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 42 No. 2 FEBRUARY 1972

THE RADIO AND ELECTRONIC ENGINEER

The Journal of the Institution of Electronic and Radio Engineers

Time for Application

AT a recent colloquium* the question was raised as to the cause of delay between new ideas being proved in the laboratory and the availability of commercial equipments. It was suggested that some kind of 'middleman' may be needed to bridge the gap between R & D and the market. It seems to the writer that the gap between the inventor and the user is often filled by people who have as their aim in life the greatest obstruction to new ideas and who believe that when their brief is carefully to examine new proposals it means that the maximum possible criticisms must be raised.

About twenty five years ago, when Industry became aware of its serious lack of good scientific brains in responsible positions, it recruited many such brains, nurtured on the scientific war effort, and placed them in charge of R & D. There have been many notable successes but in general the function that should have been outlined was 'Application', not 'Research and Development'. In the great majority of cases very little 'R' and not much 'D' is required. Ideas, new techniques and versatile electronic devices quite often need only enlightened application to bring early benefit to the user.

The R & D manager, however, is trained in Research, not Application and when he has to evaluate a new product, device or system for his company his main interest is to re-develop, exhaustively test and report upon, rather than to see how quickly some benefits can be obtained. This no doubt accounts for the 'not invented here' syndrome. In the writer's view, although 'Mark I' may not meet the ideal requirements of the R & D department, it may well be adequate for the time being and its initial cost could be written off and a great deal of working experience obtained during the time that would be lost in developing the ideal version. By deploying adequate equipment as soon as it is available the innovator is encouraged and can then proceed with his 'Mark II' development whilst enjoying the benefits of feedback on 'Mark I' performance from the user. (Unfortunately some users think that this part of their activity warrants having a 'Mark I' free of charge and justifies a twelve-months' delay in acceptance whilst it is 'evaluated'!)

There are signs that the situation is improving; commercial and financial pressures are calling for speedy results and some enlightened managements insist that R & D is a means to an end, not an end in itself. Clearly the need to apply our enormous existing fund of scientific and engineering knowledge is urgent. Chartered Engineers, particularly those who have had both academic and industrial experience, are best placed to fill the need of a 'middleman' in the chain:

Invention—Development—Application—Production—Installation.

Perhaps one can sum up the situation and the guiding brief to this middleman by inventing a new Law:

'Be adequate and you can be first. Be best and you will be last'.

J. BILBROUGH

*'Fresh Approaches to Non-destructive Testing'. Organized jointly by the I.E.E., the I.E.R.E. and others, and held in London on 13th December 1971.

Contributors to this Issue

Mr. R. T. Irish (Member 1965, Graduate 1961) was a technical officer with the Mullard Radio Valve Co. from 1954 to 1961, working in the Measurements and Applications Laboratory and later in the Technical Department (Transmitting). During this period he read for his London University B.Sc.(Eng.) degree at the Northampton College of Advanced Technology (now City University). In 1961 Mr. Irish was appointed lecturer in electronics at the Royal Military College of Science; he was promoted to senior lecturer in 1964 and to principal lecturer in 1971. His research interests and publications cover microwave interactions with gaseous plasmas, microwave filters and integrated circuits.

Mr. A. J. Allen graduated in 1950 as an external student of the University of London with first-class honours in electrical engineering. Following an apprenticeship with The General Electric Company he obtained further industrial experience with Joseph Lucas Ltd. He entered education as a lecturer at Birmingham College of Advanced Technology (now Aston University) in 1956 and was appointed senior lecturer in control engineering at Farnborough Technical College in 1960. He began research on digital controllers in 1967 and gained his M.Phil. as an external student of the University of Reading. He was appointed a lecturer in the Applied Physical Sciences Department at Reading in January 1971. Apart from his teaching and research interests he is a consultant to several industrial organizations.

Mr. P. Atkinson (M. 1962) is now a senior lecturer in the Department of Applied Physical Sciences at the University of Reading, which he joined as a lecturer on its foundation in 1964. He has contributed several previous papers to this *Journal* in recent years on control engineering subjects; a fuller note on his career appeared in the July 1969 issue.

Professor H. Sutcliffe (F. 1971) holds the chair in electronics engineering at the University of Salford where he went as reader in 1963. His earlier career was referred to in a note in the September 1970 issue of the *Journal* on the occasion of the publication of a previous paper.

Mr. John J. Morrison was educated at Albert School, Glasgow, and The Royal Technical College (now The University of Strathclyde). Following an apprenticeship in electrical engineering he joined Dobbie McInnes (R. & D.) Ltd., as a design and development engineer and was subsequently Technical Manager of this Company from 1955 until 1958. He then went to Honeywell Controls Ltd., as Chief Design Engineer and in 1963 he was appointed Engineering Manager.

He took up his present position as Technical Director of Cableform Ltd. in 1967 and since then he has been engaged largely in the field of components and systems for electric traction.

Mr. Morrison is currently Chairman of the Technical Committee of the Electric Vehicle Association.

Mr. H. A. Kemhadjian obtained a first degree in electrical engineering at Southampton University in 1956 before working for Mullard Ltd. at the Southampton Semiconductor Development Laboratories. During a period of eight years, he held a number of posts concerned with applications of semiconductor devices, their characterization and electrical development. He joined the Electronics Department of the University of Southampton as a lecturer in 1964. His main research interest is the modelling of semiconductor components for application to integrated circuit design.

Mr. M. A. Flemming obtained a B.A. honours degree in physics at New College, Oxford, in 1969. Since that time he has been pursuing research on high-frequency measurement methods applicable to integrated circuits in the Electronics Department at the University of Southampton.

Professor D. G. Tucker (F. 1953) is head of the Department of Electronic and Electrical Engineering of the University of Birmingham, the post he has held since 1955. In addition to research interests in the fields of underwater acoustics and circuit techniques, in both of which he has contributed numerous papers to this and other journals and has written several text books, Professor Tucker is keenly interested in the history of technology: this is his fourth paper in this area to be published in *The Radio and Electronic Engineer*. A note on his earlier career was published in July 1970.

U.D.C. 681.323 : 66/67.02

A Special-purpose Computer for the Direct Digital Control of Processes

A. J. ALLEN,
B.Sc.(Eng.), M.Phil., C.Eng., M.I.E.E.*

and

P. ATKINSON,
B.Sc.(Eng.), A.C.G.I., C.Eng., M.I.E.E., M.I.E.R.E.*

Presented at meetings of the West Midlands Section in Wolverhampton on 6th April 1971 and the Thames Valley Section in Reading on 29th April 1971.

SUMMARY

A general-purpose digital computer for the on-line control of large scale industrial processes serves the dual purposes of a time-shared controller for each loop by means of programmed algorithms and of a data logger and alarm scanner. The architecture of the general-purpose computer has been criticized regarding its suitability as a time-shared controller. This paper considers this problem and presents a case for the special-purpose digital controller. A research model is described in detail and its performance in a simulated system is given.

* *Department of Applied Physical Sciences, University of Reading, Reading RG6 2AL.*

1. Introduction

1.1. *Feedback Control of Processes*

The control of process plants has been achieved conventionally by use of analogue pneumatic and, at present to a lesser extent, by analogue electronic controllers. These are incorporated into a feedback arrangement[1] as shown in Fig. 1. The object of any process control system is to force the *actual value* of the process output to follow the *desired value* in spite of the presence of extraneous *disturbances*. Often the only means of influencing the actual value of the process is by means of a *controlled variable* forming the process input. In the absence of disturbances and indeterminate fluctuations in the process parameters it is possible to calculate the value of the controlled variable to give a particular actual value of the process (so long as the process parameters are known). Such a control strategy is known as 'open-loop control' and is rarely used in practice because the existence of disturbances and parameter fluctuations is inevitable. The effect of disturbances and parameter fluctuations can be reduced by the use of feedback. In a feedback system the actual value of the process output which is to be controlled to some desired value is converted into a more convenient *measured value* by the detecting unit.

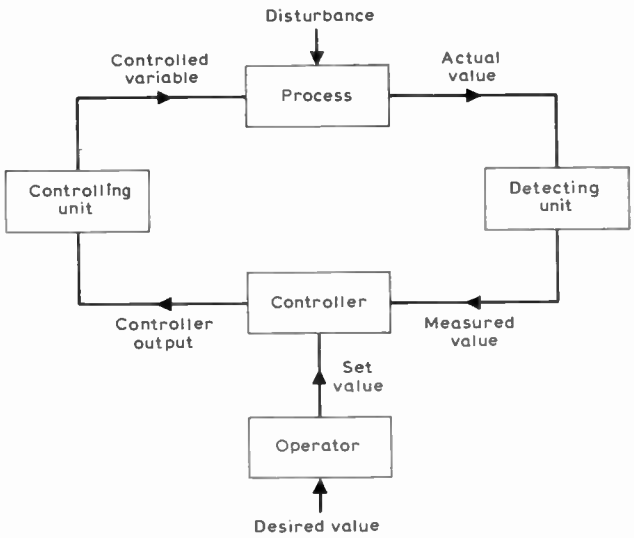


Fig. 1. Block diagram of a process control system.

It is usual to feed the controlling unit from a controller which receives the measured value and *set value* signals and within which the deviation is formed. The *deviation* is defined as the difference between the measured value and the set value. Many control engineers prefer to use the concept of an *error signal* which is defined as 'minus deviation'.

In its simplest form, the controller produces an output proportional to the error in such a sense as to reduce the error towards zero. Although the error could be made zero for one set of conditions, any change in set value or disturbance will result in a non-zero error if the loop-gain is finite. If the characteristics are sensibly constant a closer agreement between actual value and

desired value can sometimes be obtained by making the set value input to the controller different from the desired value.

To obtain more accurate control the error should tend to zero for all possible conditions of the plant, and increasing the gain of the loop does reduce the steady-state error. However, when a certain critical gain (G_c) is used it is found that any disturbance causes the complete loop to oscillate continuously[3] at a frequency f_0.

A reasonable compromise between large steady-state errors and very oscillatory transient response is often obtained using a loop gain of $G_c/2$. The required gain setting to achieve this condition could be determined in practice by altering the gain until the loop continuously oscillates when disturbed, and then doubling the proportional band setting of the controller.

1.2. *Addition of Integral Action*

With the controller adjusted as explained above, the errors are often unacceptably large. Improvement can often be obtained by making the gain of the controller different for different frequency components of the error signal.

The integral of a sinusoidal error signal increases in magnitude with decrease in the frequency of oscillation of the signal and tends to infinity for a constant value of error. This also implies that the error could be zero for any desired controller output if an error had previously existed.

An integral action term is usually added to a proportional term in a two-term controller to increase the gain at frequencies lower than f_0. It has an unstabilizing effect on the plant due to its lagging phase shift but improved response to a constant set level or disturbance can often be obtained without making the transient response unacceptable.

1.3. *Addition of Derivative Action*

The improvement obtained using integral action is in reducing the magnitude of the lower frequency components of the error signal. This tends to give the process a slow response.

The derivative of a sinusoidal error signal increases in magnitude with increase in the frequency of oscillation of the signal, and tends to infinity for infinite frequency of oscillation. It is not possible physically to realize a device having such properties, but over a reasonable range of frequencies approximation to derivative action can be obtained. Associated with the increase in gain is an advance in phase, which usually tends to stabilize the process control system.

Combinations of a proportional term with an integral action term and a derivative action term are usually made in a three-term controller.

1.4. *Analogue Three-term Controller*

An analogue three-term controller receives the measured value and set value signals from which it forms an error signal ε. It operates on this signal to form the *controller output* V as shown in Fig. 1 obeying the relationship

$$V = G\left\{\varepsilon + T_D\frac{d\varepsilon}{dt} + \frac{1}{T_I}\int \varepsilon\, dt\right\} \qquad \ldots\ldots(1)$$

where G = proportional gain factor

T_D = derivative action time

and T_I = integral action time.

Devices which produce this control action can be physically realized using various media, the most practically important being pneumatic and electronic. Analogue controllers of all varieties have received considerable attention elsewhere[2, 3] but this paper is concerned with the objectives and physical realization of digital electronic controllers.

1.5. *The Use of Digital Techniques in the Control of Processes*

Analogue control systems and recording devices have been used since the earliest times. Such devices will cheaply and reliably operate with errors in the order of 2% of full range. Such arrangements always require one set of hardware per control loop although time-shared analogue recording of variables is possible.

To achieve better accuracy without greatly increased cost it is normally necessary to use digital techniques; however there are several other advantages associated with digital systems, including better resolution, simpler and more effective data transmission, logging and display of control information, and simpler control of the controller itself. Furthermore the time-sharing of hardware is usually possible leading to a substantial reduction in cost when it is feasable to control a number of loops simultaneously.

1.6. *Direct Digital Control*

With the advent of digital computers it has become fairly common to control a number of processes simultaneously by means of a single time-shared on-line computer. The computer is normally programmed to achieve the numerical equivalent of two- or three-term control. At present for economic reasons one general-purpose computer has to serve a minimum of fifty process control loops.

However the flexibility and high speed of computation of large modern computers is such that many additional functions can be performed. These include data logging, on-line interpretation of non-linear instruments, automatic plant start-up and shut-down and on-line optimization of the system to produce the most economic process operation.

If all these operations are occurring time-shared with the central control function, certain hardware failures could lead to a catastrophic situation. There are various apparent solutions to this difficulty. One is to use two similar computers, with one continuously on stand-by with a third checking for a failure on the machines. Another is to have analogue back-up on every loop; these are switched in to perform control in the event of a computer failure. A third solution is to use small

special-purpose computers to perform more reliably the control of individual loops. The parameters of these controllers can be adjusted by a central computer; in the event of a failure in the central computer the system can be made to function reasonably satisfactorily using the settings prior to the time of failure. This paper contains a consideration of the justification of the use of the third method and its implementation.

2. Special-purpose Digital Controllers

2.1. *Justification of the Use of a Special-purpose Digital Controller*

The use of a single digital controller for the control of each loop separately combines the major advantages of the digital system with the reliability of the analogue controller. It will free the full-scale digital computer to perform the tasks the computer can do best, such as alarm scanning, data logging, process identification, optimization of controller parameters, and automatic up-dating of the set value whilst the controller continues to control its own loop reliably.

The speed at which calculations can be performed on such a controller does mean however that its application in the control of a single loop would appear to be wasteful. The cost of a digital controller is inevitably greater than the cost of an analogue controller even for the same accuracy. However, a compromise in cost-effectiveness may be achieved by addition of extra digital storage and similar facilities so as to allow the single digital controller to cater for the simultaneous control of a small number of loops. It would appear that it would be possible to compete economically with analogue controllers if about four loops were controlled from a single digital controller. It must be emphasized that there will be no need for a hierarchical general-purpose computer in such a system although it may be advantageous to have one available for the performance of the tasks mentioned above. For example, if during an alarm scan a failure of one special-purpose controller were to be detected it could be arranged that a replacement controller would be automatically substituted.

Furthermore, the use of a special-purpose digital controller is very attractive when only a few loops have to be controlled in which case it is quite impossible to justify the purchase of a general-purpose computer because of the high cost per control loop, the expense of the development of specialized software and the high degree of local expertise required.

2.2. *Philosophy of Design of a Research Prototype*

2.2.1. General principles

The output from a continuous three-term controller should effectively obey the relationship given in equation (1). A digital device must work on sampled values of the error ε and should produce at its output an approximation to the signal V defined in this equation.

Thus if the error and output are only considered to change at sampling instants time T_s apart and $\varepsilon(n)$ represents the value of ε at the nth sampling instant, then equation (1) may be approximated by

$$V(n) = G\left\{\varepsilon(n) + \frac{T_D}{T_s}(\varepsilon(n)-\varepsilon(n-1)) + \frac{T_s}{T_I}\sum_{n=-\infty}^{n}\varepsilon(m)\right\} \quad (2)$$

To provide satisfactory control action for the process, the sampled control signal $V(n)$ must be converted into a suitable form. This form may be produced from the sampled signal by means of a zero-order hold (or clamp). Such a unit could be composed of a digital store containing $V(n)$ in binary form coupled to a digital-to-analogue converter. A controlling system arranged in this way is usually referred to as a *positional* control system.

In an alternative scheme known as an *incremental* control the required *change* in the controller output ($\Delta V(n)$) can be evaluated at sampling instants. This signal is then used to actuate an integrating mechanism (e.g. a stepping motor). To obtain an overall three-term control action using this incremental scheme it is necessary to evaluate the second derivative of the error (instead of the integral of the error). A suitable algorithm for computing $\Delta V(n)$ is thus

$$\Delta V(n) = G\left\{\frac{T_s}{T_I}\varepsilon(n)+(\varepsilon(n)-\varepsilon(n-1))+ + \frac{T_D}{T_s}(\varepsilon(n)-2\varepsilon(n-1)+\varepsilon(n-2))\right\} \quad ...(3)$$

In a research model it is desirable to incorporate as many options as possible so that various techniques may be investigated. It is possible to investigate both positional and incremental control methods if the controller is arranged to generate four terms with adjustable coefficients as in equation (4):

$$V(n) = G\left\{\varepsilon(n) + \frac{T_a}{T_s}\left[(\varepsilon(n)-\varepsilon(n-1))+ + \frac{T_b}{T_s}(\varepsilon(n)-2\varepsilon(n-1)+\varepsilon(n-2))\right] + \frac{T_s}{T_c}\sum_{m=-\infty}^{n}\varepsilon(m)\right\} \quad .....(4)$$

The implementation of this algorithm can be understood by reference to an informational flow diagram as shown in Fig. 2. This diagram indicates the arithmetic functions to be performed but does not apparently contain information regarding the digital storage requirements. However, careful study of the diagram shows that a practical implementation of this scheme will require digital storage of four time-varying quantities (e.g. $\varepsilon(n)$, $\varepsilon(n-1)$, etc.) together with the four adjustable parameters of the controller (e.g. G, T_a/T_s, etc.). The other information required by the controller includes set value (S.V.), measured value (M.V.) and T_s which may also have to be stored. A program sequence for the realization of the algorithm is given in the Appendix.

2.2.2. Special requirements

When using a modern general-purpose digital computer to evaluate $V(n)$ the range of numbers that can be stored is so great that the programmer need not worry about size of constants or variables. If a special-purpose unit is to be made relatively cheaply then it will work in

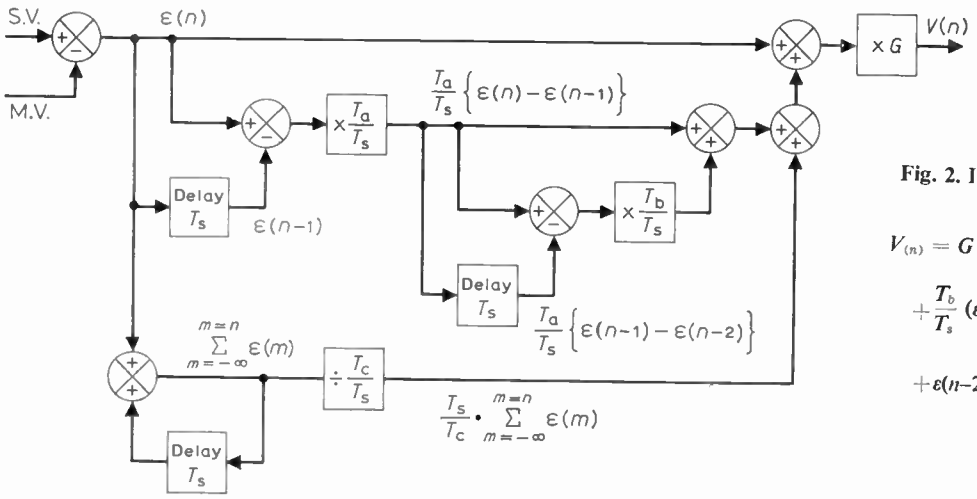


Fig. 2. Informational flow diagram.

$$V_{(n)} = G\,.\left\{\varepsilon(n) + \frac{T_a}{T_s}\left[(\varepsilon(n) - \varepsilon(n-1)) + \frac{T_b}{T_s}(\varepsilon(n) - 2\varepsilon(n-1) + \varepsilon(n-2))\right] + \frac{T_s}{T_c}\cdot\sum_{m=-\infty}^{n}\varepsilon(m)\right\}$$

natural binary code throughout, with limited word length storage.

To make resolution and accuracy of the set value at least as good as that of commercial analogue controllers, it is necessary to represent set value and measured value by 11 binary digits where the most significant bit (m.s.b.) represents −1 (100% of negative range) and the least significant bit (l.s.b.) represents 1/1024 (approximately 0·1% of full positive range).

Study of Fig. 2 shows that numbers stored inside the computer are multiplied by T_a/T_s, T_b/T_s and G while another number is divided by T_c/T_s. It has been shown[4] that to obtain a reasonable approximation to derivative and integral action the sampling must be at such a rate so as to make the values of T_a/T_s, T_b/T_s and T_c/T_s lie within the range 2 to 32. Practical experience shows that there is little advantage in having the resolution better than can be achieved by using fixed settings of 2, 4, 8, 16 and 32 for these constants. The limitations on these settings will of course simplify the hardware implementation. This simplification can be extended if the settings of G are fixed at $\frac{1}{2}$, 1, 2, 4, 8, 16 and 32.

The controller must be arranged to calculate a new value of its output $V(n)$ every sampling interval (T_s) and it is generally desirable that this interval shall be adjustable to match the dynamics of the process. The time to perform the calculations must be less than the smallest setting of T_s and it should be as small as possible in order to minimize the pure time-delay introduced by the controller.

3. Implementation of Four-term Digital Controller

3.1. *General*

In order to investigate the relative advantages of using incremental and positional control as explained in Sect. 2.2.1 a research model of a four-term digital controller (called Digicon[4]) was developed to obey equation (4). It should be noted that a four-term controller was chosen to allow the advantages of using a second derivative of error in positional control to be investigated; normal three-term incremental control could then also be readily achieved by setting T_s/T_c to zero.

Digicon is constructed using integrated circuit logic elements, the program constants and controls being determined by switches mounted on the front panel. The use of switches serves the dual purpose of displaying the values of the various constants as well as providing storage, so minimizing the number of electronic units.

A block diagram of Digicon is shown in Fig. 3.

The counter timing unit (c.t.u.) generates the timing waveforms and so controls the action of the other units in accordance with the wired program and front panel switches.

The digital error unit is fed with the set value (S.V.) from the front panel switches, and the measured value (M.V.) in the form of a square-wave signal. An error signal is formed and stored for parallel loading into the accumulator.

The calculations and overload unit (c.o.u.) operates on the numbers in the accumulator and from the various stores according to the programmed instructions from the c.t.u. The address logic controls the serial routing of numbers between the various shift registers in the storage unit and the c.o.u.

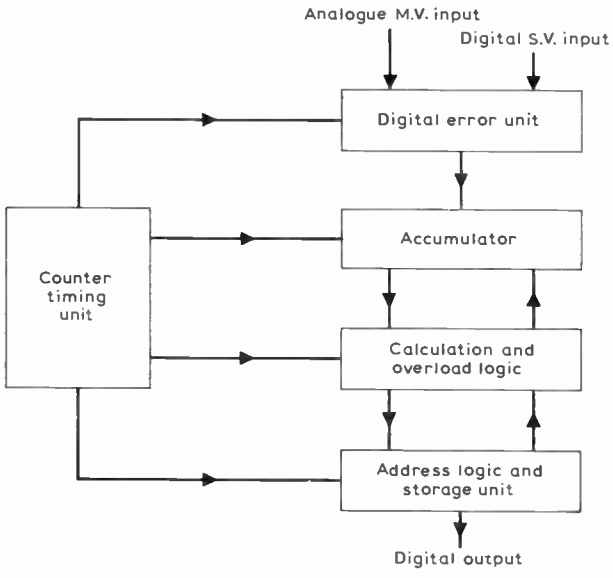


Fig. 3. Block diagram of Digicon.

