In such a procedure, finite separation between values of the continuous field parameter are assumed, the remaining steps completed, then the resultant errors examined to find the smallest interval in the continuous parameter which is acceptable in terms of allowable error and available V, T and W.

Step 2: Set down for each statement the consequent probabilistic wave field. If A, B, . . . represent the statements of experimental conditions, and X represents the set of sampled pressures, then each possible probabilistic wave field is represented by $f_A(X)$, $f_B(X)$ . . .

Step 3: Form all pertinent likelihood ratios. If there are N probabilistic wave fields, then there are $(N - 1)$ pertinent likelihood ratios:

$$l_{12}(X),\ l_{23}(X),\ \ldots\ ,\ l_{(N-1),\ N}(X)$$

It should be noted that any of the $N(N - 1)$ possible likelihood ratios can be formed from the $(N - 1)$ pertinent ratios.

Step 4: Construct the likelihood ratio measurement operator. If N is 2, then the operator is just the resultant single likelihood ratio. If N is greater than 2, then the pertinent ratios must be combined as illustrated in Fig. 4 for $N = 3$, or, more generally as indicated by Middleton.[15]

Step 5: From the probabilistic wave fields of Step 2 and the thresholds set by the likelihood ratios of Step 3, calculate the error probabilities, and from these construct the measurement error characteristics. By examining the error probabilities, see whether the m.e.c. family can be characterized by a single number, d, and if so, find what d is in terms of V, T, W and A, B, . . .

Step 6: For given V, T and W find the minimum average total error or, for maximum acceptable average total error, find minimum values of V, T and W. This is the endpoint of the likelihood ratio measurement analysis—to find the limit on error imposed by finite V, T and W and the measurement operation which realizes this limit.

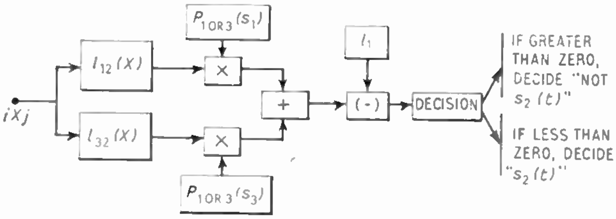


Fig. 4. Likelihood ratio processing to decide that $s_2(t)$ is or is not present, if $s_2(t)$ is one of three possible signals.

5. Examples

5.1. *Resolution in One Dimension*

In order to illustrate the likelihood ratio procedure for a composite binary hypothesis, the problem of deciding whether a Gaussian noise signal $s_2(t)$ is present will be considered. It is assumed that $s_2(t)$ is one of three possible Gaussian noise signals and that Gaussian interfering noise is also present. The single dimension, time, is used to simplify the analysis. All three signals and noise are assumed to have zero mean (e.g. the static pressure is not considered) and variances at any of the $2WT$ sample points of S_1, S_2, S_3 and N, respectively, with $S_3 > S_2 > S_1$. The $2WT$ samples are assumed independent, so the consequent probability density functions, of order $2WT$, are:

$$f_1(X) = \{2\pi(S_1 + N)\}^{-WT}$$

$$\exp\left\{-\frac{1}{2(S_1 + N)} \sum_{i=1}^{2WT} {}_ix^2\right\} \qquad \ldots\ldots(27)$$

$$\mathrm{f}_2(X) = \{2\pi(S_2+N)\}^{-WT}\exp\left\{-\frac{1}{2(S_2+N)}\sum_{i=1}^{2WT}{}_ix^2\right\} \quad\ldots\ldots(28)$$

$$\mathrm{f}_3(X) = \{2\pi(S_3+N)\}^{-WT}\exp\left\{-\frac{1}{2(S_3+N)}\sum_{i=1}^{2WT}{}_ix^2\right\} \quad\ldots\ldots(29)$$

where $S_2 - S_1 = \delta S = S_3 - S_2$

and $\quad \delta S \ll S_1 + N,\ S_2 + N,\ S_3 + N$

The decision to be made is whether or not $s_2(t)$ is present. There are two pertinent likelihood ratios, $l_{12}(X)$ and $l_{32}(X)$,

$$l_{12}(X) = \left(\frac{S_2+N}{S_1+N}\right)^{WT}\exp\left[\frac{-\delta S\sum_{i=1}^{2WT}{}_ix^2}{2(S_1+N)(S_2+N)}\right] \quad\ldots\ldots(30)$$

$$l_{32}(X) = \left(\frac{S_2+N}{S_3+N}\right)^{WT}\exp\left[\frac{\delta S\sum_{i=1}^{2WT}{}_ix^2}{2(S_3+N)(S_2+N)}\right] \quad\ldots\ldots(31)$$