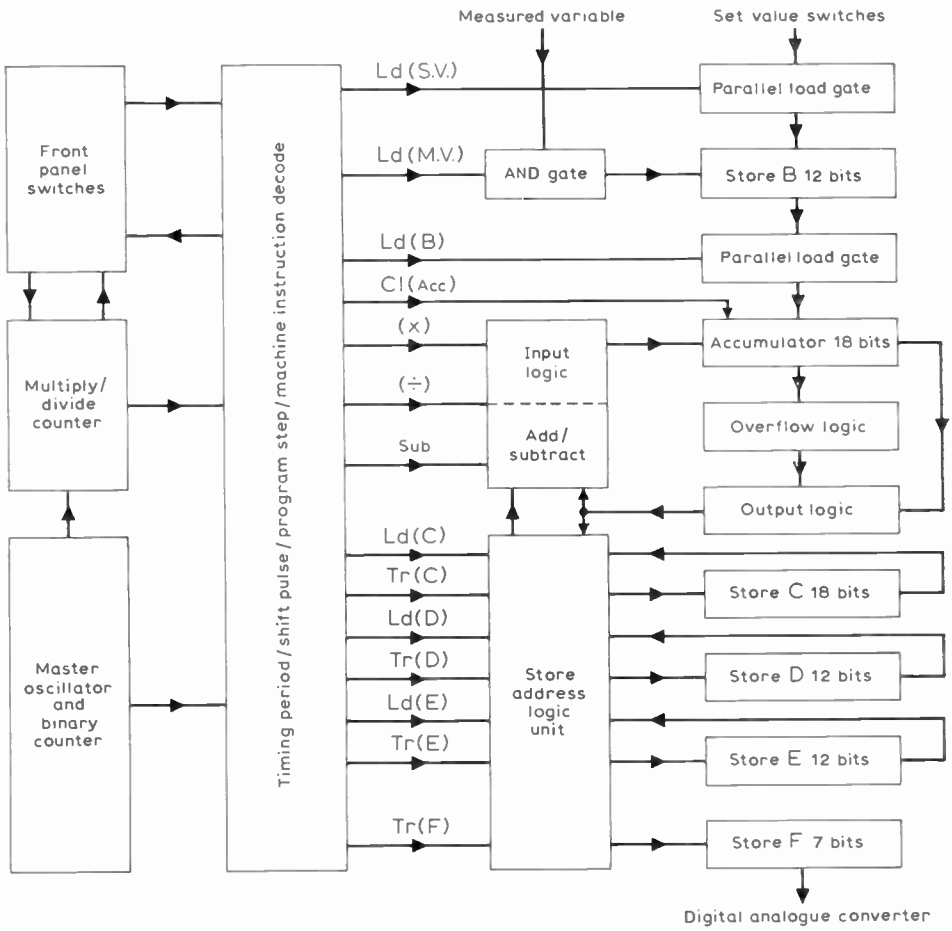


Fig. 4. Detailed arrangement of Digicon elements.

The digital output from the controller is fed in parallel from a non-circulating shift register store.

3.2. *Detailed Arrangement of the Controller Hardware*

The block diagram of Digicon shown in Fig. 3 is expanded and given in more detail in Fig. 4. This section contains a description of the purpose and operation of the sub-blocks and is intended to be read with continuous reference to Fig. 4. A full description of the electronic circuitry and detailed mode of operation can be found elsewhere.[4]

Digicon was made using dual in-line RTL integrated circuits which were commercially available at the outset of the project in 1968. The overall cost of components was in the order of £150. Technological developments since that time and the results from the work presented later in this paper indicate that a much cheaper and simpler arrangement could now be implemented.

3.2.1. The counter-timing unit

A master oscillator producing a square wave of frequency 131 kHz, binary counting chain and decoding gates are used to produce the timing waveforms which control the action of the digital error unit, and the sequence of the calculation. The sampling interval T_s is determined by selection of an output from some desirable point on the counter chain. This selection is made by means of a rotary switch mounted on the front panel giving values of T_s from 1/8 to 8 seconds.

The time for analogue to digital conversion (a.d.c.) is 1/32 s after which the calculation steps are 'obeyed' using the new value of error and previously stored values. The 16 program steps (Pr0–Pr15) each of 32 timing periods (Tp0–Tp31) take 1/128 s. During Pr14 the number stored in the digital-to-analogue converter is changed to the newly calculated value after which it is held constant for the following sampling interval.

During one program step various parallel actions can take place, such as clearing the accumulator, or parallel loading into the accumulator from store B. However, to reduce the hardware requirements most of the actions take place serially, with 12 or 18 shifting pulses. These are used to circulate the numbers in the various stores, along various paths so as to transfer numbers, or to perform addition or subtraction depending on the program step being obeyed. To enable multiplication and division to be performed either 6 or all of the usual 18 shifting pulses are inhibited and a number of shifting pulses between 0 and 6 are generated by a special counter under the control of the front panel switches on which the controller constants are set.

3.2.2. The digital error unit

An analogue measured variable signal (MVS), in the form of a variable frequency square wave, can be conveniently

converted into digital form by counting the number of cycles received in a known time. In particular the number counted in a 12-bit counter (store B) during the conversion time, determined by counting 2^{12} cycles of the master oscillator, will be effectively zero (ignoring the overflow bit) when the frequency of the measured variable signal (FMV) is the same as that of the master oscillator (FMO). If store B is considered to hold numbers in the range ± 2 with a resolution of 1/1024 then each cycle of MVS represents 1/1024, and the full 2^{12} cycles represents 4 which is out of range. However, when FMV = 1·25 FMO, the number of cycles counted during the same conversion time will be $(2^{12}+2^{10})$ which represents $+5$ but is stored as $+1$. The number in store B will therefore have had $+1$ added to it, which is equivalent to having had -1 subtracted from it. Similarly, when FMV = 0·75 FMO, the number of cycles counted during the same conversion time will be $(2^{12}-2^{10})$, which represents $+3$ but is stored as -1, and so the number in store B will have had $+1$ subtracted from it. This results in a system for subtracting the digital number representing the measured variable in the range ± 1 from a number in store B representing the set value in the range ± 1 and so producing a value of the error $\varepsilon(n)$ in the range ± 2.

The 11-bit set value number is set and displayed on front panel toggle switches and is parallel loaded into store B just before the MVS is gated into the counter.

3.2.3. The accumulator

This is an 18-bit shift register into which the 12-bit number from store B can be parallel loaded. The most significant bits are made the same as the sign bit of the original 12-bit number to form a correct 18-bit number in the range ± 128.

The accumulator can serially feed a 12-bit or 18-bit store when it receives 18 shift pulses, provided that the 18-bit store is fed with the same 18 shift pulses, and that the 12-bit store is fed with the first 11 and the last of these 18 pulses.

The number in the accumulator can be serially circulated via the arithmetic unit and so have added to it or subtracted from it the number from a 12-bit or 18-bit store.

The number in the accumulator can be divided by the mth power of two ($0 \leqslant m \leqslant 6$) by shifting the number right by m binary places, keeping the most significant bits the same as that of the original number.

The number in the accumulator can be multiplied by the mth power of two ($0 \leqslant m \leqslant 6$) by shifting left m binary places. This is effectively performed by circulating the 18-bit number with $(18-m)$ shift-right-pulses. To ensure that a correct answer is possible, the original number must be a 12-bit number and the 6 most significant bits are made zero before circulation starts. When $m = 6$ the 18-bit answer will be formed with zeros for the 6 least significant bits and the original 12-bit number forming the 12 more significant bits. When $0 \leqslant m < 6$, the correct answer will be formed with m zeros at the least significant end, followed by the original 12 bits and the $(6-m)$ most significant bits will be the same as the original sign bit. The production of the correct signals is performed by the input logic unit.

3.2.4. Number overflows

The state of each of the 7 more significant bits of the 18-bit number in the accumulator should be the same if the answer is to be correctly given by the 12 least significant bits. A test for a short number overflow is made after most program steps by checking these 7 bits and if detected then a positive or negative full house signal is used during the next program step instead of the false 12-bit number in the accumulator.

The only need for 18-bit numbers is in forming the integral term. The extra magnitude is needed because after division by T_i/T_s the weighted value of the integral term should still be capable of producing full scale output.

However, adding a number to this 18-bit number could produce a 19-bit answer with the 18 l.s.b. being inaccurate. When this condition could arise, a check has to be made to detect if a long number overflow has occurred, and if this is detected a 'full house' signal is generated and used in the next program step.

Another overflow signal is used to stop integral wind-up. If the controller output is at its maximum value, indicated by an overflow during its calculation in program step 13, a bistable is set which inhibits the further increase of the integral term when next calculated in program step 7.

The overload logic therefore controls which version of the number in the accumulator is to be fed to the stores and arithmetic unit via the accumulator output logic unit.

3.2.5. Storage of calculated values

Most stores are 12 or 18-bit shift registers, which receive 12 or 18 bit shifting pulses each program step during which the calculation is performed. The numbers normally circulate via end-around connexions but new numbers can be fed in when programmed as a transfer from the accumulator. When numbers are loaded into the accumulator via the arithmetic unit the original number usually circulates (i.e. only a copy is taken). However, a new number from the accumulator could be fed into the store while the old number is being fed out if both load and transfer orders are programmed (e.g. LdD, TrD).

Store F is effectively a 7-bit store which receives a new number from the accumulator's 7 medium significant bits serially during program step 14, and feeds in parallel the d.a.c. at all times. It therefore acts as the zero-order hold necessary for controlling analogue processes. The research model uses weighted resistance inputs to an operational amplifier which are fed from the store F bistables to produce the analogue output in the range ± 1 V.

3.2.6. The calculation in detail

The sequence of machine instructions has been outlined in the Appendix. Each instruction is labelled with

a program step symbol (PrS, PrM, etc.) and a description of the implementation of these steps within the machine is given below.

(i) *Forming the error*

PrS. The 11-bit set value number in the range ± 1 is parallel loaded into store B about 1 ms prior to PrM.

PrM. The measured variable square wave in the range 1·25 FMO to 0·75 FMO representing the range -1 to $+1$ is gated into store B which counts the cycles received in the time taken for 2^{12} cycles of FMO. This makes the number in store B = S.V.$-$M.V. = error $\varepsilon(n)$ in the range ± 2 at the end of this 1/32 s.

(ii) *Calculation of the weighted first derivative of error term*

Pr1. The new value of error $\varepsilon(n)$ is parallel loaded into the accumulator from store B to form an 18-bit number in the range ± 128. $\varepsilon(n)$ is circulated via the arithmetic unit and has subtracted from it the 12-bit number in store E which holds the previous value of error, $\varepsilon(n-1)$.

At the end of this program step the number in the accumulator will be $\varepsilon(n)-\varepsilon(n-1)$, the first derivative of error. This number must be a correct 12-bit number (i.e. all 7 most significant bits must be the same) in order to keep following calculations correct. If, say, the 6 m.s.b. were '1' and 7th was '0' representing -4 then a negative short number overflow would have occurred, and the next program step should use a '$-$fh' number having the 7 m.s.b. at '1' and 11 l.s.b. at '0' which represents -2. The overflow procedure (SNO) is repeated at each program step unless stated.

Pr2. The number in the accumulator is to be multiplied by T_a/T_s (given by 2^m; $0 \leqslant m \leqslant 6$) which is set on a front panel switch. The normal first 11 shift pulses are used together with a further m (controlled by the multiply/divide counter gated with Pr2) and the last (18th) shift pulse to circulate the number in the accumulator via the output logic, arithmetic unit and input logic units. The 6 v.m.s.b. are first set to zero, and during the m shift pulses the signal fed into the accumulator is the sign bit of the original number.

The number returned to the accumulator is

$$(T_a/T_s)(\varepsilon(n)-\varepsilon(n-1)) = (d)$$

which is the weighted first derivative of error.

(iii) *Calculation of the sum of the weighted first and second derivative terms*

Pr3. The number in the accumulator (d) is transferred to store D. This store held the previous value of the weighted first derivative which is now

$$(T_a/T_s)(\varepsilon(n-1)-\varepsilon(n-2))$$

and this number is subtracted from the number held in the accumulator to form in the accumulator

$$(T_a/T_s)(\varepsilon(n)-2\varepsilon(n-1)+\varepsilon(n-2))$$

the part-weighted second derivative term.

Pr4. The number in the accumulator is next multiplied by T_b/T_s in a similar manner to that described in Pr2, thus forming the weighted second derivative term

$$(T_a/T_s)(T_b/T_s)(\varepsilon(n)-2\varepsilon(n-1)+\varepsilon(n-2)) = (d^2).$$

Pr5. The weighted first derivative term from store D is now added to the number in the accumulator.

Pr6. The result of Pr5 is temporarily transferred to store E as

$$\frac{T_a}{T_s}\left[(\varepsilon(n)-\varepsilon(n-1)) + \frac{T_b}{T_s}(\varepsilon(n)-2\varepsilon(n-1)+\varepsilon(n-2))\right] = (d+d^2)$$

the sum of the weighted first and second derivative terms.

(iv) *Calculation of the weighted integral term*

Pr7. The accumulator is cleared, and if there was no overflow when the controller output was last calculated in Pr13 then the error is loaded into the accumulator from store B. The 18-bit number in store C is now added to the 18-bit number in the accumulator to form a new integral of error term. This could form a 19-bit number, therefore a long number overflow check must be made.

Pr8. The correct 18-bit number representing the integral of error is circulated back into the accumulator, and a copy of it is fed into store C as the updated value of the integral term. (No overflow check is required.)

Pr9. The number in the accumulator is now divided by T_c/T_s (given by 2^m; $0 \leqslant m \leqslant 6$) by a shift right of m pulses generated by the multiply/divide counter in place of the usual 18 shift pulses. (No overflow check is required.) This generates the weighted integral of error

$$\frac{T_c}{T_s} \cdot \sum_{m=-\infty}^{n} \varepsilon(m) = (i)$$

(v) *Calculation of the controller output*

Pr10. To the weighted integral term (i) in the accumulator is now added the number in store E ($d+d^2$) to form ($i+d+d^2$).

Pr11. The correct version of ($i+d+d^2$) is transferred temporarily to store E.

Pr12. The present value of the error is parallel loaded into the accumulator from store B, and this is circulated together with the number from store E via the adder to form ($p+i+d+d^2$) in the accumulator.

Pr13. The number in the accumulator is now multiplied by $G \times 2$ (where $G = 2^m$; $0 \leqslant m \leqslant 6$) as described in Pr2. The 12-bit number in the accumulator at the end of this program step is $2G(p+i+d+d^2)$ in the range ± 2. If a short number overflow is detected an extra bistable (Of) is set which inhibits LdB during the next calculation at Pr7.

Pr14. The correct 7 bits representing the most significant part of the number in the range ± 2 are transferred to the 7-bit store F where they represent the output in the range ± 1; therefore this number will be:

$$V(n) = G\left\{\varepsilon(n) + \frac{T_a}{T_s}\left[(\varepsilon(n)-\varepsilon(n-1)) + \frac{T_b}{T_s}(\varepsilon(n)-2\varepsilon(n-1)+\varepsilon(n-2))\right] + \frac{T_s}{T_c}\sum_{m=-\infty}^{n}\varepsilon(m)\right\}$$

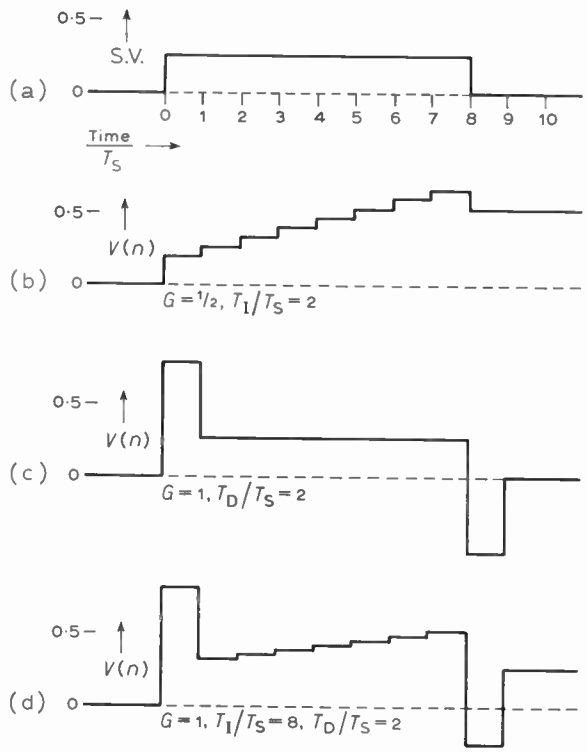


Fig. 5. Controller output due to step changes in set value for various settings.

(vi) *Preparation for next sample*

Pr15. The present error in store B is parallel loaded into the accumulator and then serially transferred into store E. Store B is then cleared and the set value is parallel loaded from the switches as outlined in PrS. The sequence repeats with a periodic time T_s which is adjusted by a front panel switch.

4. Practical Tests on Digicon and the Determination of Optimum Settings

The digital parts of Digicon can be tested most conveniently using the set value switches to produce an input and by observing the output on the neon-glow indicators. However, permanent records are more easily obtained after the digital output has been converted into analogue form. Reproducible test signals are also more easily obtained from a low-frequency waveform generator feeding the a.d.c. This implies that most tests involve the use of the simple a.d.c. and d.a.c. and some errors must be due to their limited accuracy.

In order to assess the performance of the controller in a closed loop, an electronically-simulated plant was developed and used. This allowed the effect of controller settings to be investigated for a process described by up to five exponential lags.

4.1. *Tests on the Controller*

Step response tests can readily be performed by insertion of switched changes in the set value with the controller on open loop. The waveforms in Fig. 5 show the controller outputs V for a step change in set value for (b) proportional plus integral control, (c) proportional plus derivative control, and (d) proportional plus derivative plus integral control. These outputs are in exact accordance with what would be predicted from theoretical considerations but indicate very clearly the differences between the outputs of a digital controller and an ideal analogue controller. For instance, an ideal analogue controller with proportional plus derivative plus integral action would produce an output consisting of an infinite impulse at the instant of application of the step, combined with a step and a ramp. It may be observed from Fig. 5(d) that the digital device does in fact produce a short pulse of area equal to the area of the infinite derivative component and this is subsequently followed by step and incremental ramp components.

4.2. *Closed-loop Response to Step Changes*

4.2.1. The system

In order to investigate the performance of the controller when operating in conjunction with a typical plant an electronic plant simulator having five simple lags of adjustable time-constants was connected in a closed-loop arrangement with Digicon as the controller (Fig. 6).

The parameters of the controller must be set so that the performance of the system is adequate with respect to response to step changes in both the set value and disturbance (θ_d).

4.2.2. Optimizing the controller settings

In order to obtain optimum performance from the system it is necessary to match the dynamics of the controller to that of the plant. Various authors[5, 6] have explained how this may be achieved for analogue

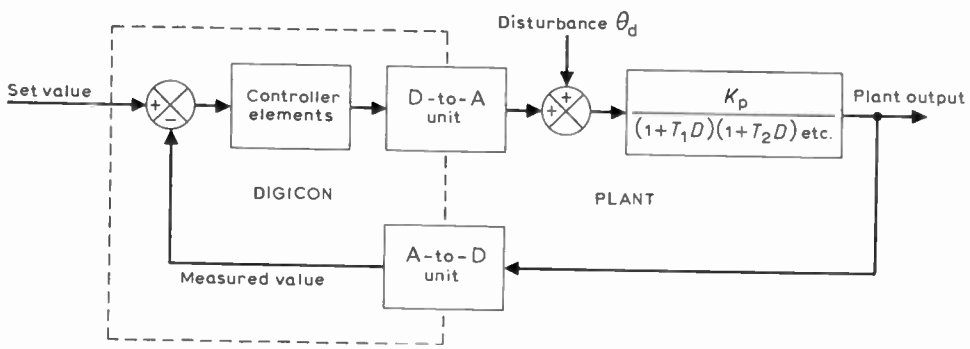


Fig. 6. Closed-loop arrangement for system appraisal.

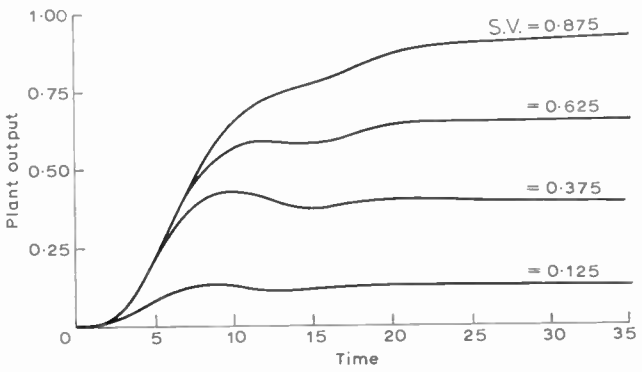


Fig. 7. Plant output due to various step changes in set value.

three-term controllers. These methods can be modified for application to systems containing digital controllers.[7] The sampling time T_s is best fixed in relation to the frequency f_0 at which the system would oscillate continuously on closed loop with proportional action alone; in practice, this angular frequency may be determined either by computation when the plant parameters are known or by experiment when they are not.

The authors have found that in practical systems sampling more frequently than sixty times per cycle of transient oscillation offers no advantage whereas sampling too slowly reduces the effectiveness of the derivative action. A suitable value of T_s is found to be given by

$$T_s = \frac{1}{32 f_0}$$

It has been suggested[5] that a suitable value of T_I is given by

$$T_I = \frac{1}{2 f_0}$$

and the optimum ratio of T_I/T_D is 4.

This results in controller settings of

$$T_I/T_s = 16 \text{ and } T_D/T_s = 4$$

The controller gain setting G may then be adjusted for adequate damping. (A setting equal to about one-half that required to make the system oscillate continuously on closed-loop with proportional action alone will normally give the required degree of damping.)

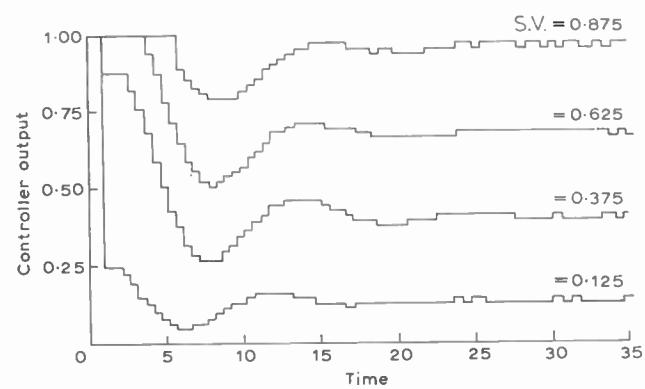


Fig. 8. Controller output due to various step changes in set value.

4.2.3 Tests on the system

Many tests were made with a variety of plants, but the following were typical and illustrate the main findings. In order to investigate the performance of Digicon, a plant having time-constants 0·5 s, 2 s, 4 s, 1 s and 1 s respectively and a zero frequency gain of unity was used. In the tests described, Digicon was used as a positional controller rather than an incremental controller and in this mode the switch setting T_a/T_s is equivalent to T_D/T_s (the effective derivative action control) and T_c/T_s is equivalent to T_I/T_s (the effective integral action control).

In accordance with Section 4.2, the controller settings were $T_s = \frac{1}{2}$ s, $G = 2$, $T_I/T_s = 16$ and $T_D/T_s = 4$. Figure 7 shows a series of plant outputs for various size step changes in set value. The non-linear effects for large signals are clearly illustrated. Figure 8 shows the controller outputs for the same tests.

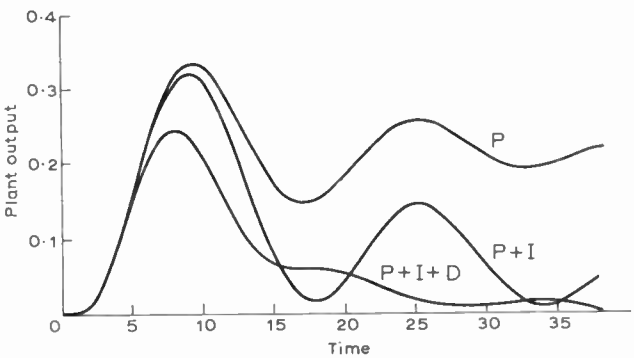


Fig. 9. Plant output due to 0·75 step disturbance for various controller settings.

In order to illustrate the advantages of integral action and derivative action, tests were performed with a step disturbance (the set value being held arbitrarily at zero). Figure 9 shows the response for proportional action alone and indicates a steady error following the initial oscillatory transient; the response for proportional plus integral action shows this error being reduced but in a highly oscillatory manner. Finally with three-term control action it is seen that the transient error is rapidly forced to decay to zero. All these results are as would be expected.

The effect of using the second derivative term in addition to the normal three terms was investigated and it appeared that for relatively low-order systems some improvement in the response could be achieved. However, with higher-order systems (i.e. greater than 4th order) the improvement in performance was found to be small.

The use of the controller in the incremental mode produces results which depend on the method of integrating the digital output. When an integration period which is short compared with T_s is used, the overall response of the system is identical to that produced using positional control. However, if the integration period is equal to T_s, or longer (which is possible when using a mark/space ratio type of d.a.c.) the extra delay reduces the advantages of the derivative action, so that

when slow sampling is used there may be no net phase advance from the controller. Thus when it is desirable to obtain effective three-term control action in the incremental mode it becomes even more essential to select an adequate sampling frequency. This point is frequently overlooked by practising engineers.[8]

5. Conclusion

This paper has shown that it is possible to achieve more effective control using a digitally-instrumented control system than by using a purely analogue arrangement. This fact has tended to persuade control engineers to adopt the general-purpose computer as a software-programmed on-line controller for simultaneously controlling many loops. Unfortunately there is evidence that such arrangements are both costly and unreliable. This has led to the present investigation which set out to produce a special-purpose digital controller with a hardware program. A reasonably inexpensive device has been designed and successfully tested in conjunction with various plants. The paper contains a description of the hardware and the results of a few typical tests.

The investigations have led the authors to believe that such a special-purpose controller may be made very much more economically than the original research model without impairing the effectiveness of the controller in practice. For instance, there appears to be no practical advantage in having a ratio of integral action time to derivative action time of other than four. This will limit the size of the stores required and reduce the need for overload detection. Also whilst it has been shown that for certain systems it is advantageous to have three-term control, it is equally true that in other systems, derivative action may have little or no advantage and proportional plus integral control is then adequate.

These practical considerations, combined with the fact that large-scale integrated logic elements are becoming progressively cheaper make special-purpose digital controllers a very attractive proposition.

6. References

1. British Standard 1523: Section 2: 1960. Section 2: 'Process Control'.
2. Atkinson, P., 'Feedback Control Theory for Engineers' (Heineman, London, 1968).
3. Young, A. J., 'An Introduction to Process Control System Design' (Longmans, Harlow, Essex, 1957).
4. Allen, A. J., 'A Wired-Logic Digital Multi-Term Controller', M.Phil.Thesis, Department of Applied Physical Sciences, University of Reading, 1970.
5. Ziegler, J. G. and Nichols, N. B., 'Optimum settings for automatic controllers', *Trans. Amer. Soc. Mech. Engrs*, **64**, pp. 759–68, 1942.
6. Atkinson, P. and Davey, R. L., 'A theoretical approach to the tuning of pneumatic three-term controllers', *Control*, **72**, No. 117, pp. 238–42, March 1968.
7. Bell, C. A. and Cutting, G. W., 'Application of a special-purpose digital controller to a chemical process'. Paper presented to 3rd International Congress of Chemical Engineers (CHISA), Prague, September 1969.
8. Thompson, A., 'Direct digital control—the state of the art in 1967', *Proc. Instn Elect. Engrs*, **115**, No. 10, pp. 1541–47, October 1968.

7. Appendix 1. Machine-language program

Store (B) only receives information (in parallel) from the set value store (front panel switches) during Program step S (PrS) and the measured variable pulses via a gate opened during PrM. Store (B) only feeds into the accumulator (A) in parallel during the early part of a program step.

The stores (C), (D), (E) and (F) can only be fed from (A), and only stores (C), (D) and (E) can feed into (A) via the arithmetic unit. The stores (S0), (S1), (S2) and (S3) which are the program constant switches G, T_c/T_s, T_a/T_s, T_b/T_s respectively, affect the number of shift pulses used in order to modify the value stored in (A).

The number from an addressed store is added to the number in the accumulator c(A) unless specifically programmed to subtract (SUB) it from c(A).

Unless specifically asked not to by $\overline{\text{SNO}}$ a short number overflow test is made at the end of each program step. If a long number overflow test is required (LNO) it has to be programmed. Convenient symbols are Ld meaning load, Tr meaning transfer.

Obeying the above rules, the program can be written as below.

PrS	Ld(SV)
PrM	Ld(MV)
Pr0	
Pr1	Ld(B), (SUB), Ld(E)
Pr2	((X); (S2))
Pr3	Tr(D), (SUB), Ld(D)
Pr4	((X); (S3))
Pr5	Ld(D)
Pr6	Tr(E)
Pr7	Ld(B).$\overline{\text{Of}}$, Ld(C), $\overline{\text{SNO}}$, LNO
Pr8	Tr(C), $\overline{\text{SNO}}$
Pr9	((÷); (S1)), $\overline{\text{SNO}}$
Pr10	Ld(E)
Pr11	Tr(E)
Pr12	Ld(B), Ld(E)
Pr13	((X); (SO)), Set 'Of' bistable if overflow detected
Pr14	Tr(F), $\overline{\text{SNO}}$
Pr15	Ld(B), Tr(E)

Manuscript first received by the Institution on 23rd August 1971 and in final form on 22nd December 1971. (Paper No. 1431/IC 59.)

U.D.C. 621.376.23: 621.391.822.2

Relative Merits of Quadratic and Linear Detectors in the Direct Measurement of Noise Spectra

Professor H. SUTCLIFFE,
Ph.D., C.Eng., F.I.E.E., F.I.E.R.E.*

SUMMARY

Initially a brief discussion is presented of the relative merits of quadratic and linear detectors as regards ease and convenience of use following narrow-band filters in the measurement of noise spectra. An analysis is then given of the noise content in the output signals of the two types of detector. The analysis uses results already available in the literature of the subject, but is presented here in a manner which allows a direct and ready comparison. The analysis is not restricted to any one type of narrow-band filter and it is concluded quite generally that neither type of detector has a significant advantage.

* *Department of Electrical Engineering, University of Salford, Salford M54 WT*

List of Symbols

a parameter of quadratic detector (A V^{-2})
b parameter of linear detector (A V^{-1})
B_c effective noise bandwidth at detector input (Hz)
B_i parameter, akin to bandwidth, of input noise v (Hz)
B_0 effective noise bandwidth of low-pass filter (Hz)
e factor, ≤ 1
f frequency (Hz)
f_c centre frequency of spectrum (Hz)
f_0 parameter used in defining spectral shape (Hz)
$G(f)$ intensity of noise spectrum at detector output (A^2 Hz^{-1})
i output current from detector (A)
i_m estimated value of $\bar{i}$ (A)
i_n fluctuating component of i at output of low-pass filter (A)
$L = G(0)$ for linear detector (A^2 Hz^{-1})
$Q = G(0)$ for quadrature detector (A^2 Hz^{-1})
$S(f)$ intensity of noise spectrum at detector input (V^2 Hz^{-1})
S_c value of $S(f)$ at frequency f_c (V^2 Hz^{-1})
S_m estimated value of S_c (V^2 Hz^{-1})
v input fluctuation to detector (V)

1. Introduction

The type of measuring equipment under discussion is shown in Fig. 1. It is assumed that at some stage in the measurement the tunable filter is set with its centre frequency at a particular value f_c Hz, so that signal v at the output terminals of the filter has a noise spectrum $S(f)$ volts^2/Hz as illustrated in Fig. 2. The value $S_c = S(f_c)$ for various settings of f_c gives the required information about the noise spectrum of the device under test. Each S_c is obtained by measuring or deducing the mean square value of v and using the following relation:

$$\overline{v^2} = \int_0^\infty S(f)\,df = S_c B_c \qquad \text{.....(1)}$$

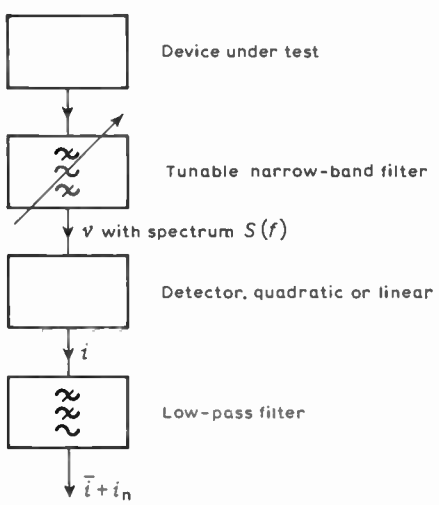


Fig. 1. Components of measuring system.

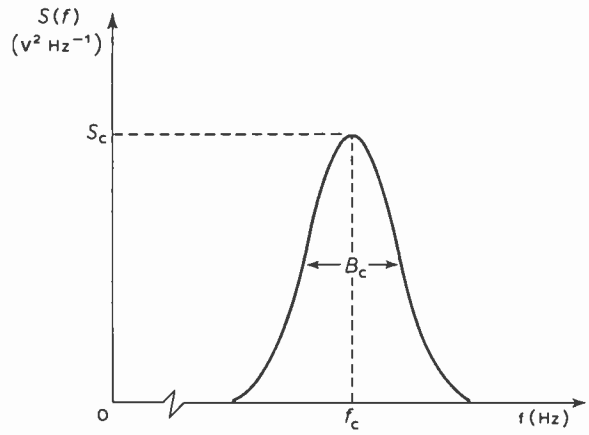


Fig. 2. Narrow band spectrum $S(f)$ and its effective bandwidth B_c. Area under curve $= S_c B_c$.

The quantity B_c in equation (1) is the effective noise bandwidth of the tunable filter when set to frequency f_c. We shall assume that its value is known, either from measurements or by calculation. As an example, in a typical wave analyser for the audio-frequency range, $B_c \simeq 7$ Hz.

It would appear at this stage of the discussion that a quadratic detector is most suitable for use in these circumstances since its performance fits naturally into the theoretical requirements. But this is perhaps too hasty a conclusion, for it is thoroughly established[1-3] that the fluctuation v at the output of a narrow-band filter is normal or Gaussian in distribution. It follows that the mean magnitude is related to the mean square value by the following simple equation.

$$\overline{v^2} = \frac{\pi}{2}\left(\overline{|v|}\right)^2 \qquad \text{.....(2)}$$

Hence, for detecting noise at the output of a narrow-band bandpass filter, a linear detector can be no less appropriate than a quadratic detector, since the linear detector responds to the mean magnitude $\overline{|v|}$ in full-wave rectifying circuits, and half this value in the more common half-wave circuits. In comparing the merits of quadratic and linear detectors we must seek other criteria.

2. Relative Convenience

One feature which could influence a choice between the two types is the relative ease and convenience of realizing the functions of quadratic and linear detection. In the writer's opinion it is much simpler to construct a precise linear detector than a precise quadratic detector. For the latter, one must use either the thermal effect as in thermojunctions or thermistors, or a non-linear squaring circuit of at least moderate complexity. For a linear detector all that is needed is a rectifying diode in conjunction with a signal source of high impedance. To a very close approximation the diode current is a rectified version of the input signal.

An argument against linear detectors is that they behave in a complicated manner when fed with an input consisting of noise plus a sinusoidal signal, whereas with quadratic detectors the relation is simple. Thus for quadratic detectors the output is proportional to the sum of the input noise $\overline{v^2}$ and signal $\overline{v_s^2}$. This leads to convenience in calibrating the measuring system with a sine wave signal generator. Yet it is not perhaps generally realized that the same procedure is almost as simple for a linear detector using published data.[4] The following table is a brief and approximate extract from the available information.

Table 1

Performance of linear and quadratic detectors with noise and sine wave inputs

r.m.s. signal in / r.m.s. noise in	mean output / mean with noise only: linear detector	mean output / mean with noise only: quadratic detector
0	1	1
0·5	1·12	1·25
1·0	1·45	2
1·5	1·90	3·25
1·6	2·00	
2·0	2·40	5

From the point of view of convenience in use, then, it appears that there is little advantage in selecting either type of detector. The following section discusses the more fundamental issue of statistical accuracy arising from noise components in the output signals of the detectors.

3. Statistical Accuracy

Fluctuations in the output signals of the detectors, especially with narrow-band measurements, impose limits on the accuracy of observations of the mean or true value. We shall consider the situation in which the detector, linear or quadratic, is followed by a low-pass filter of effective noise bandwidth B_0. This filter will pass the direct component $\bar{\imath}$ of the output i of the detector together with an additional fluctuating component i_n. If B_0 is small then i_n will also be small, but a penalty will have been paid since the response of the detector circuit will be slow. This problem is particularly acute with narrow band measurements as the following analysis shows.[5]

Let the rectifier characteristics be as follows:

Quadratic detector (Q)	Linear detector (L)	
output $i = av^2$	output $i = bv$, $v > 0$	(3)
	0, $v < 0$	

The next pair of equations quote well known relations between the mean output current of each detector and the noise spectrum of the input p.d.:

$$\text{(Q)}\quad \bar{\imath} = a\int_0^\infty S(f)\,\mathrm{d}f, \qquad \text{(L)}\quad \bar{\imath} = b\left[\frac{\int_0^\infty S(f)\,\mathrm{d}f}{2\pi}\right]^{\frac{1}{2}} \qquad (4)$$

At this stage we accept the result of some formidable mathematical theory and extract expressions for the intensity G (A^2 Hz^{-1}) of the low frequency noise components of i:

$$G(f) = 2a^2\int_0^\infty S(x)S(f+x)\,\mathrm{d}x \qquad \text{.....(5Q)}$$

$$G(f) = \frac{b^2}{4\pi\bar{v}^2} \int_0^\infty S(x)S(f+x)\,dx \quad .....(5L)$$

The discussion is now greatly simplified if it is recognized that in normal circumstances the low-pass filter bandwidth B_0 is small compared with the bandwidth of the input noise to the detector. This implies that for the estimation of i_n it will be legitimate to set $f = 0$ in equations (5) and to use, within B_0, simply the $G(0)$ values. These are given by the following expressions:

$$G(0) = Q = 2a^2 \int_0^\infty S^2(x)\,dx \quad .....(6Q)$$

$$G(0) = L = \frac{b^2}{4\pi\bar{v}^2} \int_0^\infty S^2(x)\,dx \quad .....(6L)$$

The mean square values of output noise i_n are now available

$$\overline{i_n^2} = QB_0 \quad A^2 \quad .....(7Q)$$

$$\overline{i_n^2} = LB_0 \quad A^2 \quad .....(7L)$$

The troublesome feature of the measurement is the output noise expressed as a fraction of the mean output. This is most conveniently expressed as follows:

$$\frac{\overline{i_n^2}}{(\bar{i})^2} = 2B_0 \int_0^\infty S^2(x)\,dx \left[\int_0^\infty S(x)\,dx\right]^{-2} \quad .....(8Q)$$

$$\frac{\overline{i_n^2}}{(\bar{i})^2} = \frac{B_0}{2} \int_0^\infty S^2(x)\,dx \left[\int_0^\infty S(x)\,dx\right]^{-2} \quad .....(8L)$$

Examination of equations (8) leads one to suppose that the quadratic detector is four times noisier than the linear detector. This is a mistaken assumption as will be shown later, but meanwhile the 'clumsiness' of (8Q) and (8L) demands a tidying operation. Both contain the factor $1/B_i$, where B_i (Hz) is given by:

$$B_i = \left[\int_0^\infty S(x)\,dx\right]^2 \left[\int_0^\infty S^2(x)\,dx\right]^{-1} \quad .....(9)$$

This cumbersome expression becomes meaningful if a particular example is considered. An idealized situation, commonly quoted,[6] assumes a rectangular spectrum $S(f)$, such that $S(f)$ is constant and of value S_c within bandwidth f_0 centred at f_c and $S(f)$ is zero elsewhere. Then from (9), B_i in this instance is equal to f_0 and is identified with the input noise bandwidth. Other input spectral distributions are discussed in reference 5 and in the Appendix, and it is found that the value of B_i as defined in (9) is normally close to B_c as defined by (1) and is also close in value to the '3 dB' bandwidth B_3. For the rectangular spectrum only do these three types of bandwidth coincide exactly.

We now return to equations (8) and replace the integrals by the symbol B_i, remembering that B_i is approximately the bandwidth of the tunable narrow-band filter in the measuring system.

$$\frac{\overline{i_n^2}}{(\bar{i})^2} = 2\frac{B_0}{B_i} \quad .....(10Q)$$

$$\frac{\overline{i_n^2}}{(\bar{i})^2} = \frac{1}{2}\frac{B_0}{B_i} \quad .....(10L)$$

Interpretation of the implications of these two equations requires a consideration of the experimental process. The experimenter would like to observe $\bar{i}$, but this is denied him. He can observe only the total output current $\bar{i}+i_n$. The noise component i_n is of course a function of time and our experimenter may have a lengthy time available to make a number of observations, or he may make a recording and estimate or compute its average. The fundamental aspects of the situation have been the subject of exhaustive and detailed theoretical studies.[7] The practical situation is that there will be an estimate i_m that will differ from the true value $\bar{i}$ by some error. This error is a statistical quantity with its magnitude proportional to the r.m.s. value of i_n. For single observations of i, of course, the r.m.s. error *is* the r.m.s. value of i_n. Thus from equations (10) we can obtain equations showing the errors in the estimated values i_m:

$$i_m = \bar{i}[1 + e(2B_0/B_i)^{\frac{1}{2}}] \quad .....(11Q)$$

$$i_m = \bar{i}[1 + e(0{\cdot}5B_0/B_i)^{\frac{1}{2}}] \quad .....(11L)$$

where e is a factor common to both types of measurement.

Finally, it will be recalled that the experimenter is seeking an estimate of the value S_c and assumes it to be proportional to i_m for the quadrature detector and to i_m^2 for the linear detector. Denoting the assumed value by S_m, we have:

$$\frac{S_m}{S_c} = 1 + e(2B_0/B_i)^{\frac{1}{2}} \quad \text{(for the quadrature detector)} \quad (12Q)$$

$$\frac{S_m}{S_c} = [1 + e(0{\cdot}5B_0/B_i)^{\frac{1}{2}}]^2$$

$$= 1 + e(2B_0/B_i)^{\frac{1}{2}} + e^2B_0/2B_i$$

(for the linear detector) ...(12L)

Equations (12) are identical except for the final term in (12L). This is in contradiction with the superficial conclusion which might have been drawn from equations (8). The final term in (12L) is of negligible significance in practical circumstances, as may be shown by a typical example. Suppose the value of $\bar{i}$ is estimated to within 10%, that is, $e(B_0/2B_i)^{\frac{1}{2}} = 0{\cdot}1$, then the final term is only 0·01 and may be neglected in comparison with the main error term.

It may be concluded, therefore, that for both quadratic and linear detectors the fractional error in the estimation of spectral intensity is given by:

$$\text{fractional error, r.m.s. value} = e(2B_0/B_i)^{\frac{1}{2}} \quad ...(13)$$

where B_0 is the effective noise bandwidth of the post-detector filter and is assumed to be much less than B_i.

B_i is a bandwidth-type function of the input spectrum as given by equation (9). It is similar in value to the 3 dB bandwidth of the narrow-band filter.

e is a factor never greater than unity. For a single estimate of output current, $e = 1$. For the average of a large number N of independent estimates,[2] $e = N^{-\frac{1}{2}}$.

It is of interest to consider a particular example. Suppose the detector input noise has a spectrum for which $B_i = 7$ Hz. An envelope detector is used, so the behaviour is identical to that of the linear detector except that the output fluctuations and mean value are scaled up by the factor π. A simple CR lag circuit with a time-constant of 5 seconds acts as the low-pass filter, which thus has its '3 dB' frequency at $(10\pi)^{-1}$ Hz and its effective noise bandwidth $\pi/2$ times this, that is $B_0 = (1/20)$ Hz. From equation (10L), the fractional r.m.s. output noise is $(2 \times 20 \times 7)^{-\frac{1}{2}} \simeq 0{\cdot}06$. From equation (13), fractional r.m.s. spectral error for each single observation of current is $0{\cdot}06 \times 2$ or 12%. This example exposes the deplorable situation encountered in measuring low-frequency spectra, for the measurement is inexact and also tedious because of the long time-constant. This has led to the consideration of time-varying filters.[8]

4. Conclusion

It has been shown that for noise spectra measurement there is little advantage in using either kind of detector since their statistical errors are identical. This result could perhaps have been inferred from general principles, but the detailed analysis has the merit of collecting together a quantity of relevant and useful information which is somewhat scattered in the literature of the subject.

5. References

1. Van der Ziel, A., 'Noise' (Prentice Hall, Englewood Cliffs, N.J.' 1954).
2. Davenport, W. B. and Root, W. L., 'Random Signals and Noise' (McGraw-Hill, New York, 1958).
3. Bell, D. A., 'Electrical Noise' (Van Nostrand, New York, 1960).
4. Bennett, W. R., 'Response of a linear rectifier to signals and noise', *Bell Syst. Tech. J.*, 23, pp. 97–101, 1944.
5. Smith, R. A., 'The relative advantages of coherent and incoherent detectors. A study of their output noise spectra under various conditions', *Proc. Instn. Elect. Engrs*, 92, Pt IV, pp. 43–54, 1951.
6. King, R., 'Electrical Noise' (Chapman & Hall, London, 1966).
7. Blackman, R. B. and Tukey, J. W., 'The Measurement of Power Spectra' (Dover, New York, 1958).
8. Ambrozy, A., 'Reducing the time requirement in direct reading v.l.f. noise measurement', *Proc. Instn. Elect. Engrs*, 53, pp. 1161–62, 1965.
9. Gradshteyn, I. S. and Ryzhik, I. M. (Trans. Jeffrey, A.), 'Tables of Integrals, Series and Products' (Academic Press, New York, 1965).

6. Appendix: Filters and noise bandwidth

We assume that the input noise to the narrow-band tuned filter is of uniform spectral density. Then the filter output noise spectrum $S(f)$ will take various forms depending on the response of the filter. Four types of filter, idealized to a greater or lesser extent, will be considered.

(a) Rectangular filter

$$S(f) = S_c \quad \text{within } f_c \pm f_0/2$$
$$\text{elsewhere, } S(f) = 0$$

(b) Gaussian filter

$$S(f) = S_c \exp\left[-\pi(f-f_c)^2/f_0^2\right]$$

(c) Single tuned circuit

$$S(f) = S_c[1+\{\pi(f-f_c)/f_0\}^2]^{-1}$$

(d) Critically coupled pair

$$S(f) = S_c[1+\{(f-f_c)/f_0\}^4]^{-1}$$

Three types of bandwidth associated with each noise spectrum were mentioned in the text. They are:

B_c, the effective noise bandwidth defined by equation (1).

B_i, a type of bandwidth for which no particular name has been devised. We could call it 'the noise input bandwidth parameter'. It is defined by equation (9) and plays a vital part in the concluding section of the paper. By coincidence it is equal to B_c for filter type (a).

B_3, the familiar '3 dB bandwidth', the difference between the two frequencies at which $S(f) = S_c/2$.

Expressions for B_c, B_i and B_3 may be obtained by carrying out the mathematical manipulations.[9] Results are presented in Table 2.

Table 2

	Filter type			
	(a)	(b)	(c)	(d)
B_0	f_0	f_0	f_0	$\frac{\pi}{\sqrt{2}}f_0$
B_1	$f_0(=B_c)$	$\sqrt{2}f_0$	$2f_0$	$\frac{4\pi}{3\sqrt{2}}f_0(=\frac{4}{3}B_c)$
B_3	$f_0(=B_c)$	$0{\cdot}94f_0$	$\frac{2}{\pi}f_0$	$2f_0(=0{\cdot}9B_c)$

The remarkable feature about this Table is the small amounts by which the different kinds of bandwidth differ. It may be concluded that for all types of narrow-band filter encountered in practice, the use of the convenient B_3 as a substitute for B_c or B_i will not lead to gross errors in planning an experiment.

Manuscript received by the Institution on 12th October 1971. (Paper No. 1432/CC117.)

U.D.C. 621.373.421 + 621.375.13

The History of Positive Feedback:

The Oscillating Audion, the Regenerative Receiver, and other applications up to around 1923

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

SUMMARY

Positive feedback, or regeneration, played an important part in radio engineering during the two or three decades following its development for electronic circuits around 1912–15. This paper reviews the earlier history of the subject, considers the controversial inventions of 1912–15 in some detail, and goes on to examine some other applications of positive feedback in self-oscillating detectors, e.g. the autodyne, the homodyne and the super-regenerative receivers.

* *Department of Electronic and Electrical Engineering, University of Birmingham, Birmingham B15 2TT.*

Contents

1. Introduction

1.1. *General Observations*

Positive feedback, as something deliberately intended, is nowadays of much less significance than negative feedback, which forms the basis of control systems. In terms of mechanical systems, negative feedback in the form of governors was important long before positive feedback was recognized either implicitly or explicitly. But in electronic circuits it was the other way round; positive feedback for a couple of decades from 1912 reigned supreme, and negative feedback was something 'invented' for electronic systems around 1930.

In positive feedback part of the output signal is fed back to the input in such a way as to reinforce the signal in the input. This has the effect of increasing the gain of an amplifier, and, by reducing the damping of any tuned part of the amplifier, increases the sharpness of tuning (i.e. reduces the bandwidth). If the feedback is sufficient, the oscillations in the system will become self-sustaining, the input signal can be removed, and an oscillation-generator is obtained. The liability to break into self-oscillation if parameters of the system change is one of the difficulties of using positive feedback for amplification. In the early days of radio, the ability to obtain a high sensitivity was so important that the difficulties of positive feedback (usually then called reaction or regeneration) were not regarded as serious.

As in many other aspects of electronic and radio engineering, ideas of feedback developed without much regard to theory and knowledge already existing in other fields of science and engineering. Therefore, although we must start our account of the history of positive feedback with a consideration of instability and oscillation in mechanical governors, it is nevertheless true to say that this topic had no influence on the development of electronic systems.

1.2. *Outline of the Scope of this History*

Early governor and control systems used *negative* feedback in a usually simple mechanical form in which instability was unlikely. It is therefore improbable that *positive* feedback was ever recognized before the second half of the 19th century, when Clerk Maxwell drew attention to it (although not by that name), and analysed it in terms of instability in governors.

In the early years of the 20th century telephone repeaters (i.e. amplifiers) were made essentially of telephone receivers coupled to microphones, and it was known that when the output was coupled to the input the repeater could 'howl', i.e. generate self-oscillations.

The development of the audion, the first thermionic triode, by de Forest from 1906 onwards, led to the development of electronic amplifiers, and eventually positive feedback was 'invented' in relation to such amplifiers from 1911 onwards. Two applications emerged; one, the generation of oscillations, and the second, the increase in amplification produced by feedback below the critical amount required to produce oscillation. Both these applications had enormous commercial value in the rapidly-growing field of radio communication, and there were in consequence repeated and fierce legal battles over priority of invention, mainly between de Forest and Armstrong.

From the idea of positive feedback, otherwise known as reaction or regeneration, Bolitho and Armstrong went on to develop the super-regenerative receiver, while others used the positive feedback in different kinds of self-oscillating detectors represented by the two main systems: the autodyne and the homodyne.

Well within a decade of the invention of feedback electronic oscillators, the non-linear nature of the oscillation process in real circuits was recognized and examined mathematically, notably by van der Pol.