The likelihood ratio for deciding whether $s_2(t)$ is or is not present would be $l_{2,(1\text{ or }3)}(X)$, but it is more convenient to consider the reciprocal of this ratio, with the thought that the acceptance region for $s_2(t)$ will be defined for a set of values of the reciprocal likelihood ratio *less* than a given threshold quantity. The reciprocal likelihood ratio, $l_{(1\text{ or }3),2}(X)$ is given by:

$$\begin{aligned} l_{(1\text{ or }3),2}(X) &= P_{(1\text{ or }3)}(S_1)\,l_{12}(X) + P_{(1\text{ or }3)}(S_3)\,l_{32}(X) \\ &= \tfrac{1}{2}\left(1-\frac{\delta S}{S_2+N}\right)^{-WT}\exp\left\{-\frac{WT(\delta S)\,\overline{x^2}}{(S_2+N)^2}\right\} + \\ &\quad + \tfrac{1}{2}\left(1+\frac{\delta S}{S_2+N}\right)^{-WT}\exp\left\{\frac{WT(\delta S)\,\overline{x^2}}{(S_2+N)^2}\right\} \quad\ldots\ldots(32) \end{aligned}$$

$$\overline{x^2} = (2WT)^{-1}\sum_{i=1}^{2WT}{}_ix^2,$$

$$P_{(1\text{ or }3)}(S_1) = \tfrac{1}{2} = P_{(1\text{ or }3)}(S_3)$$

The minimum value of (32) may readily be shown to be

$$l_{\min} = \left[1-\left(\frac{\delta S}{S_2+N}\right)^2\right]^{-WT/2} \quad\ldots\ldots(33)$$

and thus will be unity only if $\delta S = 0$, that is, if all three signals are identical. For $\delta S \neq 0$, the minimum threshold may be found from (33) and will occur at $\overline{x^2}]_{\mathrm{TH}}$

$$\overline{x^2}]_{\mathrm{TH}} = S_2 + N \quad\ldots\ldots(34)$$

If a value of $l_{(1\text{ or }3),2}(X) > l_{\min}$ is selected, in general two values of $\overline{x^2}$ will satisfy

$$l_{(1\text{ or }3),2}(X) = l_1 > l_{\min} \quad\ldots\ldots(35)$$

and are given by

$$\cosh\left\{\frac{WT(\delta S)\,(\overline{x^2})_1}{(S_2+N)^2}\right\} = \frac{\begin{aligned}&[\mathrm{e}^{aWT}+\mathrm{e}^{bWT}] + \\ &+ [\mathrm{e}^{aWT}-\mathrm{e}^{bWT}]\sqrt{l_1^2-\mathrm{e}^{(a+b)WT}}\end{aligned}}{2\,\mathrm{e}^{(a+b)WT}} \quad\ldots\ldots(36)$$

and

$$\cosh\left\{\frac{WT(\delta S)\,(\overline{x^2})_2}{(S_2+N)^2}\right\} = \frac{\begin{aligned}&[\mathrm{e}^{aWT}+\mathrm{e}^{bWT}] - \\ &- [\mathrm{e}^{aWT}-\mathrm{e}^{bWT}]\sqrt{l_1^2-\mathrm{e}^{(a+b)WT}}\end{aligned}}{2\,\mathrm{e}^{(a+b)WT}} \quad\ldots\ldots(37)$$

where

$$\mathrm{e}^a = \left(1-\frac{\delta S}{S_2+N}\right)^{-1},\quad \mathrm{e}^b = \left(1+\frac{\delta S}{S_2+N}\right)^{-1}$$

Since both $(\overline{x^2})_2$ decreases and $(\overline{x^2})_2$ must exceed or equal zero, as l_1 is increased, $(\overline{x^2})_2$ decreases until it equals zero, then stays constant.

The relation between the likelihood threshold l_1 and the quantity y^2:

$$\left.\begin{aligned}(S_2+N)\,y^2 &= (\overline{x^2}) \\ (S_2+N)\,y_1^2 &= (\overline{x^2})_1 \\ (S_2+N)\,y_2^2 &= (\overline{x^2})_2\end{aligned}\right\} \quad\ldots\ldots(38)$$

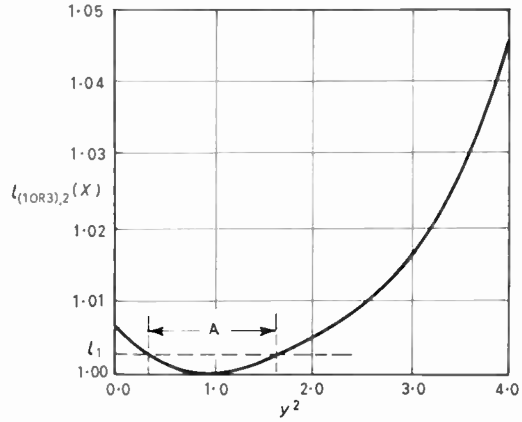


Fig. 5. Likelihood ratio (eqn. (32)) versus y^2 (eqn. (38)) showing acceptance region A for $s_2(t)$ for threshold l_1.

is shown in Fig. 5. The likelihood ratio operator is simply a mean square pressure measurement, shown in block diagram form in Fig. 4.

The average total error is given by the expression

$$P(\epsilon) \cong \frac{2}{3(2\pi)^{1/2}} \left[\int\limits_{2\sqrt{WT}[y_1-1]}^{\infty} \exp\left\{-\frac{u^2}{2}\right\} \mathrm{d}u + \right.$$

$$\left. + \int\limits_{2\sqrt{WT}\,y_2}^{2\sqrt{WT}\,y_1} \exp\left\{-\frac{u^2}{2}\right\} \mathrm{d}u \right] \quad \ldots\ldots(39)$$

derived from eqns. (17) and (27)–(29) with the assumption that $2WT \gg 1$ so that the χ^2 distribution for y^2 resulting from (27)–(29) may be approximated by the appropriate Gaussian distribution.[17] From (39) it may readily be shown that if the signals are equally likely *a priori*, then minimum $P(\epsilon)$ occurs when the thresholds for $\overline{x^2}$ are set halfway between $S_1 + N$ and $S_2 + N$ and halfway between $S_2 + N$ and $S_3 + N$.

Formally, the signals $s_1(t)$ and $s_3(t)$ may be considered in this example as either two separate signals or one composite signal. In the latter case, this example serves to lend some precision to the specification of what is meant by resolution, although consideration of other values of S_1 and S_3 may be more appropriate to other familiar types of resolution problems.

5.2. *Space–Time Signal Known Exactly in Gaussian Noise*

The processing problem here is to decide whether a signal whose wave*form* and wave*front* are known exactly (*a priori*) by the experimenter is or is not present in a background of white Gaussian noise of band (0, W). This is a simple binary hypothesis, and the two probability density functions describing the probabilistic wave fields corresponding to noise alone and signal plus noise are:

$$\mathrm{f}_N(X) = (2\pi N)^{-MWT} \exp\left\{-\frac{1}{2N} \sum_{i=1}^{2WT} \sum_{j=1}^{M} {}_ix_j^2\right\} \quad \ldots\ldots(40)$$

$$\mathrm{f}_{SN}(X) = (2\pi N)^{-MWT} \exp\left\{-\frac{1}{2N} \sum_{i=1}^{2WT} \sum_{j=1}^{M} ({}_ix_j - {}_is_j)^2\right\} \quad \ldots\ldots(41)$$

and the samples are assumed statistically independent from point to point in time and from point to point in space. The space–time likelihood ratio is:

$$l_{SN}(X) = \exp\left\{-MWT\left(\frac{S}{N}\right) + \frac{1}{N} \sum_{i=1}^{2WT} \sum_{j=1}^{M} {}_ix_j\,{}_is_j\right\} \quad \ldots\ldots(42)$$

where ${}_ix_j$ = acoustic pressure amplitude at the ith time sampling point at the jth space sampling point