With the growth of better amplifying devices, the use of positive feedback to increase amplification, with its attendant risk of instability, gradually disappeared. However, the use of positive feedback to produce controlled self-oscillation is still important.

It is hoped that the present review of the history of positive feedback, viewed from a distance of over half a century, will put the matter in reasonable perspective.

2. Mathematical Concepts of Instability in Governors

It may come as a surprise to some that the famous James Clerk Maxwell, so well known for his work in electromagnetism, should have published[1] a classic paper 'On governors' in 1868. In introducing his analysis, Maxwell says:

> 'It will be seen that the motion of a machine with its governor consists in general of a uniform motion, combined with a disturbance which may be expressed as the sum of several component motions. These components may be of four different kinds:
>
> 1. The disturbance may continually increase.
> 2. It may continually diminish.
> 3. It may be an oscillation of continually increasing amplitude.
> 4. It may be an oscillation of continually decreasing amplitude.
>
> 'The first and third cases are evidently inconsistent with the stability of the motion; and the second and fourth alone are admissible in a good governor. This condition is mathematically equivalent to the condition that all the possible roots, and all the possible parts of the impossible roots, of a certain equation shall be negative.
>
> 'I have not been able completely to determine these conditions for equations of a higher degree than the third; but I hope that the subject will obtain the attention of mathematicians.
>
> 'The actual motions corresponding to these impossible roots are not generally taken notice of by the inventors of such machines, who naturally confine their attention to the way in which it is *designed* to act; and this is generally expressed by the real root of the equation. If, by altering the adjustments of the machine, its governing power is continually increased, there is generally a limit at which the disturbance, instead of subsiding more rapidly, becomes an oscillating and jerking motion, increasing in violence till it reaches the limit of action of the governor. This takes place when the possible part of one of the impossible roots becomes positive'.

Maxwell does not specifically relate this condition to positive feedback as such, but it can be seen from his mathematical analysis that in a typical system instability takes place when the time constants and the control sensitivity (i.e. loop gain) are high enough—in other words when the feedback is positive and large enough. So the role of positive feedback in generating oscillations is at least implicitly established by 1868, even though it is not yet explicit.

In passing, it is interesting to note Maxwell's modesty in not regarding himself as a mathematician, and his gentle reproof of inventors for insufficient attention to the real modes of operation of their machines.

3. Positive Feedback and Oscillations in Electromechanical Systems

It has been stated by many writers that the ability of an amplifying system to oscillate when the output is coupled back to the input was discovered by the users of the early telephone repeaters, which were electromechanical in nature. Although there were several proposals for telephone repeaters just after 1900,[2] the

commonest was that based essentially on a telephone receiver coupled mechanically to a microphone. It was to be expected that at some time the output of one of these devices would get coupled to the input and an oscillation would result (at a frequency where the phase relations were right). Judge Mayer referred specifically to this effect in his judgement in the Armstrong–de Forest litigation and also referred to a book by Miller[3] where it is mentioned.

With this prior knowledge of oscillation, it is perhaps a matter for surprise that feedback was not applied deliberately to make an oscillation generator earlier than it was. Its use in increasing the gain of an amplifier was rather more subtle and was more understandably delayed.

4. Positive Feedback in Electronic Circuits, 1911–1915

4.1. *The General Situation*

The story of the invention (or discovery), understanding and development of the idea of positive feedback in electronic circuits is a very complicated and confused one. Because of the commercial importance of the idea, there was a great deal of patent litigation over it, and a great deal of public attention was drawn to it. It is one of the famous (or notorious!) cases of engineering history, and consequently much has been written about it. Practically every book which discusses the history of radio[4] gives some account of it; yet there is little consistency among the various accounts (except where one is copied directly from another, as happens occasionally), and all seem to be inaccurate or incomplete to a greater or lesser extent.

The question of who really was the inventor of positive feedback is not now of much importance. The word 'invention' is itself not entirely appropriate for positive feedback; 'discovery' would be better. The legal wrangles were concerned with awarding patent priority in the United States, and therefore considered only inventors who had filed patents in the United States. Scant (if any) attention was given to inventors who had either filed no patent, or had failed to extend their patent cover to the United States. Nevertheless, the judgement of District Judge J. M. Mayer[5] in 1921 is one of the most interesting accounts of the invention available. This judgement awarded priority to Edwin H. Armstrong, but was subjected to several appeals[6] and counter-appeals, finishing in the Supreme Court of the U.S.A., which in 1934 awarded priority to Lee de Forest.

Whatever the legal judgements were at any time, the radio engineering profession seemed to consider Armstrong the true inventor, or at any rate the man who made the biggest contribution to the development of the reaction circuit and the generation of oscillations. In 1917 (or at any rate, for the year 1917) the Institute of Radio Engineers (U.S.A.) awarded its first Medal of Honor to Armstrong 'in recognition of his work and publications dealing with the action of the oscillating and non-oscillating audion.'[7] In 1934, following the Supreme Court's award of priority to de Forest, Armstrong wished to return the medal to the Institute. The Board of the Institute, however, insisted on his keeping it, with the words[8]

> 'That the present Board of Directors, with full consideration of the great value and outstanding quality of the original scientific work of yourself and of the present high esteem and repute in which you are held by the membership of the Institute and themselves, hereby strongly reaffirms the original award, and similarly reaffirms the sense of what it believes to have been the original citation.'

Armstrong, as is well-known, went on to make many other important inventions and developments, e.g. super-regeneration and frequency-modulation, and his reputation greatly increased; he was awarded the Edison Medal of the American Institute of Electrical Engineers in 1942.[9]

The men named by various authors as having been concerned in the invention of positive feedback are (in alphabetical order):

Edwin H. Armstrong (U.S.A)	Eugen Reisz (Austria)
Lee de Forest (U.S.A.)	Henry J. Round (U.K.)
Charles S. Franklin (U.K.)	Wilhelm Schloemilch
Irving Langmuir (U.S.A.)	Siegmund Strauss (Austria)
Charles V. Logwood (U.S.A.)	H. B. Van Etten (U.S.A.)
Fritz Lowenstein (U.S.A.)	Georg von Arco (Germany)
Alexander Meissner (Germany)	Otto von Bronk

Of these, we can dismiss from further consideration, on the grounds that their patents, cited in the literature, show no use of, or knowledge of, positive feedback, the following:

Langmuir[10] Schloemilch and von Bronk.[11]

We have also to dismiss Strauss, as the only citation[12] gives no detailed reference, and the author has so far found no other information on him or his work.

Logwood and Van Etten were assistants to de Forest, and inventions and discoveries attributed to de Forest may well be actually due to his assistants.

Meissner and von Arco appear jointly in patents, but it is believed that Meissner was the inventor, and thus we shall refer no more to von Arco.

4.2. *The Technical Background*

By 1912 the audion, as the triode thermionic valve was then known, was six years old, having been invented in 1906 by de Forest.[13] In passing we may note that a bitter legal battle was fought over priority of invention here too, for Fleming held that his invention[14] of the thermionic diode in 1904 gave him priority in all developments of the thermionic valve.

The audion was not well understood—it seems clear from his own writing that de Forest did not understand it—but it did develop slowly. It was Armstrong who first put the design of circuits using it on a sound quantitative basis by introducing the idea of characteristic curves and oscillographic examination of waveforms.[15,16] The gain of a single audion stage was low, so the invention of cascade or multi-stage circuits was important.[17] The use of negative grid-bias was patented by Lowenstein in 1912,[18] and the grid-leak and

capacitor by Round in 1914[19], although demonstrated by him in 1913.[20]

With this amount of development going on, with the need for ever-greater amplification and the need for better sources of high-frequency oscillations than the spark, the Poulsen arc and the rotating alternator, it is really not surprising that amplifiers were found to oscillate and that feedback should be discovered around 1912–13.

Against this background we shall now examine the contribution of the main inventors, Armstrong, de Forest, Franklin, Lowenstein, Meissner, Reisz and Round.

4.3. *Armstrong*

Armstrong's patent application[21] for his feedback amplifier was filed on 29th October 1913. The specification is rather a complicated one, but basically it relates to an amplifier–detector in which the detection sensitivity is increased by positive feedback of the *audio* signals in the wing (i.e. anode) circuit to the grid circuit.

Professor Edwin H. Armstrong
1890–1954.

Coupling between anode and grid circuits is generally by means of a common reactance in the cathode circuit which is arranged to be effectively a short-circuit at the radio frequencies. Thus only the rectified and smoothed (i.e. audio) signals are fed back. However, Armstrong does also envisage r.f. feedback. Two claims (Nos. 9 and 15) make these two types of feedback quite clear:

> '9. An audion wireless receiving system having a wing circuit interlinked with a resonant grid circuit upon which the received oscillations are impressed, and an inductance through which the current in the wing circuit flows, the grid circuit including connections for making effective upon that circuit the potential variations resulting from a change of current in the wing circuit.'
>
> '15. An audion wireless receiving system having a wing circuit interlinked with a resonant grid circuit upon which the received oscillations are impressed, and means supplementing the coupling of the audion to facilitate transfer of energy from the wing circuit to the grid circuit, whereby the effect upon the grid of high frequency pulsations in the wing circuit is increased.'

The matter is made much clearer in his paper of 1915, where indeed he shows a circuit[22], redrawn here as Fig. 1, which provides specifically for feedback at both radio and audio frequencies, using a separate transformer for each.

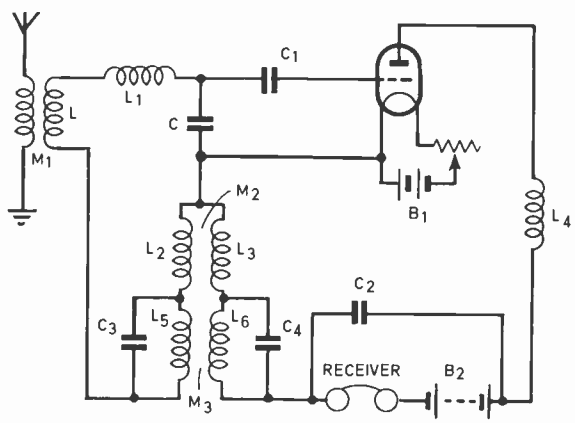


Fig. 1. Armstrong's circuit for combined r.f. and audio feedback.

> 'Here M_2 represents the coupling for the radio frequencies and the coils are of relatively small inductance. M_3 is the coupling for the audio frequencies, and the transformer is made up of coils having an inductance of the order of a henry or more. The condensers C_3 and C_4 have the double purpose of tuning M_3 to the audio frequency, and of by-passing the radio frequencies. The total amplification of weak signals by this combination is about 100 times, with the ordinary bulb. On stronger signals, the amplification becomes smaller as the limit of the audion's response is reached.'

There is no room for doubt that Armstrong understood positive feedback and saw its main importance in terms of amplification. He was perfectly aware, however, that it could generate high-frequency oscillations, for in his patent specification he says:

> 'I find that the audion is made more stable and shows less tendency to become a high frequency generator and to set up oscillations in the interlinked circuits, if the tuned grid circuit is grounded....'

In the Armstrong *v.* de Forest litigation, the courts were at pains to establish which of them first discovered feedback in this context, not who first patented it. This meant that the whole story of how Armstrong came to work on the topic had to be stated, and Judge Mayer gives an excellent account of this in his judgement of 1921.[23] In brief, Armstrong started radio experimenting at the age of 15, and was working on the audion amplifier in a scientific manner while still an undergraduate at Columbia University, where he graduated at 22 in 1913. By the autumn of 1912 he had made an audion receiver of extreme sensitivity, and demonstrated it during the winter without disclosing his circuit. He had insufficient money to file a patent, but on 31st January 1913 he had a drawing of the circuit witnessed by a notary. Judge

Mayer therefore established the date of invention as 'at least as early as January 31st, 1913'.

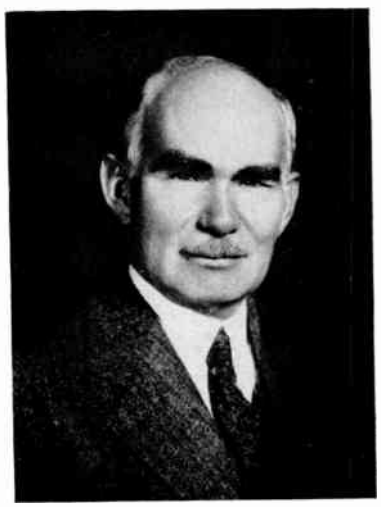
Armstrong's patent was quickly taken up commercially, licences being granted to the Atlantic Communication Co., the Goldschmidt Co. and the Marconi Co.

One point of interest is that Armstrong did not mention (in his patent or in his 1915 paper) the sharper tuning provided by feedback; the emphasis was entirely on increasing the amplification. The concept of positive feedback reducing the damping of the circuit apparently escaped him.

4.4. *de Forest*

In contrast to Armstrong's very professional and scientific approach to radio, de Forest appears almost as a fumbling amateur. In his patent specifications as in his published papers, he shows little understanding of what he is doing. Judge Mayer was not impressed by him:

> 'On the one side is...Armstrong...and he produces a sketch which is extraordinary for its clear and unmistakable description...
>
> 'On the other side is a then experienced and able worker... who is unable to rely solely on notebook entries which are not clear but require construing and who supplements these entries by recollection which is fallible and not certain.'



Photograph Science Museum, London

Dr. Lee de Forest
1873–1961.

As far as patents and publications are concerned, de Forest showed no explicit use of positive feedback until 26th March 1915 when he filed a patent[24] showing in one figure a feedback coil in a mercury-valve circuit (merely as an option to enhance the oscillation) and 13th May 1915, when Fig. 2 of his patent filed that day[25] showed a rather casual use of a feedback transformer:

> 'If desired the grid and plate circuit may be inductively associated with the plate and filament circuit through inductance L, as shown.'

In fact L was a transformer. 'Plate' is, of course, the anode. The circuit was that of an 'oscillation generator' in an aerial circuit and is redrawn in Fig. 2 of the present paper. De Forest used rather peculiar electrode connexions, as can be seen. The oscillation current in the aerial could be keyed or modulated by a microphone as shown. Actually de Forest used an audion with two grids connected in parallel and two anodes also in parallel, but this could not have been fundamental to the operation.

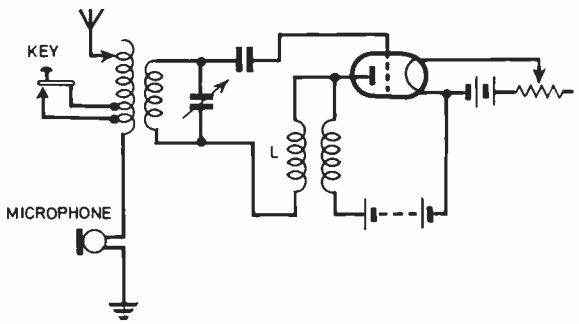


Fig. 2. De Forest's circuit of 1915 for oscillating audion.

It is clear that at this stage, de Forest still did not use feedback explicitly. In most of his circuits he obtained oscillation because of the anode-grid capacitance associated with inductive loads elsewhere—but he never recognized this; not at any rate until many years later. In his published paper of 1914[26] he similarly does not mention feedback. Note that he had earlier patents on the oscillating audion[27,28], but these had no explicit feedback.

Judge Mayer regarded de Forest's patent of September 1915[29] as the first showing feedback claimed as such.

In the light of all this the evidence produced in the patent suits that de Forest had invented feedback in 1912 seems extraordinary. In the Mayer judgement, reference is made to work done by de Forest's assistant Van Etten, who kept a notebook.

> 'Van Etten was working on the audion as a telephone relay and amplifier. On July 23rd (1912), he entered in this notebook a two-way telephone circuit using two audions...they did not have two audions available and on July 24th Van Etten attempted to set up this two-way circuit of his with a single audion having two grids and two plates. The arrangement did not work....
>
> 'The following day, August 6th, he connected the input circuit of a double audion to the output circuit and in this accidental way found that the audion would howl or sing....
>
> 'From this point on his notes show a continuing attempt to produce a telephone relay circuit in which the audion would not howl and in this effort he succeeded on or about August 29th, 1912.
>
> 'The August 6th, 1912, entry of Van Etten was not copied in the laboratory notebook nor was anything done which showed that any one appreciated the phenomenon. De Forest's testimony on the point is not sufficiently clear and definite to be satisfactory.'

A photograph of the Van Etten notebook entry is reproduced in an article[30] on a later suit, which also reproduces a sketch made by Dr. John Stone (American Telephone and Telegraph Co.) of a feedback oscillator circuit shown him by de Forest two years earlier in 1913.

The fact that successive courts so often alternated in their decisions on the case (an oscillation not due to feedback?) is sufficient indication of how contradictory was the evidence.

We should note before leaving de Forest that in 1950 he published a long autobiography[31] in which, as would be expected, there is substantial discussion of the feedback invention, one chapter being devoted to it and another to the litigation over it. It is very fascinating reading. We can, of course, hardly regard this autobiography as objective history and consequently shall not quote from it. Sufficient is it to say that he portrays Judge Mayer (and other judges who ruled against de Forest) as completely prejudiced in Armstrong's favour and unconcerned with justice, he speaks of Armstrong as an enemy, and explains how the case had to go before the Supreme Court twice. Perhaps one can understand how, in a case involving so many millions of dollars, feeling ran so high and one may suppose that Armstrong also had strong feelings.

4.5. *Franklin*

After the confusion of de Forest's work it is refreshing to turn to that of Charles S. Franklin of the British Marconi's Wireless Telegraph Company. His patent specification[32] of 12th June 1913 describes a regenerative amplifier thus:

> '...we make the circuit, in which the magnified oscillations occur, react on the circuit, in which the oscillations to be magnified occur, by coupling these circuits, either electrostatically or electromagnetically, to a certain degree.
>
> 'If the coupling be too strong the tube will be unstable and will itself tend to produce oscillations but there is a certain critical strength of coupling below which the tube is unable to maintain oscillations. At a coupling a little below this critical strength the tube and circuits are stable but act while receiving oscillations as though the resistance in the circuits was very small.
>
> 'The result is that the damping of the receiving system can be reduced to any required degree and the tuning of the system is made very sharp'.

Franklin's circuit is redrawn in Fig. 3 and can be seen to show the feedback quite explicitly by a special loop connexion.

It is not quite certain that Franklin invented positive feedback (as such) independently, for he had apparently visited Meissner in Germany not very long before[33] and had learned that Meissner had succeeded in making a thermionic valve oscillate continuously. However, as we shall see, Meissner appeared to be concerned only with the generation of oscillations, and it is more than likely that Franklin thought out for himself the other useful application of feedback to make sensitive radio receivers.

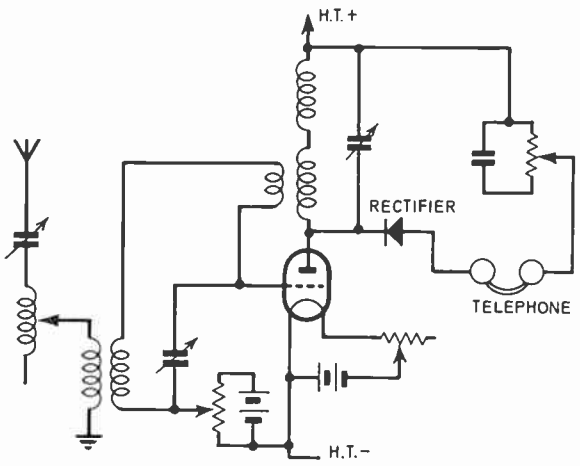


Fig. 3. C. S. Franklin's regenerative receiver of 1913.

It is interesting that Franklin is the only one of the inventors we are considering who mentions the effect of feedback in reducing the damping and sharpening the tuning. Yet this was an important property of the system and could be both a benefit and a nuisance in practice—a benefit in making the receiver more selective, and a nuisance in making it difficult to maintain a receiver in tune for a long period.

The idea of resistance causing damping in a resonant circuit was, of course, quite old even in 1913. As far back as 1853, Kelvin[34] had developed the equations for what was effectively an R, L, C circuit in studying the discharge of a capacitor through a conductor. He established that the energy of the charged capacitor at any instant during discharge was partly dissipated as heat (in the resistance) and partly conserved as current energy in the circuit. Blanchard[35] gives a full discussion of the history of electrical resonance, so we need only say here that the first application of resonance to obtain selectivity in radio receivers was by Sir Oliver Lodge[36] in 1897. The idea that an amplifier could be used to reduce the resistance or damping of a tuned system and so sharpen the tuning does seem to be due to Franklin, and it is an important concept.

The author has so far been unable to establish when Franklin started work on positive feedback. His patent gives every appearance of having been based on experimental trials, and it is possible he was as early as Armstrong; his patent was much earlier than Armstrong's. He did not, however, take out a United States patent (as far as the author can trace) and this accounts for his omission from the U.S. litigation.

4.6. *Lowenstein*

The evidence concerning Lowenstein's contribution to the invention or development of positive feedback is both intriguing and unsatisfactory—intriguing because his work appears to be the very earliest of all work on the oscillating audion, and unsatisfactory because there is no record by Lowenstein himself nor any proper documentary evidence of what he did. All that we have are recollections by his contemporaries, recorded long

afterwards. But these do add up to a fairly strong case for believing that Lowenstein had a reliable audion oscillator towards the end of 1911 or early 1912.

The recollections concerned are contained in a long article by Hammond and Purington,[37] a long discussion on this by Espenschied,[38] and in a book by Miessner.[39] All three of these are mainly subjective or personal views of the early history of radio by people who were themselves very much involved in making that history and they are very far from being impartial; indeed they are extremely controversial. There is agreement among them, however, that Lowenstein had an audion oscillating by the end of 1911, and that his circuit and arrangements are not recorded or patented.

Espenschied says:

> 'One wonders too that the evidence of Lowenstein's having the audion as an oscillation generator during the winter of 1911–1912 was not presented to the courts in the long de Forest–Armstrong litigation over the oscillating tube. Such evidence would have demonstrated the natural tendency of an amplifier to oscillate and hence how little invention there was in the oscillating audion *per se* once it had become an amplifier.'

One may question the philosophy of this remark. Was there really only very little invention in realizing that positive feedback was necessary for oscillation, and working out how best to arrange it? Certainly one can appreciate that invention *was* involved in the feedback or regenerative amplifier, as distinct from the oscillator, and there is nothing fortuitous or fumbling in the patents of Armstrong and Franklin.

4.7. *Meissner*

Alexander Meissner worked in Berlin with a slightly different type of valve, the Leiben–Reisz valve, but like the audion, it was a triode and his circuits are therefore comparable. There seems no doubt that he had a working oscillator by at latest the beginning of 1913, and his German patent[40] is based on several applications from 10th April 1913 onwards. His circuit shows that his positive feedback was quite deliberate; it is redrawn in Fig. 4. There seems no provision here for using it as an amplifier. However, in British and United States patents[41,42] filed early in 1914, Meissner (and his co-patentee von Arco) gives a masterly treatment of the applications of positive feedback in oscillators, amplifiers, amplifier-detectors and heterodyne receivers. For scope and adequacy of understanding, these descriptions are the best of the whole lot dealing with positive feedback; only sharpness of tuning seems to be omitted.

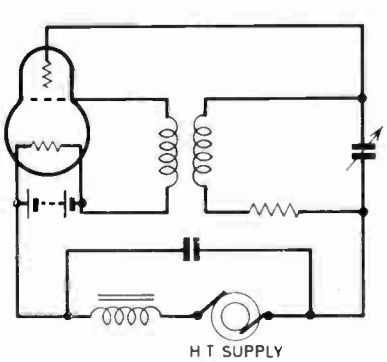


Fig. 4. Meissner's feedback oscillator circuit.

Without further evidence one cannot tell how priority should be allocated between Meissner and Franklin; since they are known to have discussed the subject together, it is possible their inventions were not strictly independent of each other.

4.8. *Reisz*

Eugen Reisz, of Austrian nationality but working in Germany, filed a patent in the U.S.A. on 9th April 1913 in which he showed an amplifier provided with a feedback connexion by means of an anode–grid transformer coupling. There is no doubt that he intended this to be a regenerative amplifier, for he says that when the phases are right

> 'then the amplitudes of the currents reinforced by the relay [i.e. valve] always become greater.'

Judge Mayer said that this patent did not call for discussion in relation to the Armstrong *v.* de Forest priority dispute. Perhaps on some legal grounds he may have been right, but on technical grounds Reisz's patent seems very relevant. Admittedly it was presented in terms of an amplifier for telephone lines and not for a radio receiver, but this was a trivial distinction and the patent covered other applications in general terms.

4.9. *Round*

Although H. J. Round is mentioned in some of the texts as an inventor of positive feedback, it seems from his patent specifications that he was rather an inventor of new systems which involved positive feedback as a prior art. His earlier relevant patent[44] seems to have been filed in December 1913, and this covers what later became known as the autodyne method of reception, using a feedback valve circuit for the local oscillator. His claims do not involve the feedback, but only the method of reception and the circuit detail. A somewhat later patent[45] includes a feedback oscillator as part of a transmitter.

4.10. *Conclusions regarding the Invention of Positive Feedback in Electronic Circuits*

From the account of the work of the various inventors given above, it seems reasonable to conclude that:

(a) The self-oscillating audion was discovered in late 1911 by Lowenstein.