${}_is_j$ = signal pressure amplitude at ith point in time and jth sampling point in space

$$S = (2MWT)^{-1} \sum_{i}^{2WT} \sum_{j}^{M} {}_is_j^2.$$

The likelihood ratio measurement operation (42) can be considered in two steps. For the ith instant of time, the data ${}_ix_j$ are multiplied by the respective values ${}_is_j$. Now, if the signal is a travelling wave, this multiplication compensates or delays the inputs to match the wavefront of the signal at the ith instant, the delayed inputs then being added up. This is one way of describing pattern formation (in which identical wavefront matching is usually done over the entire measurement period—the delay for the jth space sampling point, in other words, is assumed independent of time). The space-summed set

$$\sum_{j=1}^{M} {}_ix_j\,{}_is_j$$

is then summed over time, which corresponds to filtering the data $x(t)$ in time with a matched filter in the sense of North[18] and VanVleck and Middleton.[19] This matched filtering is identical to cross-correlating the input data with the signal waveform, but illustrates the possibility that the matched filter for any space sampling point j may change with j. To make the time processing clearer, the order of space and time summation may be reversed (see Section 4.3), in which case matched filtering takes place at each spatial sampling point output, followed by pattern summation. In Fig. 6, both orders of processing are shown. The average total error probability may be found using the m.e.c. curves given in Fig. 7, with the m.e.c. parameter d given by

$$d = 2WTM\left(\frac{S}{N}\right) \quad \ldots\ldots(43)$$

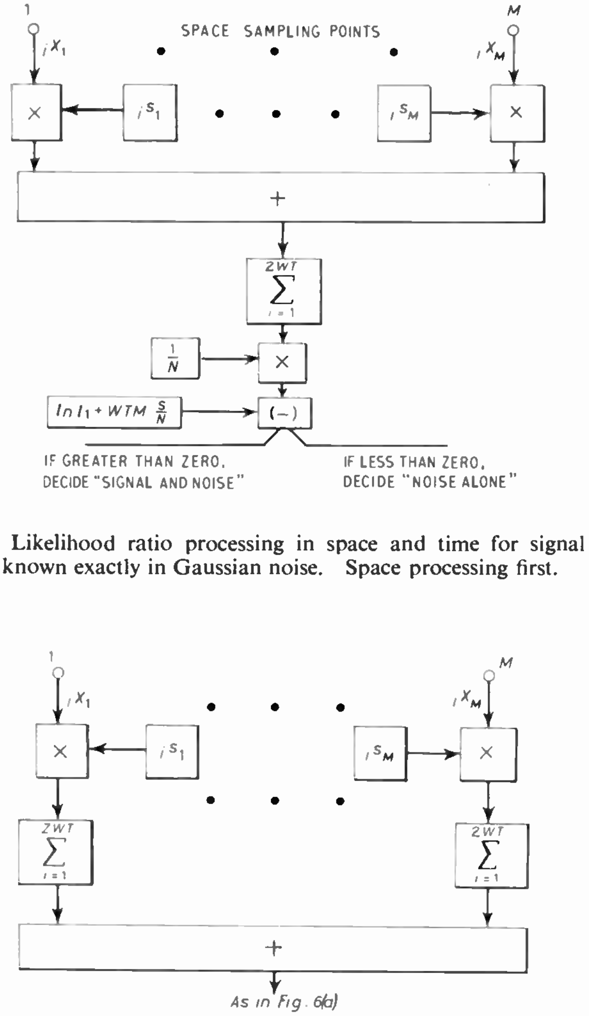
In (43) the number of space sampling points, M, enters in exactly the same manner as the number of time sampling points, $2WT$, due to the equivalence of the space and time knowledge of both signal and noise. Relation of eqn. (43) to more familiar terms may be made by taking 10 $\log_{10}$ of both sides:

$$10 \log_{10} d = 10 \log_{10} (2WT) +$$

$$+ 10 \log_{10} M + 10 \log_{10} \left(\frac{S}{N}\right) \quad \ldots\ldots(44)$$

In eqn. (44), 10 $\log_{10} d$ is a constant for a constant error performance—that is, $P(\epsilon_{\mathrm{I}})$ and $P(\epsilon_{\mathrm{II}})$, 10 $\log_{10}$

$(2WT)$ is the well-known improvement in input signal/noise ratio for matched filtering in the presence of white Gaussian noise; $10 \log_{10} (M)$ is the directivity index of the array of M elements, perfectly matched to the signal wavefront in independent (from point to point in space) Gaussian noise; $10 \log_{10} (S/N)$ is the input signal/noise ratio in decibels.