(b) The fact that feedback could produce oscillation in an audion was discovered by Van Etten in August 1912.

(c) The *use* of feedback as such was invented early in 1913 by Armstrong, Reisz, Meissner and Franklin, all within a month or two of one another.

(d) The concept of feedback affecting the damping of the resonance and sharpening the tuning was due to Franklin in 1913.

It is interesting to speculate on the soundness of the principles of law which led to the award of the priority (and hence profits) of invention to de Forest who did not 'invent' feedback (although his assistants may have done so in a limited sense) and who did not understand, exploit or patent the principle until long after several others did understand, did exploit and did patent it.

5. Self-oscillating Detectors: Autodyne and and Homodyne

In the development of feedback amplifiers and oscillators, some special properties and advantages of oscillating receivers were noticed and developed. These followed the ideas of heterodyne reception, which had been invented by R. A. Fessenden in 1902[46] but not used extensively in pre-audion days. The burst of new development produced by the valve circuits led to a great deal of concurrent theoretical analysis of heterodyne reception[47–50] and before long to the invention by Armstrong [51] of the 'superheterodyne' (as it later became called) in which a signal which is of too high a frequency for direct amplification is changed to a lower frequency where amplification can readily be provided. However, the oscillating receivers to which reference was made above are the *autodyne* and the *homodyne*—to use names which were introduced a little later.

The autodyne may be defined as a receiver in which self-oscillations are generated at a frequency different from that of the incoming signal; these oscillations heterodyne with the signal in the non-linearity of the receiver to give a beat (or 'heterodyne') tone at a convenient frequency. The system was used basically for radio-telegraphy, and the beat tone was thus in the audio range, to be heard on headphones.

The homodyne may be defined as a receiver in which self-oscillations are generated at the same frequency as that of the incoming signal; provided the phase-relationships are correct, interaction in the non-linearity of the receiver gives an audio output corresponding to the modulation on the incoming signal. This system is therefore useful for receiving speech transmissions.

5.1. *The Autodyne*

As with the basic concept of positive feedback, so with the autodyne there was a spate of patents from different inventors all at more or less the same time—in this case, however, even more closely spaced.

The autodyne depends, of course, on the receiver having a continuous self-oscillation to beat with the incoming signal. It is clear, therefore, that the development of the autodyne was dependent on not only the audion valve itself, but also on feedback to make it oscillate. It is thus particularly interesting that the basic idea of the autodyne was first put forward, by H. J. Round,[52] in terms of a circuit which not only had no feedback but also had no audion! The basic circuit of this is shown in Fig. 5; here C is a crystal rectifier. The aerial is out-of-tune with the incoming waves by a frequency difference f, say. Circuit B is 'preferably aperiodic'. Each burst of signal (telegraphic) excites a transient oscillation in the input tuned system of frequency equal to the natural frequency of resonance, i.e. at a frequency differing from the signal by f. The rectifier therefore produces a beat note of f, and circuit D is tuned to f so that a tonal pulse of this frequency is heard in the telephone receiver.

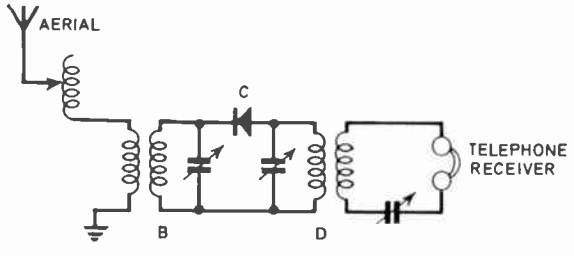


Fig. 5. Round's anti-atmospheric receiver showing the principle of the autodyne.

Marconi photograph

Mr. Charles S. Franklin, C.B.E., (1879–1964) and Captain Henry J. Round, M.C., (1881–1966). Taken in 1962 during the making of a film on their contributions to radio engineering.

Round's object in designing this system was to provide immunity from 'atmospherics' which comprise bursts of interference of very short duration. The idea is that such interference cannot produce the beat tone f of sufficient duration to be recognized as a tone in the telephone receiver. But Round says that 'with such a simple arrangement however it may be difficult to get the beat frequency unaffected by atmospherics and yet low enough to be within the limits of audibility' and so he proposed that the frequency difference f be made considerably

greater, and a second heterodyne stage be provided to change f down to an audible frequency, using a local oscillation generator of unspecified type.

The first autodyne using a valve to produce a continuous local oscillation was also due to Round and was mentioned earlier in Section 4.9. It was patented[53] only a few days after the system just described; the basic circuit is shown in Fig. 6. Here the tuned circuits A and E are tuned to slightly different frequencies, and by varying the coupling between them, the self-oscillation may be produced at a frequency suitable for producing an audible beat-note in the telephone receiver.

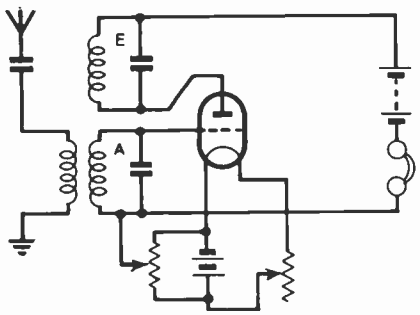


Fig. 6. Round's autodyne receiver.

Round's system was closely followed by a patent by Armstrong[54] only nine days later covering substantially the same idea, except that Armstrong used his preferred method of obtaining feedback by connecting the telephone receiver as a common impedance between anode and grid circuits.

Meissner's patent,[55] already referred to in Section 4.7, includes, among other uses of positive feedback, not only the autodyne receiver, but also a double or reflex autodyne receiver which uses the autodyne principle for both radio-frequency and audio-frequency stages in one and the same valve!

Both Armstrong and Meissner refer to the increased sensitivity of the autodyne system (although they do not use that name for it).

As with positive feedback itself, de Forest enters the autodyne patent scene rather later than the others. The patent[28] of March 1914 by himself and C. V. Logwood shows what seems to be an autodyne receiver, but the explanations are not clear. In an article[56] early in 1915 he describes what is clearly an autodyne under the name 'ultraudion'. The sub-title is significant: 'Modification of the audion which makes of it an extraordinary sensitive instrument for this purpose and offers great possibilities for high-speed wireless records'. The audion is made to oscillate by overheating the filament, it is tuned to be slightly different from the incoming frequency, and the sensitivity is then claimed to be increased by a factor of from 10 to 50. But there is no science here, no explanations.

De Forest gives a more personal account of his invention of the ultraudion in his autobiography[57] and describes how it came to be made.

5.2. *The Homodyne*

It is difficult to think that the homodyne could ever have been a true invention, because the autodyne would obviously have become one whenever the self-oscillation pulled in or synchronized to the incoming signal. However, so long as the autodyne was used only for receiving telegraph signals (which was all it was suited for) this synchronized condition would have been an undesirable nuisance, a condition to be avoided. As Armstrong said in his patent already cited:

> 'So long as the oscillations thus generated in the detector circuit are of the same frequency as the received oscillations no signals will be heard in the telephones....'

He also referred to the need for the coupling between the aerial and the receiving circuit to be extremely loose, probably to avoid pull-in of the oscillation.

Presumably what inventive process there was in the concept of the homodyne lay in the appreciation that it could efficiently demodulate a carrier wave modulated by speech. This seems to have been realized first by by Burton W. Kendall[58] in 1915. It is possible that he approached the invention rather from the point of view that increased sensitivity could be obtained in a receiver if a local carrier were added to reinforce the incoming wave, for this is the first thing mentioned in his specification; but he quickly goes on to describe the self-oscillating detector, which was then the only practicable form of the system. The appropriate circuit arrangement from his specification is shown in Fig. 7.

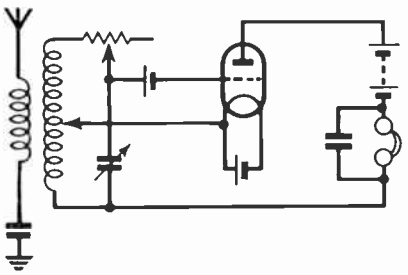


Fig. 7. Kendall's homodyne receiver of 1915.

The homodyne, unlike the autodyne, was the subject of much interest and development later. In 1923, Hartley[59] gave a mathematical discussion of the subject, showing the effect of phase errors in the local oscillation, etc. In an article in 1924 Colebrook[60] described the homodyne substantially as Kendall specified it. From about this time onwards, however, the system developed towards one of greater refinement, with the non-linear oscillation circuit separated from the desirably-linear signal circuit; the name 'synchrodyne' was later coined by the present author. The history of this development from the early 1920s onwards has been separately published[61] and will therefore not be pursued here.

6. Super-regeneration

In addition to the autodyne and homodyne there were many other circuits and systems developed which utilized positive feedback. Many examples are sum-

marized in Blake's book.[4(a)] Here we shall confine ourselves to what was almost certainly the most important—the super-regenerative receiver.

The name was given to the system by Armstrong, who is generally held to be the inventor. It is, however, difficult to see the difference in principle between Armstrong's system[62,63] and the considerably-earlier British one due to J. B. Bolitho.[64,65] A somewhat similar principle was involved in L. B. Turner's 'valve relay'.[66] The general idea is to exploit the extremely high gain which is obtained in a feedback valve circuit which is just on the point of oscillating. At this point, the application of the signal, even of almost infinitesimal magnitude, will produce a substantial amplitude of the oscillation, and this amplitude will depend on the signal amplitude. The problem is, of course, that a valve circuit cannot normally be held just at the point of oscillation. It is normal for the amplitude to build up until limited by non-linear action, at which point the extreme sensitivity to the applied signal has been lost. So the object of all these super-regenerative inventions is to keep the circuit just at the point of oscillation.

This object is achieved by using an auxiliary circuit such that as soon as the oscillation starts to build up it is quenched by a brief alteration of a parameter of the circuit; it then starts to build up again, and so on. If the quenching is caused to occur at intervals which are too close to cause any audible modulation of the output audio signal, but which are sufficiently spaced in relation to the cycles of the radio signal, then the envelope of the oscillation waveform reproduces the modulation (audio) signal with extremely high amplification and usually acceptable distortion. Turner used an electromechanical relay to short-circuit the feedback as the oscillation built up, but this was obviously inadequate for speech. Bolitho and Armstrong used separate valve oscillators to give the periodic quenching, the former by causing the feedback to be largely cancelled as the separate oscillator went through its positive half-cycles, the latter by applying the output of the separate oscillator to the grid of the feedback valve, so that the amplification was cut off on negative half-cycles of the separate oscillator. One of Armstrong's circuits is shown in Fig. 8. Here the circuit A is the radio-frequency amplifier and detector, and circuit B is the separate oscillator of lower frequency.

There were many other versions of the super-regenerative receiver, one even using gas-flames as oscillating grid leaks,[67] and some of these are summarized in Blake's book already cited.

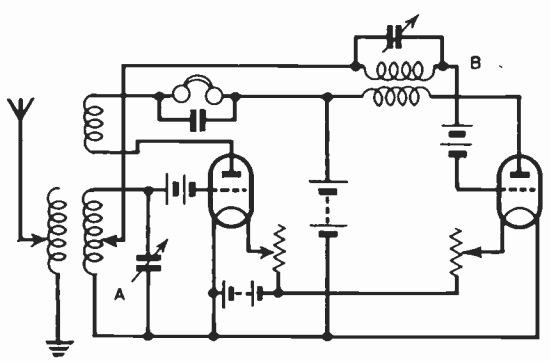


Fig. 8. Armstrong's super-regenerative receiver.

7. The Non-linear Theory of Oscillators and Synchronization of Oscillators

With the increasing use of oscillating electronic circuits, it was to be expected that theoretical and mathematical studies of oscillators should be undertaken. The first of these appears to be a paper by Hazeltine[68] in 1918, which is a very thorough and admirable study, introducing the useful concept that positive feedback is equivalent to a negative resistance, but dealing mathematically with the circuits in purely linear-circuit terms. In his discussion of the behaviour of oscillators in respect to the incoming signal to which they may be coupled in an autodyne circuit, he describes what appears to be the pull-in effect (later called synchronization of oscillators) but his account is not satisfactory on this matter, nor is he (naturally) able to analyse it properly with his linear representations.

The recognition that the non-linear characteristic of the valve dominates the study of oscillating circuits was probably due to B. van der Pol,[69] and another early study was made by E. V. Appleton,[70] partly in collaboration with van der Pol. Synchronization of oscillators was recognized as part of this non-linear behaviour.[71,72]

Subsequent work in this field has been prolific and need not be discussed further here.

8. Conclusions

We have now traced the history of positive feedback from its origins up to the early 1920s. It can be seen that the topic played a vital part in radio development from about 1911 or 1912 onwards, and that numerous nventors were almost simultaneously, and usually competitively, 'inventing' the same thing at each stage. We have made some sort of judgement on the priorities (in a technical rather than legal context) in the basic inventions of positive feedback in electronic circuits, and these give no place to de Forest, who nevertheless was awarded priority by the U.S. Supreme Court.

9. References

1. Maxwell, J. C., 'On governors', *Proc. Roy. Soc., London*, 16, pp. 270–283, 1868. Reprinted in 'Selected Papers on Mathematical Trends in Control Theory', ed. Bellman, R. and Kalaba, R. (Dover Publications, New York, 1964).
2. These are briefly summarized in Tucker, D. G., 'The early history of amplitude modulation, sidebands, and frequency-division multiplex', *The Radio and Electronic Engineer*, 41, p. 43, January 1971.
3. Miller, Kempster B. (Unfortunately it has so far proved impossible to locate a copy of this work in any British library and no further reference can be given).
4. See, for example, the following:
   (a) Blake, G. G., 'History of Radio Telegraphy and Telephony', p. 260. (Radio Press, London, 1926).
   (b) Archer, G. L., 'History of Radio to 1926', p. 113. (American Historical Society, New York, 1938).

(c) Maclaurin, W. R., 'Invention and Innovation in the Radio Industry', pp. 77, 106, 119. (Macmillan, New York, 1949).

(d) Dunsheath, P., 'A History of Electrical Engineering', p. 277. (Faber and Faber, London, 1962).

(e) Sharlin, H. I., 'The Making of the Electrical Age', p. 111. (Abelard-Schuman, New York, 1963).

(f) Finn, B. S., 'Electronic communications' in 'Technology in Western Civilization', ed. Kranzberg, M. and Pursell, C. W., p. 300. (Oxford University Press, New York, 1967).

(g) Baker, W. J., 'A History of the Marconi Company', p. 151. (Methuen, London, 1970).

5. For a long extract of the important parts see 'The "feedback" or "regenerative" valve circuit', *Radio Review*, **2**, p. 424, 1921.
6. e.g., Arwin, W. B., 'Dr. Lee de Forest and reaction patents', *Wireless Weekly*, p. 168, 11th June 1924.
7. See *Proc. Inst. Radio Engrs*, **7**, p. 95, 1919.
8. See *Proc. Inst. Radio Engrs*, **22**, p. 812, 1934.
9. See *Electrical Engineering*, **62**, p. 147, 1943.
10. U.S. Patent No. 1,282,439 filed 29th October 1913, issued 22nd October 1918.
11. U.S. Patent No. 1,087,892 filed 14th March 1913, issued 17th February 1914.
12. Finn, B. S., *loc. cit.*
13. de Forest, L., 'The audion', *Trans. Amer. Inst. Elect. Engrs*, **25**, p. 735, 1906.
14. British Patent No. 24,850 filed 16th November 1904.
15. Armstrong, E. H., 'Operating features of the audion: explanation of its action as an amplifier, as a detector of high-frequency oscillations and as a "valve"', *Electrical World (New York)*, **64**, p. 1149, 12th December 1914.
16. Armstrong, E. H., 'Some recent developments in the audion receiver', *Proc. Inst. Radio Engrs*, **3**, p. 215, 1915.
17. Langmuir, I., U.S. Patent No. 1,282,439 filed 29th October 1913, issued 22nd October 1918.
18. Lowenstein, F., U.S. Patent No. 1,231,764 filed 24th April 1912, issued 3rd July 1917.
19. Marconi W.T. Co. and Round, H. J., British Patent No. 13,248 filed 29th May 1914, issued 27th May 1915.
20. Baker, W. J., *loc. cit.*, p. 151.
21. U.S. Patent No. 1,113,149 filed 29th October 1913, issued 6th October 1914.
22. Armstrong, E. H., 'Some recent developments in the audion receiver', *Proc. Inst. Radio Engrs*, **3**, p. 215, 1915. (See Fig. 14 on p. 223.)
23. *Radio Review*, **2**, pp. 424–430, 1921.
24. U.S. Patent No. 1,221,034 filed 26th March 1915, issued 3rd April 1917.
25. U.S. Patent No. 1,201,273 filed 13th May 1915, issued 17th October 1916.
26. de Forest, L., 'The audion—detector and amplifier', *Proc. Inst. Radio Engrs*, **2**, p. 15, 1914.
27. U.S. Patent No. 1,201,270 filed 14th September 1914, issued 17th October 1916.
28. U.S. Patent filed 12th March 1914, co-patentee C. V. Logwood; also British Patent No. 3,950 filed 12th March 1915, issued 13th March 1916, Convention date 12th March 1914.
29. U.S. Patent No. 1,311,264 filed 4th September 1915, issued 29th July 1919.
30. Arwin, W. B., *loc. cit.*
31. de Forest, L., 'Father of Radio—the Autobiography of Lee de Forest' Chapters 29 and 40, pp. 290–299 and 375–386. (Wilcox and Follett, Chicago, 1950).
32. British Patent No. 13,636 filed 12th June 1913, issued 11th June 1914.
33. Baker, W. J., *loc. cit.*, p. 151.
34. Thomson, W. (later Lord Kelvin), 'On transient electric currents', *Phil. Mag.* **5**, p. 393, 1853.
35. Blanchard, J., 'The history of electrical resonance', *Bell Syst. Tech. J.*, **20**, p. 415, 1941.
36. British Patent No. 11,575 filed 10th May 1897, issued 10th August 1898.
37. Hammond, J. H. and Purington, E. S., 'A history of some foundations of modern radio-electronic technology', *Proc. Inst. Radio Engrs*, **45**, p. 1191, 1957 (see particularly p. 1199).
38. Critique of above by Espenchied, Lloyd, *Proc. Inst. Radio Engrs*, **47**, p. 1253, 1959 (see particularly p. 1255).
39. Miessner, B. F., 'On the Early History of Radio Guidance' see particularly p. 21. (San Francisco Press, 1964).
40. German Patent No. 291,604, application date 'from 10th April 1913 onwards', issued 23rd June 1919, in name of Gesellschaft für Drahtlose Telegraphie m.b.H.
41. British Patent No. 252 filed 5th January 1914, issued 19th August 1915, in name of Graf Georg von Arco and Dr. Alexander Meissner.
42. U.S. Patent No. 1,314,102 filed 10th January 1914, issued 26th August 1919, in name of Georg von Arco and Alexander Meissner.
43. U.S. Patent No. 1,234,489 filed 9th April 1913, issued 24th July 1917.
44. British Patent No. 28,413 filed 9th December 1913, issued 9th December 1914.
45. British Patent No. 13,248 filed 29th May 1914, issued 27th May 1915.
46. Sharlin, H. I., 'The Making of the Electrical Age', p. 114 (Abelard-Schuman, New York, 1963).
47. Hogan, J. L., 'The heterodyne receiving system', *Proc. Inst. Radio Engrs*, **1**, Part 1, p. 75, 1913.
48. Liebowitz, B., 'The theory of heterodyne receivers', *Proc. Inst. Radio Engrs*, **3**, p. 185, 1915.
49. Armstrong, E. H., 'A study of heterodyne amplification by the electron relay', *Proc. Inst. Radio Engrs*, **5**, p. 145, 1917.
50. Howe, G. W. O., 'The amplification obtainable by the heterodyne method of reception', *Proc. Inst. Radio Engrs*, **6**, p. 275, 1918.
51. U.S. Patent No. 1,342,885 filed 8th February 1919, issued 8th June 1920.
52. British Patent No. 27,480 filed 28th November 1913, issued 26th November 1914.
53. British Patent No. 28,413 filed 9th December 1913, issued 9th December 1914.
54. British Patent No. 24,231 filed 17th December 1914, issued 17th December 1915, (Convention date 18th December 1913).
55. British Patent No. 252 filed 5th January 1914, issued 19th August 1915, also U.S. Patent No. 1,314,102, filed 10th January 1914, issued 26th August 1919.
56. de Forest, L., 'The ultraudion detector for undamped waves', *Electrical World, (New York)*, **65**, p. 465, 20th February 1915.
57. de Forest, L., ref. 31, Chapter 33, pp. 316–319.
58. U.S. Patent No. 1,330,471 filed 29th November 1915, issued 10th February 1920.
59. Hartley, R. V. L., 'Relations of carrier and side-bands in radio transmission', *Proc. Inst. Radio Engineers*, **11**, p. 34, 1923.
60. Colebrook, F. M., 'Homodyne', *Wireless World & Radio Rev.*, **13**, p. 645, 1924.
61. Tucker, D. G., 'The history of the homodyne and synchrodyne', *J. Brit. Instn Radio Engrs*, **14**, p. 143, 1954.
62. U.S. Patent No. 1,424,065 filed 27th June 1921, issued 25th July 1922.
63. Armstrong, E. H., 'Some recent developments of regenerative circuits', *Proc. Inst. Radio Engrs*, **10**, p. 244, 1922.

64. British Patent No. 156,330 filed 6th October 1919, issued 6th January 1921.

65. 'The Bolitho circuit', *Wireless World & Radio Rev.*, **12**, p. 266, 2nd June 1923.

66. British Patent No. 130,408, filed 16th February 1918, issued 7th August 1919.

67. Blake, G. G., 'Some suggested lines for experimental research', *Wireless World*, **14**, p. 316, 11th June 1924.

68. Hazeltine, L. A., 'Oscillating audion circuits', *Proc. Inst. Radio Engrs*, 6, p. 63, 1918.

69. van der Pol, B., 'A theory of the amplitude of free and forced triode vibrations', *Radio Review*, **1**, p. 701, 1920.

70. Appleton, E. V. and van der Pol, B., 'On a type of oscillation–hysteresis in a simple triode generator', *Phil. Mag.*, **43**, p. 177, 1922.

71. Appleton, E. V., 'Automatic synchronization of triode oscillators', *Proc. Cambridge Phil. Soc.*, **21**, p. 231, 1922–3.

72. van der Pol, B., 'Forced oscillations in a circuit with non-linear resistance', *Phil. Mag.*, **3**, p. 65, 1927. (Published in Dutch in *Tijdschr. van het Ned. Radiogenootschap.*, October 1924).

Manuscript received by the Institution on 19th July 1971. (Paper No. 1433/CC 118).

STANDARD FREQUENCY TRANSMISSIONS—January 1972

(Communication from the National Physical Laboratory)

Jan 1972	Deviation from nominal frequency in parts in 10^{10} (24-hour mean centred on 0300 UT)			Relative phase readings in microseconds N.P.L.—Station (Readings at 1500 UT)		Jan 1972	Deviation from nominal frequency in parts in 10^{10} (24-hour mean centred on 0300 UT)			Relative phase readings in microseconds N.P.L.—Station (Readings at 1500 UT)	
	GBR 16 kHz	MSF 60 kHz	Droitwich 200 kHz	*GBR 16 kHz	†MSF 60 kHz		GBR 16 kHz	MSF 60 kHz	Droitwich 200 kHz	*GBR 16 kHz	†MSF 60 kHz
1	+0·3	0	−0·1	583	626·7	17	−0·5	−0·2	−0·2	599	638·9
2	+0·2	−0·1	−0·1	581	627·2	18	+0·1	−0·2	−0·2	598	640·0
3	−0·4	−0·1	−0·2	585	624·1	19	−0·3	0	−0·2	601	639·7
4	−0·1	−0·1	−0·2	586	625·3	20	−0·1	−0·1	−0·2	602	633·9
5	0	−0·1	−0·2	586	626·2	21	−0·1	−0·1	−0·2	603	634·8
6	−0·1	−0·1	−0·1	587	626·6	22	0	−0·1	−0·1	603	635·8
7	0	−0·3	−0·2	587	630·3	23	+0·3	−0·1	−0·1	600	643·6
8	−0·1	−0·1	−0·2	588	634·0	24	−0·5	−0·1	−0·2	605	644·3
9	−0·1	−0·1	−0·2	589	635·2	25	−0·3	0	−0·2	608	644·0
10	−0·1	−0·1	−0·1	590	636·1	26	−0·2	−0·2	−0·2	610	—
11	−0·1	−0·1	−0·2	591	635·5	27	—	—	−0·3	—	—
12	−0·1	−0·1	−0·2	592	634·1	28	—	—	0	—	—
13	−0·1	−0·1	−0·1	593	635·2	29	−0·2	0	+0·1	612	643·7
14	+0·1	−0·1	−0·2	592	636·1	30	+0·1	−0·1	0	611	644·2
15	−0·2	−0·1	−0·2	594	636·6	31	−0·4	−0·1	0	615	644·1
16	0	−0·1	−0·2	594	637·4						

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

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

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

U.D.C. 621.317.733:621.382

High-frequency Measurement of Integrated Circuit Components

H. A. KEMHADJIAN,
M.Sc. (Eng.), C.Eng., M.I.E.E.*
and
M. A. FLEMMING,
B.A.*

SUMMARY

A jig for extending a 50 Ω high frequency measuring system to the pads of a chip device is described. No permanent wire bonds are used and contact may be made with normal aluminium pads. The electrical characterization of the jig up to 1 GHz has been carried out using a General Radio transfer function and immittance bridge, and *Y*-parameter measurements on typical chip devices have been obtained. The results are applicable to the design of integrated circuits and to device modelling.