(*a*) Likelihood ratio processing in space and time for signal known exactly in Gaussian noise. Space processing first.

(*b*) Likelihood ratio processing in space and time for signal known exactly in Gaussian noise. Time processing first.

Fig. 6. Two equivalent space–time likelihood ratio processing procedures for signal known exactly in Gaussian noise.

Several straightforward extensions can be made on this example. Consider the case of one of two equally likely signals being present in Gaussian noise, for which the two probability density functions in space–time are:

$$f_1(X) = (2\pi N)^{-MWT} \exp\left\{-\frac{1}{2N}\sum_{i=1}^{2WT}\sum_{j=1}^{M}({}_ix_j - {}_is_{j1})^2\right\} \qquad \ldots\ldots(45)$$

$$f_2(X) = (2\pi N)^{-MWT} \exp\left\{-\frac{1}{2N}\sum_{i=1}^{2WT}\sum_{j=1}^{M}({}_ix_j - {}_is_{j2})^2\right\} \qquad \ldots\ldots(46)$$

The space–time likelihood ratio for this case is:

$$l_{12}(X) = \exp\left\{- MWT(S_1 - S_2) + \right.$$
$$\left. + \frac{1}{N}\sum_{i=1}^{2WT}\sum_{j=1}^{M} {}_ix_j({}_is_{j1} - {}_is_{j2})\right\} \qquad \ldots\ldots(47)$$

$$S_1 = (2\ WTM)^{-1}\sum_{i=1}^{2WT}\sum_{j=1}^{M} {}_is_{j1}^2,$$

$$S_2 = (2\ WTM)^{-1}\sum_{i=1}^{2WT}\sum_{j=1}^{M} {}_is_{j2}^2.$$

Here, instead of the space–time cross-correlation of the previous case, the input is cross-correlated with the difference in the two signals. From the point of view of space processing first, the data ${}_ix_j$ are delayed by the difference between the delays for ${}_is_{j1}$ and ${}_is_{j2}$. Otherwise, the procedure is identical to the previous case. With the signal mean square values denoted by S_1 and S_2, then the m.e.c. curves for this case are

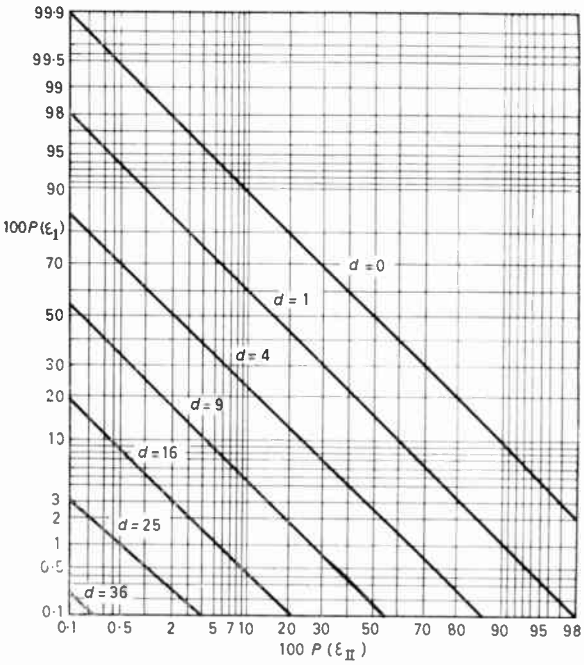


Fig. 7. The measurement error characteristics (adapted from Reference 14, Figure 3).

given by those of Fig. 7, with

$$d = 2\,WTM\,\frac{(S_1 + S_2 - 2\,\rho\sqrt{S_1 S_2})}{N} \qquad \ldots\ldots(48)$$

where $\rho = \dfrac{1}{2\,WTM}\sum_{i=1}^{2WT}\sum_{j=1}^{M} {}_is_{j1}\,{}_is_{j2}/\sqrt{S_1 S_2}$

$=$ normalized cross-correlation in space and time between s_1 and s_2.

To maximize d (and so minimize error probabilities), ρ, essentially the average product of s_1 and s_2 over space and time, should be made as small as possible. If the two signals have identical wavefronts, this means making them orthogonal in time—that is, their waveforms should be uncorrelated. If the two signals differ only in direction, then a rough argument for making the array as large as possible—that is, narrowing the beamwidth (for plane wavefronts) an appropriate amount, may be made as follows. s_1 and s_2 are assumed to be band-limited signals differing only in direction. Then to decrease ρ, one generally increases the *average* delay between ${}_is_{j1}$ and ${}_is_{j2}$—that is, the difference in delays required to match ${}_ix_j$ to the s_1 wavefront and to the s_2 wavefront—as much as possible. The delay for each is taken with respect to a reference space sampling point, and as the distance from this reference point to any other point increases, the difference in relative delay increases. Thus, the average difference in delay increases as M, if M is increased by adding space sampling points so as to increase the size of the array. This argument is more basic than one based on pattern formation, but in terms of the latter, minimum ρ corresponds to having an array such that one signal is on the major lobe, and the second is in deepest (usually the first) null.

Finally, as shown in Reference 20, the single signal in noise case can be extended to noise which is correlated from hydrophone to hydrophone in space. Considering the spatial processing only, the d equation becomes

$$d_{\text{space only}} = \frac{S}{N}\left[1 + (M-1)\,\frac{(1-\rho_N)}{1+\rho_N}\right] \qquad \ldots\ldots(49)$$

where $\rho_N =$ normalized inter-space sampling point noise correlation.

In eqn. (49), if ρ_N goes to zero, d becomes SM/N as would be expected from the spatial part of (43). As ρ_N goes to unity, corresponding to completely correlated noise, d in (49) approaches S/N, that is, there is no advantage to be gained by spatial processing, since in space the noise looks exactly like the signal.

6. Acknowledgments

The author is grateful for advice and encouragement given him by Professor F. V. Hunt of Harvard University, Dr. H. Wysor Marsh, Mr. Theodore Birdsall and Dr. David Middleton during the progress of the work.

7. References

1. E. T. Whitaker, "On the functions which are represented by the expansions of the interpolation theory", *Proc. Roy. Soc. Edinburgh*, **35**, p. 181, 1915.
2. Stanford Goldman, "Information Theory", Section 2.1, p. 67. (Prentice-Hall, New York, 1953).
3. P. L. Stocklin, "Limits of Measurement of Acoustic Wave Fields", Appendix I, page 78. Doctoral dissertation, University of Connecticut, Storrs, 1962.
4. Stocklin, ref. 3, Appendix II, p. 85.
5. W. P. Davenport, Jr. and W. L. Root, "An Introduction to the Theory of Random Signals and Noise", Section 4–5 p. 60. (McGraw-Hill, New York, 1958).
6. Davenport and Root, ref. 5, Section 6–4, p. 93.
7. Leon Brillouin, "Science and Information Theory", Chapter 8, Sections 6 and 7. (Academic Press, New York, 1956).
8. J. L. Stewart and E. C. Westerfield, "A theory of active sonar detection", *Proc. Inst. Radio Engrs*, **47**, p. 872, May 1959.
9. R. A. Fisher, "On an absolute criterion for fitting frequency curves", *Messenger of Mathematics*, **41**, p. 155, 1912.
10. Abraham Wald, "Sequential Analysis", Chapters 1 and 10. (John Wiley, New York, 1947).
11. U. Grenander, "Stochastic processes and statistical inference", *Ark. Math.*, **1**, p. 195, 1950.
12. J. Neyman and E. S. Pearson, "On the problem of the most efficient test statistical hypothesis", *Trans. Roy. Soc. (London)*, A, **231**, p. 289, February 1933.
13. D. VanMeter and D. Middleton, "Modern statistical approaches to reception in communication theory", *Trans. Inst. Radio Engrs (Information Theory)*, No. PGIT-4, p. 119, September 1954.
14. W. W. Peterson, T. G. Birdsall and W. C. Fox, "The theory of signal detectability", *Trans. Inst. Radio Engrs (Information Theory)*, No. PGIT-4, p. 171, 1954.
15. David Middleton, "Introduction to Statistical Communication Theory", Section 19.1–1. (McGraw-Hill, New York, 1960).
16. J. L. Lawson and G. E. Uhlenbeck, "Threshold Signals" Section 7.5. (McGraw-Hill, New York, 1950).
17. Nils Arley and K. Rander Buch, "Introduction to the Theory of Probability and Statistics", p. 96. (John Wiley, New York, 1950).
18. D. O. North, "Analysis of factors which determine signal/noise discrimination in pulsed carrier systems", R.C.A. Lab. Report PTR–6C June, 1943.
19. J. H. Van Vleck and D. Middleton, "A theoretical comparison of visual, aural and meter reception of pulsed signals in the presence of noise", *J. Appl. Phys.*, **17**, p. 940, November 1946.
20. P. L. Stocklin, ref. 3, Example 5.3, p. 69.