* *Department of Electronics, University of Southampton, Southampton SO9 5NH.*

1. Introduction

The characterization of a high-frequency device may be made in terms of a number of different parameters. This work is described in terms of *Y*-parameters measured on a General Radio 1607A transfer function and immittance bridge, but is equally applicable to other measuring systems of 50 Ω characteristic impedance, such as the *S*-parameter equipments currently available.

The essence of a high-frequency admittance measurement is the determination of the effects of all discontinuities, relative to the characteristic impedance of the measuring system, when viewed from a defined reference plane within the measuring system.

Thus the measurement includes the effects of any parasitic discontinuities introduced between the 50 Ω system and the component whose properties are to be measured. When measuring on integrated circuit components mounted in conventional headers using wire bonding techniques, the parasitics introduced are variable and difficult to isolate; thus extraction of the intrinsic device parameters by calculation gives uncertain results. For example, the inductance of each bonding lead when using a TO-5 header is about 4–6 nH, depending on the length of bonding wire used. Such variations mask intrinsic device variations, particularly in the region of a resonance between the lead inductance and the device capacitance. For the above range of lead lengths and a typical device capacitance of 5 pF, the resonance falls in the range 900–1100 MHz, which is the region of device cut-off frequency, where intrinsic parameter changes are of maximum interest.

In order to eliminate the parasitic elements, it is necessary to extend the 50 Ω measuring system to the pads of the component on its semiconductor chip. The component must be contacted by both conductors of the transmission line in the same plane since a physical separation between the device and the ground plane of the line introduces parasitic common lead inductance. Thus the separation of the conductors must be comparable with the chip dimensions. For a 50 Ω transmission line of chip scale the planar microstrip structure offers a suitable solution. The construction of lines of this type is described below.

So that measurements may be made on the same device in different configurations, permanent bonds must be avoided. High-frequency microstrip packages have been described but require the use of wire bonds for connecting to the device.[1] The measuring jig to be described uses probe contacts which may be pressed against the device pads.

2. Microstripline Design

The requirements for the substrate of the microstripline are uniform thickness with good surface finish and dielectric properties. The material must also be flexible, to allow pressure contact to be made to the device. For the characteristic impedance to be maintained, the strip conductor width must be great enough for its accurate definition. 'Kapton' polyimide film of 5 'thou' (125 μm) thickness meets these requirements.

The material has excellent flexibility and the dielectric constant of 3·55 gives a line width of 285 μm for 50 Ω characteristic impedance,[2] which width is readily defined by a number of methods. The use of a thinner substrate or a material of higher dielectric constant would require a narrower line, giving reduced accuracy of characteristic impedance. The rapid increase of the dielectric loss of polyimide above 1 GHz restricts the frequency range which may be used.[3] 'Polyguide' film is currently being evaluated for higher frequency use.

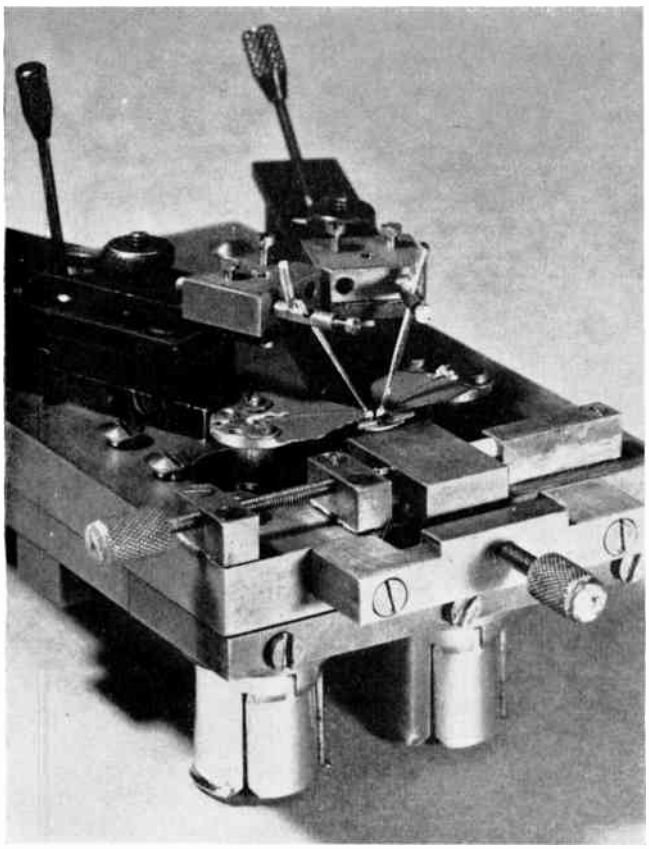


Fig. 1. The microstripline measuring jig.

Vacuum-deposited films were used for the conductors, gold proving the most suitable metal. Thorough cleaning is necessary before evaporation and the polyimide was maintained at 200°C during the evaporation. Subsequent baking at 250°C for several minutes was found to improve adhesion. The strip conductor width was defined by photomechanical etching, the evaporated film being subsequently electroplated to reduce series resistance. Other methods, including the use of out-of-contact masks, were also used but gave less reproducible results. The photomechanical method has an error of less than 5 μm.

'Monotherm' copper-clad polyimide has also been used. The presence of the bonding adhesive makes the dielectric properties inferior. Also, owing to the thickness of the cladding, undercutting occurs during etching which must be allowed for. Capacitance measurements at 900 MHz gave loss factors of 0·01 using evaporated conductors, and 0·016 for the clad material.

3. The Measuring Jig

The measuring jig which carries the chip and the microstripline connectors is designed to plug directly into the G.R. bridge terminals. The minimum possible length of microstripline is used, this being equal to the separation of the bridge terminals which is 30 mm. The calculated total dielectric loss[4] of this length of line just becomes comparable with the total losses in the bridge air lines[5] at 1 GHz. At lower frequencies the microstripline losses are negligible.

The transition into the microstripline is made via coaxial reducers and Sealectro 'Con-Hex' surface mounting stripline launchers. The centre contacts of the launchers are modified to make pressure contact on the microstripline conductor by soldering a short length of gold-plated wire perpendicularly to the centre pin. At the ends of the microstriplines, probes are mounted to contact the chip situated below the lines. The probes are 300–600 μm in length and of 100 μm diameter. The material used is tungsten which may be electrolytically etched in potassium hydroxide solution to give point radii of 1–10 μm. The probes are attached to the lines using silver-loaded Araldite conducting cement. Quartz rods, carried in micromanipulators above the lines, are used to press the probes into contact with the device pads. Pads as small as 30 μm square have been contacted in this way.

At the same time the ground connexion must be made as close to the line tip as possible. This is achieved by placing gold-plated blocks, slightly thicker than the chip, immediately adjacent to the chip edges. Thus, as the probes are pressed down the ground plane contact is made simultaneously.

The chip, which may be on one of the standard types of header or a simple gold block, is mounted on a carriage, the position of which may be adjusted by drive screws. The stripline launchers may also be rotated to further facilitate alignment of the probes with the device pads.

The layout of the complete jig is shown in Fig. 1.

4. Evaluation of the Jig Parameters

The performance of the microstrip launchers was measured using a Hewlett-Packard time domain reflectometer. Small adjustments were made to minimize the reflexion coefficient. The reflectometer trace for a 75 mm section of microstripline mounted between two transitions of the type described above and inserted into a 50 Ω line shows the maximum reflexion to be about 2%. The characteristic impedance of the stripline is accurate to within 1%.

The electrical length of one of the 15 mm microstriplines mounted in the jig was measured using the G.R. bridge. The line was measured with both open- and short-circuit terminations at its tip, the General Radio WO5 and WN5 open- and short-circuit terminations being used as references. The results are given in Table 1.

Table 1. Measurement of jig parameters

OPEN-CIRCUIT STRIPLINE

Frequency (MHz)	Jig length–WO5 (mm)	Admittance with WO5 (mS)	Admittance with jig (mS)
330	11·7 ± 0·4	0·08	0·08
500	11·8 ± 0·2	0·12	0·16
900	11·7 ± 0·2	0·24	0·36

1 mS = 1 millisiemen = 1 $m\Omega^{-1}$

SHORT-CIRCUIT STRIPLINE

Frequency (MHz)	Jig length–WN5 (mm)	Impedance with WO5 (Ω)	Impedance with jig (Ω)
330	11·7 ± 0·4	0·25	0·50
500	11·7 ± 0·2	0·40	0·80
900	11·5 ± 0·2	0·60	1·40

The series resistance measured in the short circuit case comprises both line loss and the lumped impedance of the imperfect short-circuit at the end of the line. This short-circuit was made using a small amount of conducting silver paint. The d.c. resistance of the shorted line was 0·25 Ω but a rise at high frequency is expected owing to the reducing skin depth. Since the length difference between the microstriplines and the standards is constant, when making measurements the bridge may be set up using a standard termination and the appropriate difference added to the line stretcher lengths. The length calibration for the striplines may be obtained in the open circuit condition so that the need for short-circuiting the line is avoided.

The shunt losses are small compared with those inherent in the bridge except at the highest frequencies. When measuring small admittances the standard correction procedure for the bridge may be applied. The accuracy of the system depends on the magnitude of the impedance being measured so that no general figure can be given. Most of the admittances encountered in high frequency device measurements are of the order of 10 mS where the system accuracy is 8%, this being set by the limitations of the G.R. bridge.[5]

The only parasitic element of the jig between the end of the microstripline and the chip component is the tungsten probe. The calculated value for the inductance of this is 0·1–0·25 nH. To obtain a practical value for this inductance, the transadmittance was measured for a pair of lines aligned in the jig with their pins in contact, and with their ground planes connected. Values obtained at frequencies up to 900 MHz were less than 0·4 nH for the pair of probes.

The results show that the reference plane at the end of the microstripline is maintained to 0·2 mm, which equals the best repeatability of the bridge, and that the parasitic discontinuity between this plane and the device pad is an inductance of less than 0·2 nH. The impedance of this inductance at 1 GHz is 1·25 Ω, which is much less than device impedances. Resonance with a 5 pF typical capacitance is above 5 GHz.

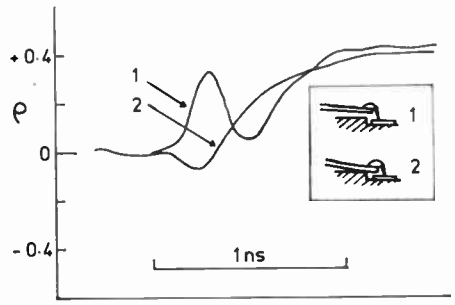


Fig. 2. Time domain reflectometer traces for 100 Ω diffused resistor mounted in the jig.
Curve 1: ground plane contact not made, showing effect of additional inductance.
Curve 2: proper contact made by both probe and ground plane.

5. Measurements on Integrated Circuit Components

The jig has also been evaluated by making measurements on chip components and comparing the results with values obtained by calculation and by measurement in conventional headers.

For these measurements diffused resistors were used. These have simple structure and are thus readily represented by a theoretical equivalent circuit.[6] The resistance and total capacitance of the devices were first measured at low frequency. The devices were mounted on TO-5 headers which can be accommodated on the jig platform and measurements of Y_{11} and Y_{21} were made over the frequency range 50–1000 MHz using the microstrip probes. The reflectometer is useful for ensuring that contact is made by both probe and ground plane. The effect of the inductance present when the ground plane is not connected is shown by curve 1 of Fig. 2. The results for a typical resistor were compared with those obtained when the identical resistor was bonded into the header in the normal way. For the in-can measurements, General Radio TO-5 transistor mounts were used.

The input admittance values shown in Fig. 3 correspond with low frequency values to within 5% at 50 MHz. At high frequencies the elimination of the input port

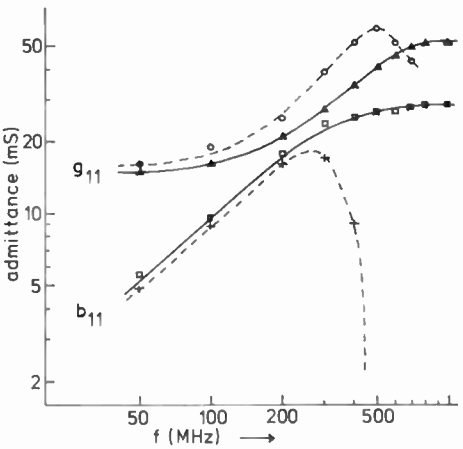


Fig. 3. Input admittance for 70 Ω diffused resistor.
Solid curve: resistor mounted in microstrip jig.
Dashed curve: same resistor bonded into TO-5 can.

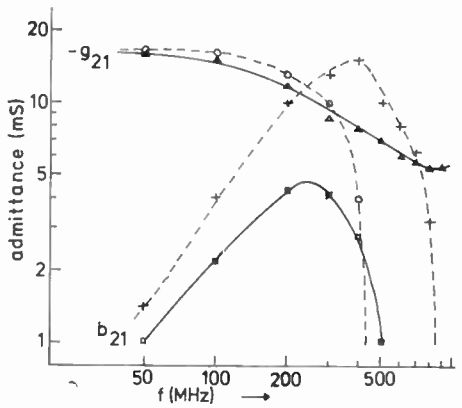


Fig. 4. Transfer admittance for 70 Ω diffused resistor.
Solid curve: resistor mounted in microstrip jig.
Dashed curve: same resistor bonded into TO-5 can.

resonance shown by the in-can measurements near 500 MHz is demonstrated by the jig measurements. The values obtained using the jig lie within 10% of the calculated intrinsic values throughout the frequency range, and within 5% below 300 MHz.

The jig measurements of transfer admittance (Fig. 4) show that the susceptance crosses zero near 500 MHz. This effect is due to common-mode impedance arising from the substrate resistivity. The computed results are very sensitive to small changes in the value of resistance included, computed values with and without a 3 Ω common-mode resistance being shown in Fig. 5. The addition of 5 nH inductors at the input and output ports of the equivalent circuit model to simulate the bonding leads yields good agreement with the in-can results. Although it is relatively easy to obtain the properties of the encapsulated device by adding a value of inductance to give a best fit, the opposite process is not possible in

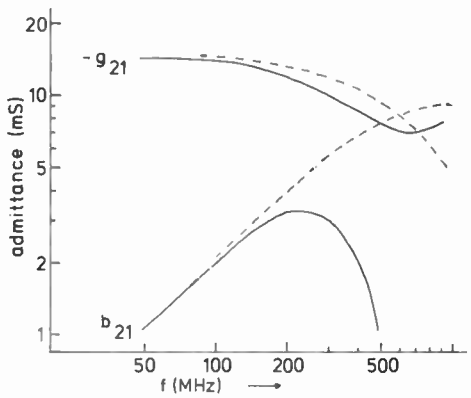


Fig. 5. Computed values of transfer admittance for 70 Ω diffused resistor.
Solid curve: including 3 Ω common-mode resistance.
Dashed curve: without common-mode resistance.

the general case when inductance is present in all leads since it cannot be determined whether a measured effect is due to intrinsic or parasitic causes. For example, the effect of the common-mode resistance in the resistors considered above is similar to that of lead inductance. Thus the two are confused in the in-can measurement and cannot be determined separately.

6. Applications

The jig may be used for measurements on both passive and active components. By using tungsten probes, the need for large bonding pads with their associated capacitance is avoided. Hence measurements can be made on devices designed as elements of integrated circuits, where pads are not required between components. The results obtained on the individual elements may be used to predict the performance of the complete circuit. The method is also useful for device modelling as the results are close to those for the intrinsic device, the normal package parasitics being eliminated. As an example of this use, measurements on m.o.s. transistors have yielded good agreement with transmission line models of the m.o.s.t.[6,7]

7. Acknowledgment

The authors gratefully acknowledge the support of the Ministry of Defence (Procurement Executive) which made possible the continuance of this work started with support from the Science Research Council.

8. References

1. Yanai, H., *et al.*, '*h*-parameters of high frequency transistors in microwave region and their measurement', *Electronics and Communications in Japan*, 51-C, No. 4, pp. 104–111, April 1968.
2. Schneider, M. V., 'Microstrip lines for microwave integrated circuits', *Bell Syst. Tech. J.*, 48, No. 5, pp. 1421–44, May–June 1969.
3. Kemhadjian, H. A., Negandhi, A. and Lewis, B. J., 'Measurements on microcircuits in the range 100–1000 MHz', *The Radio and Electronic Engineer*, 35, No. 4, pp. 217–26, April 1968.
4. Schneider, M. V., 'Dielectric loss in integrated microwave circuits', *Bell Syst. Tech. J.*, 48, No. 7, pp. 2325–32, September 1969.
5. General Radio Manual for 1607-A Transfer Function and Immittance Bridge.
6. Pearce, T., 'High frequency properties of field effect transistors for wide-band integrated circuit applications', M.Sc. dissertation, University of Southampton, December 1968.
7. Haslett, J. W. and Trofimenkoff, F. N., 'Small signal high frequency equivalent circuit for the metal-oxide-semiconductor field effect transistor', *Proc. Instn Elect. Engrs*, 116, No. 5, pp. 699–702, May 1969.

Manuscript first received by the Institution on 12th May 1971 and in revised form on 2nd November 1971. (Paper No. 1434/CC119.)

U.D.C. 621.372.852.1

Annular Resonant Structures and their uses as Microwave Filters

R. T. IRISH,
B.Sc.(Eng.), C.Eng., M.I.E.R.E.*

SUMMARY

Recent trends in the development of microwave integrated circuits have tended to evolve systems with microstrip or triplate interconnexions between the active sections. It is thus desirable that any components incorporated into such microwave systems should be fully compatible with this general circuit form.

A new configuration is described for the realization of a microwave band-pass filter. This configuration is particularly suitable for construction in 'printed-circuit' form and is thus compatible with microstrip and triplate systems. Major advantages accrue from the attenuation characteristic of the filter which does not possess recurrent pass-bands at harmonics (or other overtones) of the design pass-band—as do simple transmission-line structures. The filter is thus eminently suitable for use in frequency multipliers, parametric amplifiers and other systems in which the harmonic performance of the circuits is of paramount importance.

* *Royal Military College of Science, Electronics Branch, Shrivenham, Swindon, Wilts.*

1. Introduction

Modern microwave, band-pass filter systems may be realized in many ways, two of the most common being coupled cavities and one half-wavelength line resonators. In many cases it is desirable that these filters should be able to be 'printed' on the surfaces of a conductor-clad substrate, thus rendering them compatible with other components in microstrip form.

Although the performance of these filters may be tailored to match most responses, the disadvantage of pass-bands which occur at harmonics of the central design frequency often render them unacceptable (in parametric amplifiers and harmonic generators for example). The use of disk resonators[1] does not, in principle, ease the problem as higher-order resonances still occur, although at anharmonic intervals.

Solutions to this problem have been obtained using suitable modes of an annulus. This geometry contains *two* variables which control the modes—the inner and the outer radii. By suitable choice of these radii, resonators may be constructed which have fundamental resonances which correspond, but whose higher-order resonances differ. (This is not the case for the disk resonators referred to above, where the higher order resonances correspond once the fundamental resonant frequency has been selected.) Recurrent pass-bands at frequencies above the central design frequency may thus be eliminated. Measurements of this annular resonator indicate that its Q_0 is high—higher than the equivalent disk resonator.

A range of filters has been constructed using these annuli with very satisfactory results.

2. Theory

2.1. *The Resonant Modes of an Annulus*

The annulus may be analysed by considering it to be a section of radial waveguide, supporting only the dominant (z-independent) modes. Usually the annulus–ground plane spacing is small and the z-dependent modes are cut off. They have not been considered further.

The wave equation for the electric Hertzian vector, Π_e, taken in the z-direction, is

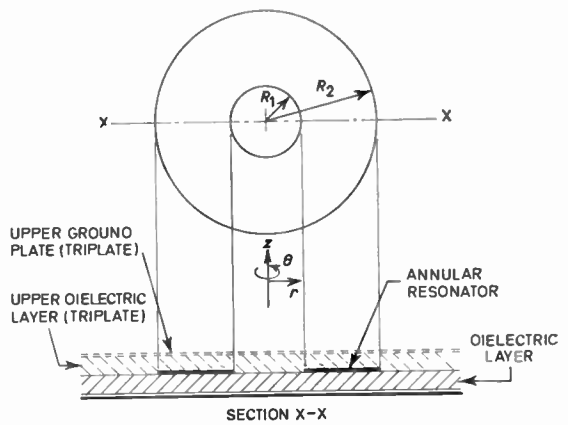


Fig. 1. Basic resonator configuration in stripline form (triplate form shown dotted).

$$\nabla^2\Pi_e - \mu\varepsilon\frac{\partial^2\Pi_e}{\partial t^2} = 0$$

and in the circular cylindrical coordinate system has solutions of the form:

$$\Pi_e = e^{jn\theta}\{A_1 J_n(kr) + B_1 Y_n(kr)\}\, e^{-j\omega t}$$

At $r = R_1$ and $r = R_2$ the radial currents, and thus the angular magnetic fields, must be zero (Fig. 1).

Thus

$$H_\theta = -\varepsilon\frac{\partial}{\partial t}\left(\frac{\partial\Pi_e}{\partial r}\right)\Bigg|_{r=R_1,R_2} = 0.$$

or

$$A_1 J_n'(kR_1) + B_1 Y_n'(kR_1) = 0$$

and

$$A_1 J_n'(kR_2) + B_1 Y_n'(kR_2) = 0$$

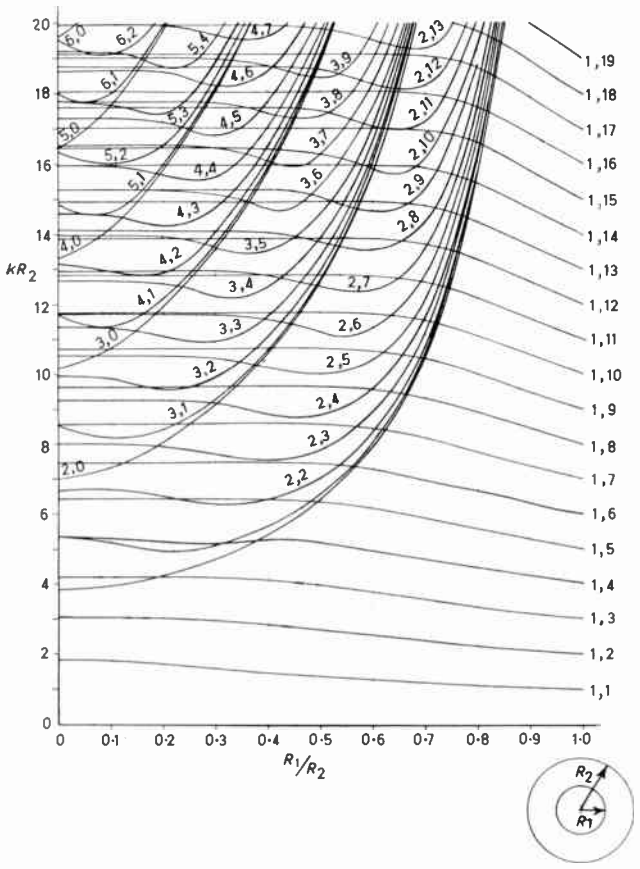
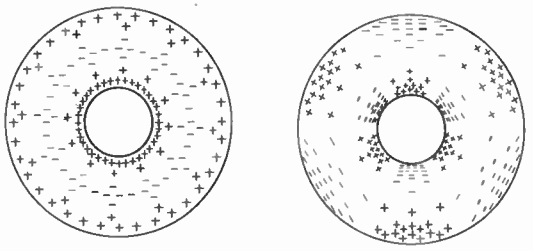


Fig. 2. Mode chart for annular resonators.

Solving these simultaneously gives the characteristic equation of the system:

$$J_n'(kR_1)Y_n'(kR_2) - J_n'(kR_2)Y_n'(kR_1) = 0$$

The eigenvalues of this transcendental characteristic equation for given values of R_1, R_2 and n represent the resonances of the annulus. Calculated results are summarized in the mode chart (Fig. 2). In this chart, the designation (m, n) is used to represent the mth resonance of the nth order TEM mode. Two such modes are sketched as Fig. 3 to clarify this nomenclature.



(a) (b)

Fig. 3. Illustration of some modes of the annular resonator.
(a) The TE_{20} mode (second $n = 0$ mode).
(b) The TE_{23} mode (second $n = 3$ mode).