Manuscript received by the Institution on 3rd July 1962 (Paper No. 832/SS17).

POINTS FROM THE DISCUSSION

Dr. E. J. Risness†: The theory of optimum acoustic signal processing based on a decision theory approach is developing rapidly, and within the next few years will undoubtedly increase our understanding of what is possible in the signal processing field—though the main conclusion may well be that optimum methods cannot in most practical cases do significantly better than we are doing now. Meanwhile, more conventional approaches to signal processing, as exemplified in many papers at this Symposium, can continue to yield useful results. It is important, however, that where the so-called optimum approach produces any basic principles or theoretical limitations their significance in more conventional calculations should be fully realized. One example of this is non-uniform sampling in space or time. By sampling a waveform across a finite aperture in space much more closely than $\lambda/2$, and using optimum processing, it is possible in theory to produce a gain and directivity equivalent to that obtained from a much larger aperture if conventional processing is used. As Dr. Vanderkulk's paper‡ shows, this improvement may be very difficult to realize in practice. Nevertheless, it should be borne in mind that so-called Nyquist sampling (in space or time) does not represent a boundary beyond which it is *impossible* to obtain any improvement.

A second sample of a basic principle produced by optimum processing considerations is the well-known one that, if the interfering noise is Gaussian and stationary, then for making decisions on signals, whether detection, location, resolution or any other 'hypothesis testing', whatever the nature of the signal itself, or the spatial and temporal configuration of the data collected, the optimum processor is never more than 'second-order'; i.e. never employs more than one squaring, multiplication or rectification process and sequence on the incoming data.

Dr. Stocklin (*in reply*): Dr. Risness' comment regarding Nyquist sampling in space–time not posing a limit to processing improvement from increased sampling is quite true, but perhaps some further comment may clarify the ground on which, I believe, we both stand.

Space–time sampling, as described in this paper, may be thought of as *uniform* sampling which, for finite space-time volume, permits reconstruction of the space–time field everywhere within the finite volume, with a small (normally) but finite reconstruction error. This, together with a description of the space–time probability density functions, permits an estimate of the number of space-time degrees of freedom available for processing gain. Now, increased sampling within the same volume essentially cuts down on the reconstruction error, increasing the state of knowledge of the field and increasing the processing gain to be realized. Since the errors are small to begin with, however, large increases in the number of samples are necessary for significant gains. Such gains also are quite sensitive to the structure of the space–time probability density functions, in the same way that super-directive array gain is quite sensitive to system bandwidth. Another way to think of the effect of increased sampling is that it is roughly equivalent (in terms of processing gain) to having sampled a larger volume in a uniform manner. From the point of view of the larger volume, only a portion of it has been sampled, and to reconstruct the field in the unsampled portion, sampling functions large in magnitude in the sampled portion are required. Greatly increasing sampling in the sampled portion results in greatly increasing the magnitude of the sampling functions relative to those for uniform sampling. The relative accuracy and stability required to generate these functions is consequently great, which is similar to characteristics of the shading functions required for super-directive arrays.

† Admiralty Underwater Weapons Establishment.

‡ W. Vanderkulk, "Processing for Optimum Acoustic Array Gain". Paper presented at the Symposium on "Sonar Systems", Birmingham, 1962. (To be published in *The Radio and Electronic Engineer.*)

Radio Engineering Overseas . . .

The following abstracts are taken 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. Information on translating services will be found in the Institution publication "Library Services and Technical Information".