2.2. *Modal Expansion*

Details of the various modes existing on an annulus may be achieved from the Hertzian vector mentioned above[2] as:

$$E_r = E_\theta = H_z = 0$$

$$E_z = \frac{\partial^2\Pi_e}{\partial z^2} - \mu\varepsilon\frac{\partial^2\Pi_e}{\partial t^2}$$

$$= \omega^2\mu\varepsilon\, e^{jn\theta}\{A_1 J_n(kr) + B_1 Y_n(kr)\}\, e^{-j\omega t}$$

$$H_r = \varepsilon\frac{\partial}{\partial t}\left(\frac{1}{r}\frac{\partial\Pi_e}{\partial\theta}\right)$$

$$= j\frac{\omega\varepsilon n}{r}\, e^{jn\theta}\{A_1 J_n(kr) + B_1 Y_n(kr)\}\, e^{-j\omega t}$$

$$H_\theta = -\varepsilon\frac{\partial}{\partial t}\left(\frac{\partial\Pi_e}{r}\right)$$

$$= j\omega\varepsilon k\, e^{jn\theta}\{A_1 J_n'(kr) + B_1 Y_n'(kr)\}\, e^{-j\omega t}$$

2.3. *The n=0 mode*

The construction of microwave filters consisting of concentric annuli has been considered using the $n = 0$ mode. In this case, angular field variations, which may be ill-defined relatively for two or more such coupled resonators, are eliminated.

With $n = 0$ in the above modal equation and the following boundary conditions (boundaries considered abrupt) apply (Fig. 4):

At

$$r = R_1, \quad E_Z = E_S \quad \text{and} \quad H_\theta = H_S$$

and at

$$r = R_2, \quad E_Z = E_R \quad \text{and} \quad H_\theta = H_R$$

the constants A_1 and B_1 may be evaluated to yield:

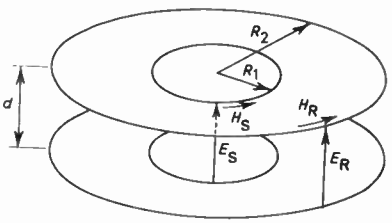


Fig. 4. Boundary conditions for the $n = 0$ mode.

$$\begin{bmatrix} E_S \\ H_S \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} E_R \\ H_R \end{bmatrix}$$

where

$$\alpha = \frac{\pi k R_2}{2} \{J_1(kR_2)Y_0(kR_1) - J_0(kR_1)Y_1(kR_2)\}$$

$$\beta = j\sqrt{\frac{\mu}{\varepsilon}} \cdot \frac{\pi k R_2}{2} \{J_0(kR_1)Y_0(kR_2) - J_0(kR_2)Y_0(kR_1)\}$$

$$\gamma = j\sqrt{\frac{\varepsilon}{\mu}} \cdot \frac{\pi k R_2}{2} \{J_1(kR_1)Y_1(kR_2) - J_1(kR_2)Y_1(kR_1)\}$$

and

$$\delta = \frac{\pi k R_2}{2} \{J_1(kR_1)Y_0(kR_2) - J_0(kR_2)Y_1(kR_1)\}$$

Thus the sending and receiving voltages and currents may be evaluated from $E = -V/d$ and $I = \oint \bar{H} \cdot \bar{dl}$. (These results are consistent with Marcuvitz's expressions for radial waveguides.[3])

2.4. *An Equivalent Circuit for the n=0 mode*

As indicated in Sect. 2.3, the $n = 0$ mode is the most attractive for the construction of microwave filters and the derivation of an equivalent circuit, for use in the neighbourhood of this resonance is thus desirable. It would, in fact, be possible to derive such an equivalent circuit in a straightforward T or Π form, but here, an ideal transformer has been added to this to suggest the natural reduction in potential as the radius is increased (Fig. 5).

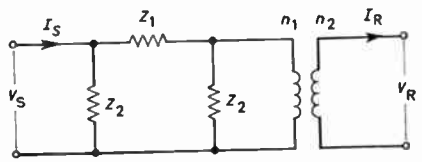


Fig. 5. An equivalent circuit for the $n = 0$ mode.

Analysis of this equivalent circuit in terms of its A, B, C, D parameters gives:

$$Z_1 = \frac{AB}{\sqrt{BC+1}}; \qquad Z_2 = \frac{AB}{BC+1-\sqrt{BC+1}}$$

and

$$\frac{n_1}{n_2} = \frac{A}{\sqrt{BC+1}}$$

Under the resonance condition, $C = 0$.

Thus $Z_1 = AB$, $Z_2 = \infty$ and $n_1/n_2 = A$. If the impedance Z_2 may be represented by a parallel-tuned circuit, capacitance C' and inductance L:

$$C' = \frac{-j\,dC}{4A\,d\omega}\bigg|_{\omega_0}$$

$$= \frac{\pi\varepsilon R_1}{2d}\left\{\frac{J_1(kR_1)Y_0(kR_2) - J_0(kR_2)Y_1(kR_1)}{J_1(kR_2)Y_0(kR_1) - J_0(kR_1)Y_1(kR_2)} \cdot R_2 - R_1\right\}$$

$$L = \frac{1}{\omega_0^2 C'}$$

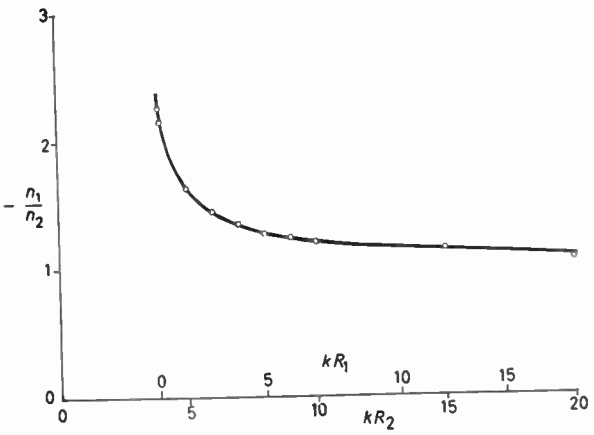


Fig. 6. Turns ratio (n_1/n_2) v. $kR_{1,2}$ for the first zeroth-order mode resonance.

Figure 6 shows how n_1/n_2 varies with the annulus dimensions R_1 and R_2 at resonance.

The above calculations relate to the resonator in microstrip form. If the design is modified to triplate, identical calculations will apply, but C' will be doubled in value (and L halved). Sample calculations of the parameter Z_1 $(= AB)$ show it to be negligible compared with the other circuit impedance under normal conditions, thus simplifying the equivalent circuit considerably.

2.5. *The Resonant Annulus as a One-port Device*

Thus far, the resonator has been considered as a transmission type of device—from one radius to another. It is, however, possible to use the resonator as a simple one-port device and the analysis for this case has been included for completeness.

Considering the port to be at $r = R_2$ and the device to be open-circuited at $r = R_1$:

$$y_{in} = \frac{H_\theta}{E_z} \cdot \frac{2R_2}{d} \qquad \text{for } n = 0$$

for microstrip configurations (double these values apply to triplate line) and

$$y_{in} = \frac{H_\theta}{E_z} \cdot \frac{R_2}{nd} \qquad \text{for } n \geqslant 1$$

Substitutions from the equations for H_θ and E_z give the wave admittance:

$$y_{ow} = j\sqrt{\frac{\varepsilon}{\mu}} \cdot \frac{J_n'(kR_2)Y_n'(kR_1) - J_n'(kR_1)Y_n'(kR_2)}{J_n(kR_2)Y_n'(kR_1) - J_n'(kR_1)Y_n(kR_2)}$$

The evaluation of the equivalent circuit in terms of a parallel inductance/capacitance combination may be found by comparison of the reactive slopes at resonance, i.e.

$$C = \frac{1}{2j} \cdot \frac{dy_{in}}{d\omega}\bigg|_{\omega=\omega_0}$$

Differentiation and the subsequent double reduction[4] of the Bessel–Neumann functions in the calculation of C are straightforward but lengthy in all cases except where $n = 0$. Under this limited condition C (stripline)

is given by

$$C=\frac{\pi\varepsilon R_2}{d}\left[\frac{\begin{array}{l}Y_1(kR_2)\{J_0(kR_1)+\frac{1}{kR_1}J_1(kR_1)\}\\ \quad -J_1(kR_2)\{Y_0(kR_1)\\ \quad\quad +\frac{1}{kR_1}Y_1(kR_1)\}\end{array}}{J_1(kR_1)Y_0(kR_2)-J_0(kR_2)Y_1(kR_1)}.R_1+R_2\right]$$

2.6. *The Effects of Fringing Capacitance*

Associated with the radial discontinuities of this resonator system are the fringing capacitances indicated below (Fig. 7). The exact solution for these capacitances are not currently known although Sneddon[5] has solved, with some complexity, the case of an electrostatically-charged, infinitely-thin disk between two ground-planes. The Schwarz–Christoffel transformation, applied to a charged centrally-mounted, semi-infinite plane between two infinite-conducting planes is, however, well documented[6,7] and this has been used here to obtain a coarse indication of the fringing capacitance relating to the edges of the annular resonator. The fringing capacitance is given by:

$$C_f=\frac{\varepsilon}{\pi}\left\{\left(\frac{2}{1-t/b}\right)\log_e\left[\frac{1}{1-t/b}+1\right]-\left[\frac{1}{1-t/b}-1\right]\log_e\left[\frac{1}{(1-t/b)^2}-1\right]\right\}F/m.$$

where t is the thickness of the centre plane and b is the separation of the outer planes.

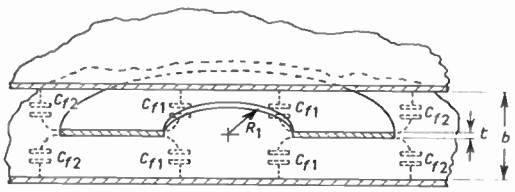


Fig. 7. Fringing capacitances associated with the annular resonator.

The precise value of the effective fringing will however differ slightly from the value obtained from the above equation, which relates to 'static' conditions. No significant discrepancy which could be attributed to this expectedly-small error was observed in any subsequent filter design.

3. Circuits Utilizing the Annular Resonator

Two trial devices were built to confirm the above theoretical results and the expected higher-order mode rejection properties of this type of system:

3.1. *The Concentric Ring Filter*

A Chebychev third-order filter with 1 dB ripple in the pass-band was designed utilizing the equivalent circuit indicated in Fig. 5 (Sect. 2.4). This filter was nominally centred at 10 GHz and was to have a bandwidth of 100 MHz.

The filter, as indicated in Fig. 8, was constructed of three concentric rings, each resonating in the (1, 0) mode. The fringing capacitance coupling, from one resonator to the next, proved to be much in excess of the desired

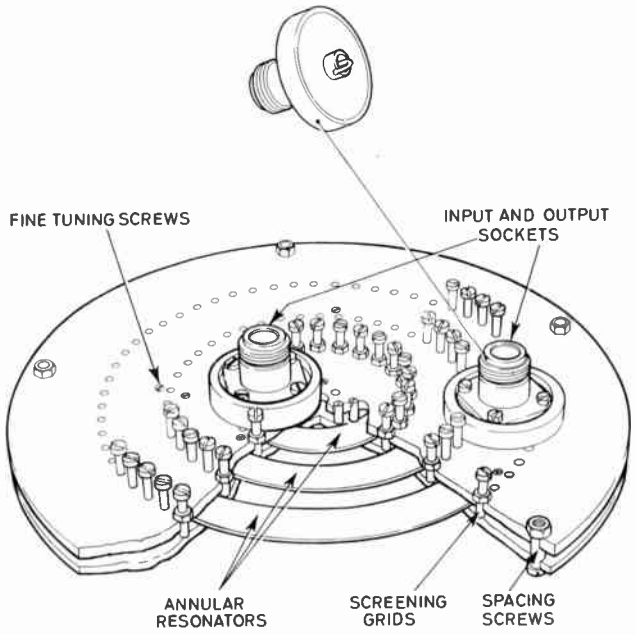


Fig. 8. The concentric-ring filter.

value and a system of screening-posts were used to reduce this to the required value.[8] Input and output connexions were made using loop coupling at the centre of each resonator. These loops were able to be adjusted by rotating them about their axes. An odd number of narrow radial slots (not shown in the figure) were cut in the disks to suppress any tendency to resonate in other modes.

The transmission characteristic of the filter is recorded in Fig. 9 and shows close agreement with theory.

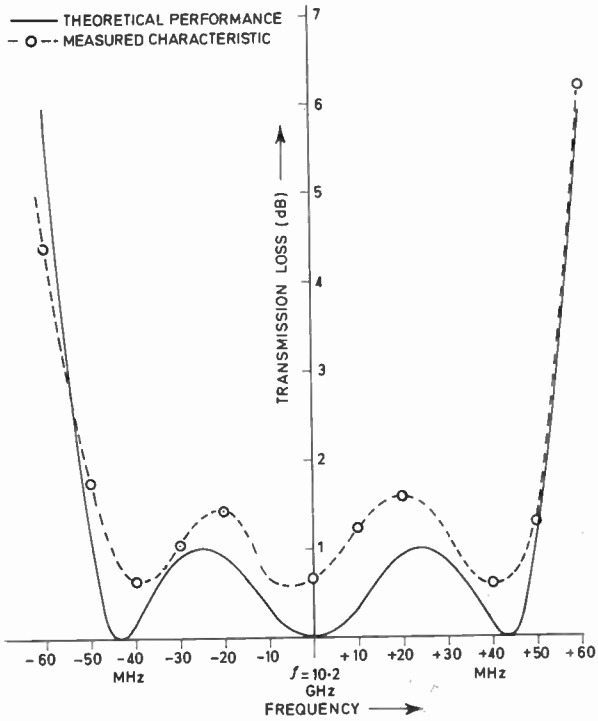


Fig. 9. The concentric-ring filter. Chebychev, third-order response ($\varepsilon_{rip}=1$ dB).

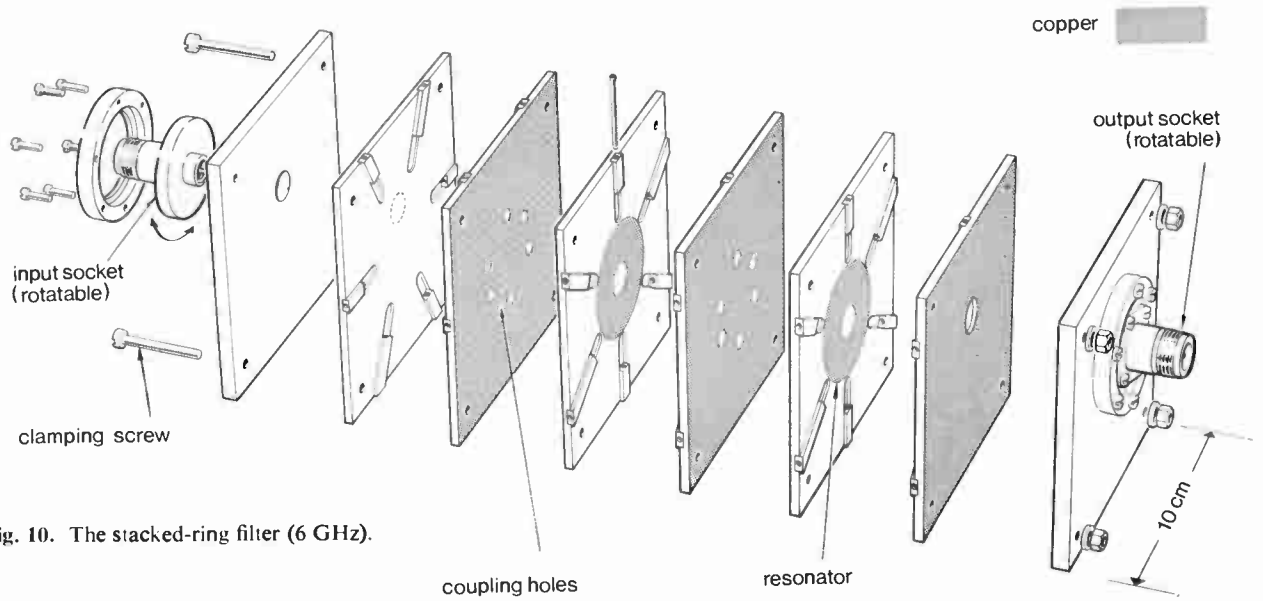


Fig. 10. The stacked-ring filter (6 GHz).

The unloaded Q_0 of each resonator has been calculated, from the electromagnetic field equations given in Section 2, and is in the order of 5000, confirming the low measured attenuation characteristic of this filter within its pass-band. In the case of practical microstrip systems, this value of Q_0 may be slightly degraded from this figure by dielectric losses and by radiation from the resonator.

No spurious filter transmission was observed over the frequency range 5 to 12 GHz. This covers a considerable number of other modes of the individual resonators and thus confirms the rejection properties of the device.

3.2. *The Stacked-ring Filter*

Whilst the above type of filter was electrically satisfactory, its construction was far from simple and would not be commercially viable. A different construction, however, enabled the device to be realized, using 'polyguide' material. This is a copper-clad, high-quality dielectric sheet which may be accurately photo-etched to produce the ring resonators. These resonators are stacked, one on top of the other and coupling is achieved by perforations in the copper screening layers (Fig. 10).

The results from this filter again agree well with theory except that the pass-band ripple is slightly greater than the design figure of 1 dB (Fig. 11). This is directly attributed to the size of the coupling holes which were slightly enlarged over their theoretical dimensions to take into account the (overestimated) field losses incurred by the finite thickness of the ground plane. The higher-order mode rejection performance was good, no spurious transmissions at a level > 50 dB (limit of measuring equipment) being detectable over the available frequency range 5 to 12 GHz.

4. Comments and Conclusions

A new resonator design, compatible with microcircuit techniques has been described which has several significant advantages over the conventional, half-wave resonator systems. Filter units have been constructed using this technique and its viability confirmed both in the pass and stop bands. The resonator configuration may also be used for frequency multiplication and parametric amplifier systems by utilizing two or more modes whose resonant frequencies are interrelated in the desired manner.

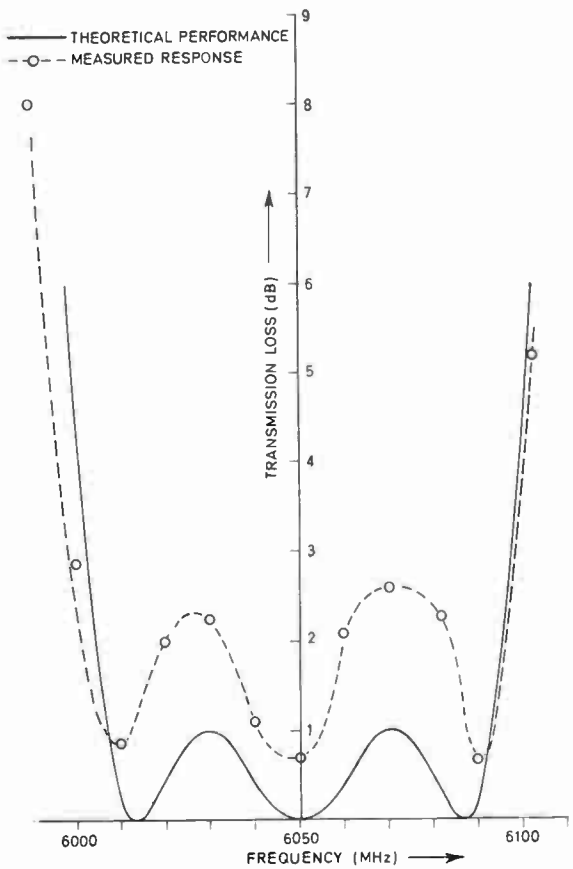


Fig. 11. The stacked ring filter.
Chebychev, third-order response ($\varepsilon_{rip} = 1$ dB).

This form of resonator is particularly well-suited to fabrication in microcircuit form, either in conventional stripline or in triplate form, particularly if high permittivity substrates, such as sapphire or rutile are used. An X-band resonator on a sapphire substrate would be approximately 0·42 cm in diameter (1, 1 mode) allowing adequate residual area on a conventional substrate disk (about 1 inch diameter) for the deposition of any associated active or passive components.

The manufacture of filters using this form of resonator would be commercially more attractive if the annuli were deposited in a single plane, either concentrically (without the need for the somewhat complex screening described in the trial device) or in line, where the coupling between resonators is provided by their respective proximities.[9] Such arrangements require a very significant extension of the work presented herein and it is hoped to make this the subject of a future communication.

5. Acknowledgment

The continuing encouragement and many fruitful discussions with Professor M. H. N. Potok are gratefully acknowledged.

6. References

1. Watkins, J., 'Circular resonant structures in microstrip', *Electronics Letters*, 5, pp. 524–5, 1969.
2. Stratton, J. A., 'Electromagnetic Theory', p. 349 *et seq.* (McGraw-Hill, New York, 1941).
3. Marcuvitz, N., 'Waveguide Handbook', p. 89. (McGraw-Hill, New York, 1950.)
4. McLachlan, N. W., 'Bessel Functions for Engineers'. (Oxford University Press, 1934.)
5. Sneddon, I. N., 'Mixed Boundary Value Problems in Potential Theory'. (North Holland, Amsterdam, 1966.)
6. Matthaei, G. L., Young, L. and Jones, E. M. T., 'Microwave Filters Impedance-Matching Networks and Coupling Structures', p. 163. (McGraw-Hill, New York, 1964.)
7. Moon, P. and Spencer, D. E., 'Field Theory for Engineers', pp. 339–41. (van Nostrand, New York, 1960.)
8. Mumford, W. W., 'Some technical aspects of microwave radiation hazards', *Proc. Inst. Radio Engrs*, 49, pp. 427–47, 1961.
9. Clar, P., 'The Application of Dielectric Resonators to Microwave Integrated Circuits'. Digest of Technical Papers; International Microwave Symposium, G-MTT 1970, pp. 19–23.

Manuscript received by the Institution on 11th August 1971. (*Paper No. 1435/CC120.*)

U.D.C. 621.337.49 : 621.86 : 629.11

Electronic Control of Battery Electric Vehicles

JOHN J. MORRISON,
M.I.M.C., C.Eng., F.I.E.E.*

Based on a paper presented at the Conference on Electronic Control of Mechanical Handling held in Nottingham from 6th to 8th July 1971.

SUMMARY

The paper defines the requirements of a traction control system and considers the characteristics of shunt wound and series wound motors for traction duties. Step control systems are briefly reviewed and their limitations discussed and methods of applying electrical braking to d.c. motors are outlined with particular reference to industrial trucks. Basic pulse control systems are reviewed and a recently introduced electronic vehicle controller, which incorporates several new techniques, is described and the novel features are examined in detail.

* *Cableform Limited, Romiley, Near Stockport, Cheshire SK6 3JQ.*

1. Introduction

An important element of the total mechanical handling field is the industrial truck which has been highly developed in the post-war decades and now includes a wide array of special-purpose vehicles.

These may be categorized as:

Pedestrian Controlled Trucks (Pallet, Platform or Fork Lift).

Rider Controlled Trucks (Pallet, Platform, or Fork Lift which may be counterbalanced, straddle or reach, or side loader types, Towing Tractors, or Order Pickers).

Most of these vehicles are battery electric-powered for reasons of efficiency, economy and reliability, and for internal works transport the absence of noise and air pollution is a further important factor.

Since the energy source is a storage battery the motive power is produced by a d.c. motor and the control requirements are:

(a) Limit the starting current to provide safe controlled acceleration.

(b) Provide speed control over an adequate range.

(c) Provide selection of running direction.

(d) Provide electrical braking.

(e) Protect against fault conditions.

In any d.c. motor the torque $T \propto \Phi I$ and the speed $N \propto E/\Phi$, where I is the armature current, Φ is the magnetic flux and E is the back e.m.f.

D.c. motors are classified according to the connexion of the field coils relative to the armature circuit.

Shunt-wound motors have the main field winding connected in parallel with the armature, so that when connected to a constant voltage supply the machine is virtually separately excited and the flux will be approximately constant at all loads.

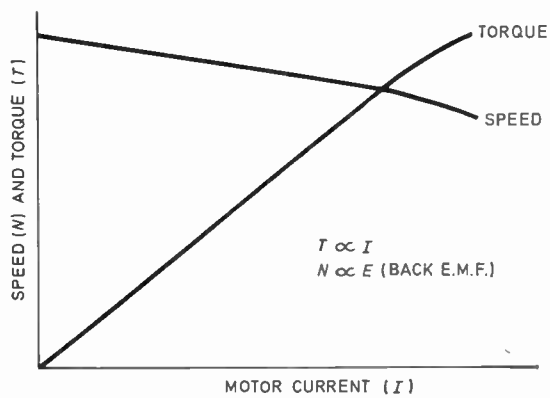


Fig. 1. Shunt motor characteristic.

Then

$$T \propto I$$

and large values of torque require high armature currents, and

$$N \propto E$$

which is approximately constant since

$E = V$—armature volt drop,

which is normally less than 5% of V at full load.

The shunt motor characteristic is given in Fig. 1.

Speed control of a shunt-wound motor is usually achieved by field weakening and the maximum speed range to maintain stability is approximately 4 : 1.

Although a few battery vehicles have been designed with shunt motors and systems of electronic field control are now being examined, in general shunt motors are best suited to applications requiring constant speed independent of the load and where starting torque requirements are not severe.

Series-wound motors have the main field connected in series with the armature and any changes in armature load current cause a corresponding change in magnetic flux.

While the magnetic circuit is unsaturated,

$$T \propto I^2$$

and

$$N \propto 1/I.$$

As the iron begins to saturate the rate of increase in flux with current decreases and the torque and speed curves flatten off.