PROPAGATION INVESTIGATIONS

Dependable technical solutions have to be found for the increasingly complex problem of wavelength allocation in the medium wave and long wave broadcast bands. One approach lies in obtaining exact knowledge of the indirect waves during the night. A Yugoslav engineer has described the organization and the research methods employed at the Institute M. Pupin, Belgrade, in the period from 1952 up to 1960. The results obtained by measurements of electromagnetic fields and by statistical processing data concerning field-strength variation for varying geophysical, meteorological and other influences facilitate the determination of propagation characteristics of indirect waves.

"Indirect hectometre and kilometre waves of electromagnetic fields", M. Masirevic. *Elektrotekniski Vestnik* (*Ljubljana*), **30**, pp. 10–16, 1962–63.

EFFECT OF SOLAR ECLIPSE ON THE D-LAYER

Measurements were carried out at Bucharest Polytechnic Institute to determine the variations in the reflexion coefficient for medium and long waves which occurred in the ionospheric D layer during the total eclipse of the sun on 15th February, 1961. Conclusions are drawn concerning different phenomena, such as ionization and recombination in the lower layers of the ionosphere under the action of sun beams.

"Results and conclusions of measurements concerning the reflexion coefficient of ionospheric waves carried out during the solar eclipse on 15th February, 1961", C. Serbu. *Telecommunicatii* (*Bucharest*), **6**, pp. 247–56, November-December 1962.

INFRA-RED AND MICROWAVE RADIOMETRY

From consideration of the laws governing the emission, transmission and collection of thermodynamic radiation, the authors of a French paper show that the passive detection of bodies around us is restricted to a very small number of frequency bands, the 35 and 100 Gc/s bands for microwaves and 30 and 80 Tc/s for infra-red. The existence of these favoured bands is the result of compromise between the emissivity of the bodies themselves, the atmosphere's transparent windows and receiver performance. A number of numerical applications show what can be expected from passive detection of an aircraft on the background of the sky, of the ground seen from the aircraft, of the horizon, the Sun and the Moon.

"Comparison of the relative values of infra-red and radio microwave interferometry", G. Broussaud and B. Richard. *Annales de Radioéléctricité*, No. 72, pp. 89–111, April 1963.

TRANSMISSION OF ANGULAR INFORMATION

An arrangement for transmitting the angle of orientation of a surveillance radar is described in a French paper. The voltages obtained from the aerial synchro units are converted to variable phase signals which are used to modulate sub-carriers in amplitude. These sub-carriers are transmitted over a single telephone channel. At the receiving end, after the channels have been separated and the signals detected, these variable phase signals are used to actuate a control for copying the aerial motion. Stress is laid on difficulties encountered when very high dynamic precision is desired.

"Angle retransmission system", M. Poliet and J. Girault. *Annales de Radioéléctricité*, No. 72, pp. 130–52, April 1963.

VISUAL FLIGHT SIMULATION

A visual simulator has been developed in Canada that gives a pilot under training the same view that he would have in an aircraft in flight. An image projected from a moving photographic slide is scanned by a television camera, then projected on to a screen in front of the trainee. The functional stages of the visual simulator can be defined as terrain storage element, data extraction system, perspective derivation element, and image transportation system. One of the most valuable uses of a visual simulator is to train the pilot to fly in bad weather conditions.

"Pilot's eye-view is projected on screen in visual simulation flight trainer", P. M. Carey. *Canadian Electronics Engineering*, 7, pp. 31–40, February 1963.

U.H.F. PROPAGATION OVER VARIOUS TERRAINS

Great differences have been observed in the received signal in comparative propagation tests using horizontally and vertically polarized waves over level, hilly, open and timbered terrain. These differences are particularly striking in forest areas when a change is made between vertical and horizontal polarization, the received signal for vertical polarization being smaller than for horizontal polarization. A number of tests carried out in Germany under various topographical conditions with different measuring methods revealed that the difference in the propagation of vertically and horizontally polarized waves is due to an 'extinction' of the vertically polarized waves by the vegetation growing on the terrain.

"Propagation of horizontal and vertical polarization ultra short waves over various terrains for low aerial heights", H. U. Widdel. *Archiv der Elektrischen Ubertragung*, **17**, pp. 145–50, March 1963.