The series motor characteristic is shown in Fig. 2.

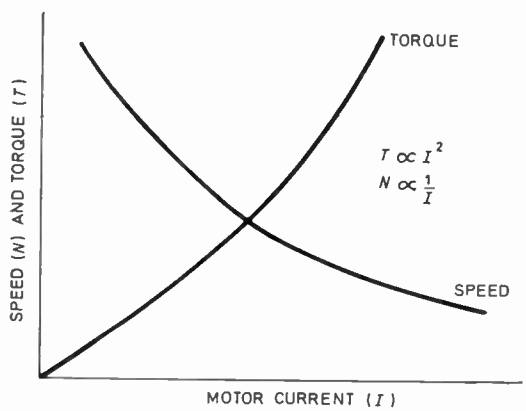


Fig. 2. Series motor characteristic.

The rapid increase in torque as the speed decreases makes this type of motor particularly suitable where large masses have to be accelerated and the series-wound motor is almost universally used for traction duties.

2. Resistance and Step Control Systems

If a standing motor were switched directly on to full voltage the current and torque would be excessive and the voltage applied to the motor must therefore be brought to a low value and gradually increased as the motor speeds up. A method widely employed is to insert a resistance in series with the motor and this is cut out in stages by a series of contactors until the motor accelerates naturally on the motor curve at full voltage. This is shown in Fig. 3 which is known as the notching curve.

The main function of this type of controller is to limit starting currents and it has only a limited value as a speed controller because of the heat dissipation in the series

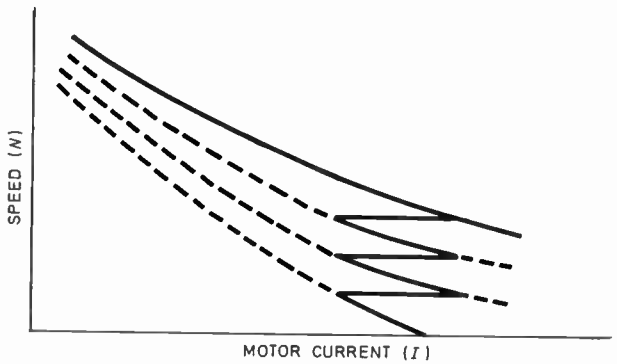


Fig. 3. Notching curve.

resistors. All stepped controllers produce acceleration jerks at each step and considerable power is lost in series resistance.

A carbon pile rheostat can be used to eliminate the acceleration steps but there is still considerable power wasted in heat dissipation which limits its use as a speed controller to relatively short time ratings, and the only economical running speed is the full voltage curve.

To reduce starting losses and provide additional economic running notches series–parallel switching of battery sections and/or motors (in multiple motor vehicles) is frequently adopted and Fig. 4 shows a typical notching diagram for a battery switching system. A comparison of starting losses is shown in Fig. 5 for full battery starting and parallel–series battery starting. The diagram shows that the starting losses are greatly reduced and an additional 'loss free' low speed notch is obtained.

This method of improving efficiency provided one of the earliest uses of electronic semiconductors on battery traction by using silicon diodes as blocking devices on the switching network. A typical control circuit using battery switching with diode routing is shown in Fig. 6.

Electric vehicles can be conveniently provided with electrical braking to improve manœuvrability and greatly extend the life of the mechanical braking system. Electrical braking of a d.c. motor is achieved by driving the motor as a generator,

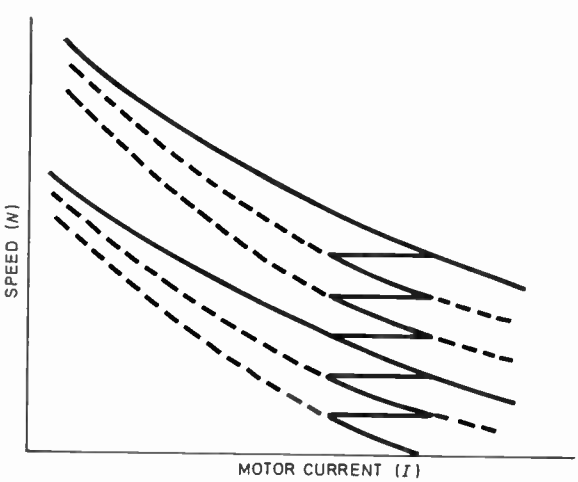
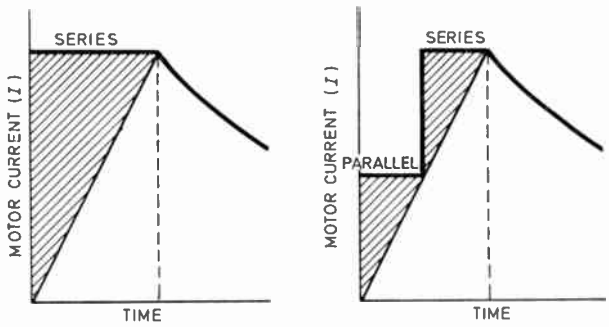


Fig. 4. Series–parallel notching curve.



(A) Full battery starting. (B) Series–parallel starting.

Fig. 5. Starting losses.

thus converting the kinetic energy of the moving mass to electrical energy and the product of current, voltage and time (IEt) represents the braking energy absorbed. When this conversion is performed at low voltage, high current and dissipated in a resistance it is known as rheostatic braking and when the conversion is performed at high voltage, low current and returned to the supply it is known as regenerative braking. The fields of application of these two methods are quite distinct.

In bringing a moving vehicle to a stop the energy to be absorbed is relatively small and rheostatic braking is appropriate. On the other hand, holding a vehicle, for instance a locomotive, to a specified speed down a long steep gradient represents a sustained output of a much higher order and regenerative braking is most suitable.

Regenerative braking can be applied fairly easily to a shunt-wound motor but on series-wound motors the circuit requirements are more complex and for industrial trucks where gradients and braking times are generally short, it is not usually an operational advantage or economically attractive.

There are then two appropriate methods of applying electrical braking to a series-wound motor on an industrial truck.

2.1. *Rheostatic Braking*

This involves disconnecting the motor from the supply, reversing the field connexion relative to the armature and connecting a resistance across the motor. The motor then acts as a series generator supplying the resistance load: braking torque $\propto N\Phi^2$, i.e. the braking torque decreases as the motor slows down. This method of braking is used extensively on battery locomotives and sometimes on industrial trucks, and frequently the starting resistors are also used as braking resistors.

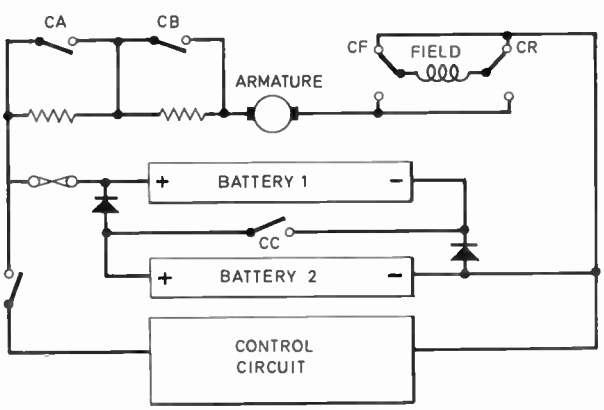


Fig. 6. Parallel–series switching scheme.

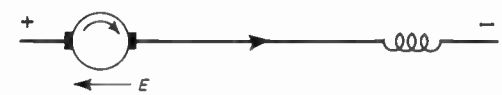
2.2. *Plug Braking (Direction Reversal)*

In this form of braking the motor connexions are reversed so that the motor tends to run in the opposite direction to the vehicle motion. The supply voltage and the back e.m.f. are then acting in the same direction and a resistance is included in the circuit to limit the current:

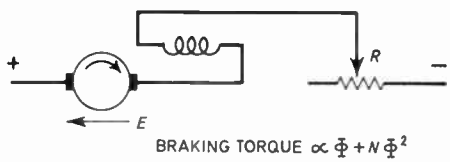
$$\text{braking torque} \propto \Phi + N\Phi^2,$$

i.e. braking torque is available to standstill.

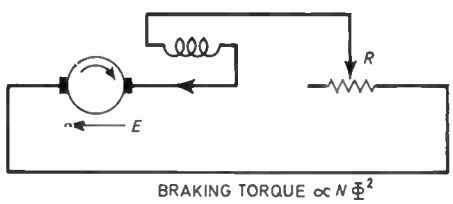
This method gives a greater braking torque but current is drawn from the battery during the braking period and this energy as well as the kinetic energy must be dissipated in the resistors. The braking resistors require to be approximately twice the value of the starting resistors. The circuits of these conditions are shown in Fig. 7.



(a) Motoring.



(b) Rheostatic braking.



(c) Plug braking.

Fig. 7. Running and braking conditions.

The rheostatic braking described above, using the motor as a series generator, produces very high currents and braking torques, and it is possible, by electrically separating the armature and field circuits, to have effectively a separately excited generator and control the braking torque by inserting resistance in the field circuit. The field current will be typically less than 10% of the series generator case and the resistor values will be much higher.

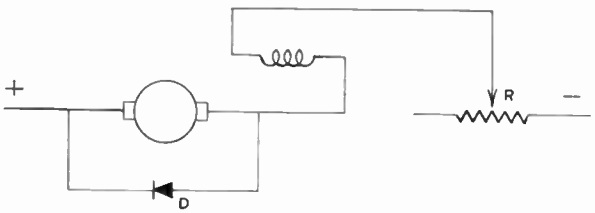


Fig. 8. Dynamic braking conditions.

Figure 8 shows a method of achieving dynamic braking which is a combination of rheostatic and plug braking. The armature current circulates through the diode D and the field current is controlled by the resistance R. At high speeds the armature current is larger than the field current and the diode D is always forward biased. In this case the braking is rheostatic and the motor is operating as a generator.

The braking energy is dissipated in the armature resistance and this is normally satisfactory on industrial trucks where gradients and braking times are short. On vehicles which can brake for a considerable time, such as road vehicles or locomotives, a resistor can be inserted in the armature circuit for the duration of braking and shorted out by a contactor for normal running.

At low speeds the current generated by the armature will fall until it is less than the field current and current will flow from the battery through the armature and field and the machine will motor in reverse to the vehicle motion. This is the plug braking mode and will bring the vehicle to a standstill.

Figure 9 illustrates a typical step controller which consists of a contactor control panel and an electronic speed control unit which provides automatic acceleration sequence, controlled braking and protection circuits.

While these step control systems are widely used and are completely satisfactory on many vehicles, in recent years electronic methods have been developed offering operational advantages which are important on industrial trucks which have very frequent starting cycles and require good low-speed manœuvrability.

The principal advantages are:

(a) Smooth stepless control.
(b) Increased efficiency.
(c) Simple effective braking.
(d) Reduced maintenance and service costs.

3. The Pulse Controller

The basic electronic method of controlling a traction motor is the pulse controller developed since the introduction of high-power semiconductors. The fundamental circuit is shown in Fig. 10.

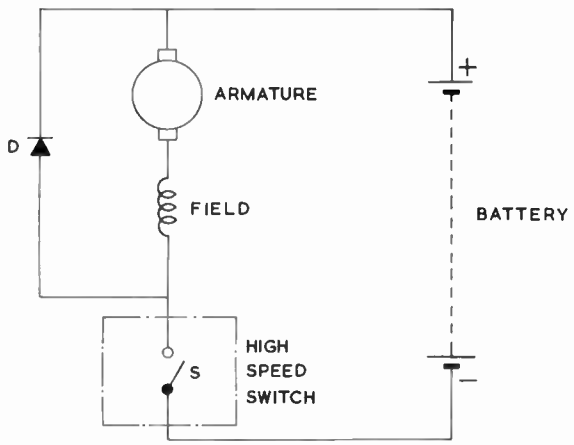


Fig. 10. Fundamental pulse controller.

The vehicle is started by closing the switch S which applies the full battery voltage to the motor. The current

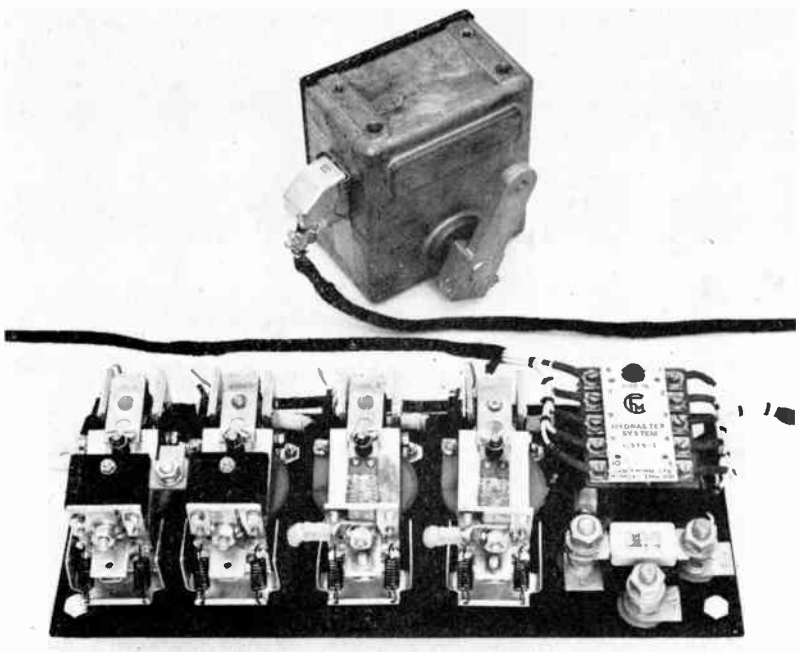


Fig. 9. Hydrastep control system.

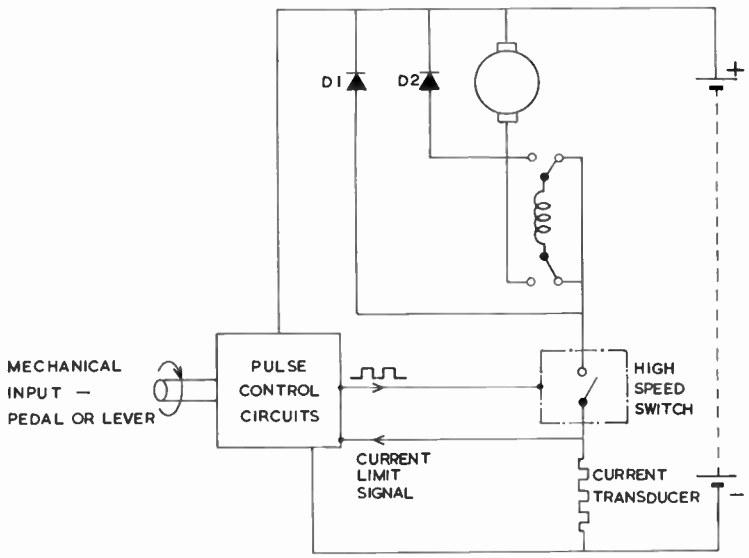


Fig. 11. Complete pulse control system.

rises inductively and after a predetermined time the switch is opened and the current decays through the motor and the free wheel or inter-pulse diode D; no battery current flows during this period. The switch is then re-closed and the cycle is repeated. The average motor current is determined by the mark/space ratio of the switch and the system acts like a d.c. transformer, transforming high battery voltage, low battery current to low motor voltage, high motor current.

The voltage applied to the load is a series of pulses but the diode D allows the motor inductance to smooth the motor current which then appears as a d.c. level with a superimposed ripple.

Since the mark/space ratio can be easily varied an infinite number of motor voltages can be obtained without the heavy power losses of resistance step control. Typical efficiency of a pulse controller is 95% to 98%.

To the basic controller of Fig. 10 a number of functions have to be added to produce a practical vehicle controller:

(a) Contactors are provided on the field winding to allow direction of motion or braking to be selected.

(b) A diode is added across the armature as described earlier to allow the motor to be used for dynamic braking.

(c) A control circuit to vary the mark/space ratio of the switch and consequently the voltage across the motor.

(d) A monitoring circuit so that the maximum current may be limited to ensure that the controller is not overloaded.

A complete pulse control system is shown in Fig. 11.

3.1. *The High-speed Switch*

While a mechanical high-speed switch would fulfil the requirements of a pulse control system it is clearly not a feasible engineering solution at the currents and switching speeds required, and semiconductor switching devices are used for this purpose.

Two types are available:

(1) *The Transistor.* It is possible to drive a suitable transistor as a switch by arranging it to operate between the two extremes of zero base current and high base current in which case the emitter-collector resistance varies between a very high value and a very low value.

If the transition from 'on' to 'off' is achieved rapidly by driving the device with a square pulse waveform, the power losses in the switch are relatively small. While the base pulse is applied the switch is closed and when the pulse is removed the switch immediately opens.

Transistor pulse controllers are usually applied to the smaller range of vehicles with relatively low battery voltages. Typically, maximum motor currents are 200 to 250 A at 24 V.

(2) *The Thyristor.* The thyristor (or silicon controlled-rectifier) is similar to a silicon diode in that it has a reverse direction in which it has a very high resistance

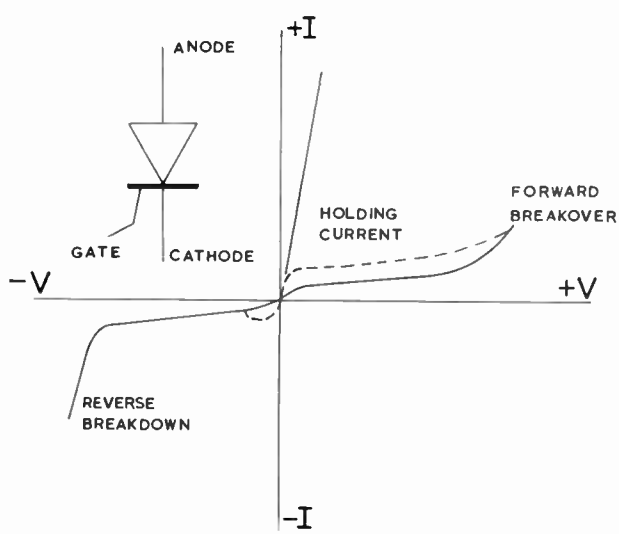


Fig. 12. Thyristor characteristic.

and a forward direction in which it can have a very low resistance. It can be switched from a high forward resistance to a low forward resistance by applying a small control signal to the gate electrode (Fig. 12).

Unlike the transistor, when the gate signal is removed the thyristor switch remains 'on' and can only be turned 'off' by reducing the main anode cathode current to less than a very small current known as the holding current. It is therefore not necessary for the gate signal to be maintained over the whole of the conduction period and it is common practice to control a thyristor by short 'turn-on' gate-firing pulses.

Thyristors are available in a wide range of voltages and currents and they are the most widely used switching element in electronic pulse controllers.

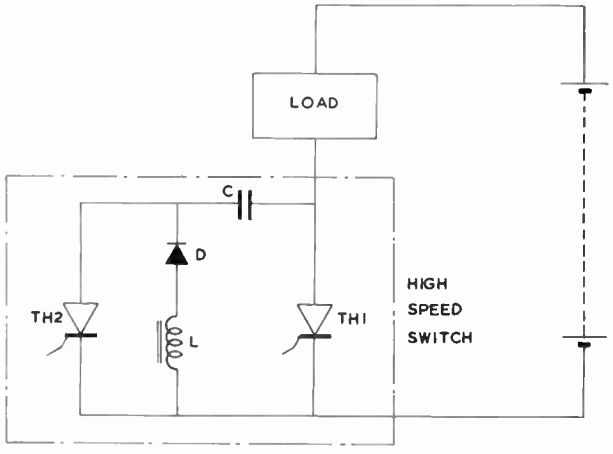


Fig. 13. Thyristor commutation.

3.2. *Commutation*

The process of turning off a thyristor is known as commutation and on d.c. circuits, since mechanical interruption is obviously not feasible, this is accomplished by reversing the voltage across the thyristor and diverting the current.

This is achieved by storing energy in a capacitor during the 'on' period and discharging this capacitor across the thyristor in the reverse polarity to turn it 'off'. A typical commutation circuit is shown in Fig. 13.

Consider initially that the main thyristor TH1 is 'off' and the commutation thyristor TH2 is 'on'. The capacitor C charges to battery voltage via the load impedance and TH2 turns 'off' naturally as soon as the capacitor is charged.

TH1 is then turned on and load current flows from the battery through TH1 and C discharges via TH1, the choke L and the diode D; these components form a resonant circuit to reverse the charge on C which is then held in this condition by the commutation diode D.

When thyristor TH2 is fired on, C discharges in the reverse direction through TH1, thus turning it off and the charging cycle is repeated.

When the load is inductive, as in the case of a traction motor, the free wheel diode forms an effective short circuit across the load and due to the source inductance C charges to a higher value than battery volts. Since TH2 has now turned off, C discharges via the commutation choke and diode to a value less than battery volts and remains at this low level.

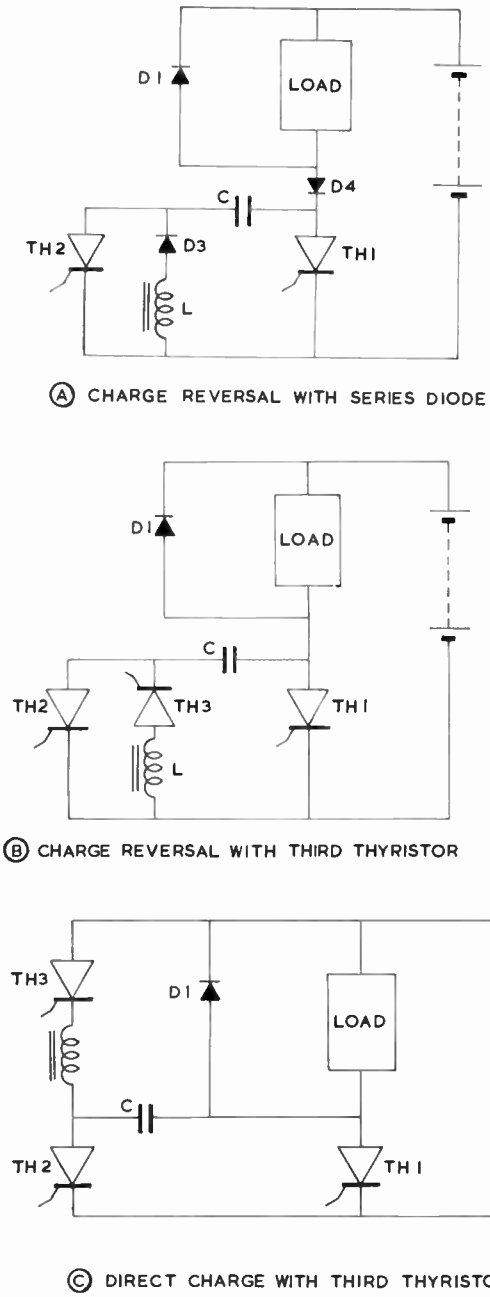


Fig. 14. Commutation methods on inductive loads.

Several methods have been adopted to overcome this effect and these are shown in Fig. 15.

On Fig. 14(a) a diode D4 has been added in series with the motor to prevent the discharge of C through D1, D3 and L.

This method reduces system efficiency due to the power loss in D4 and by substituting a thyristor TH3 in place of D3, which is fired at the same time as TH1, the same effect is achieved with no power loss as in Fig. 14(b).

An alternative method is shown in Fig. 14(c) where the charge thyristor TH3 is fired at the same time as TH1 and C charges directly from the source. This method is hazardous since if due to a fault TH2 and TH3 are on together a short circuit is applied across the battery.

All of these systems produce a voltage on the capacitor which is higher than battery voltage, due to the transfer of energy from the source inductance, and the final voltage on the capacitor:

$$E = I\sqrt{\frac{L_s}{C}} + V_B$$

where L_s is the distributed source inductance and V_B is the battery voltage.

This stored voltage can be 200% or more of battery voltage and this is often claimed as an advantage since theoretically the value of C and therefore the size can be reduced. However, there are also some disadvantages. First, to avoid the risk of voltage breakdown the working voltage rating of the capacitor should be considerably increased and although the value of C may be reduced there is little change in the physical size. Secondly, more complex circuit arrangements are required to operate over the full range into full conduction since the voltage on the capacitor would leak away when the switch ceased pulsing. This problem is frequently avoided by restricting the range of the controller to some value below full conduction and then short circuiting the main thyristor by a contactor to achieve full speed.

4. The Pulsomatic† Controller

This is a recently introduced vehicle controller incorporating several new techniques which will be discussed.

4.1. *Control Characteristics*

Most electronic controllers have a current limit system which operates at some predetermined but fixed level, as shown by the ordinate CF in Fig. 15 which shows a typical series traction motor characteristic.

The Pulsomatic Controller employs a differential amplifier to shape the current limit, and in effect the speed/current characteristic of the motor, to the curve A B D. This technique gives the following important advantages over conventional methods:

(1) The unsaturated series motor characteristic is maintained over the entire useful speed/torque range of the machine. This gives a constant power output at high torques and this has been suggested as the ideal characteristic for electric traction by Mangan and Griffith.[1]

(2) Considerably reduces the maximum average current drawn from the battery and gives a continuously rated controller at all values of torque. The average battery current is given by the curve A B E in Fig. 15.

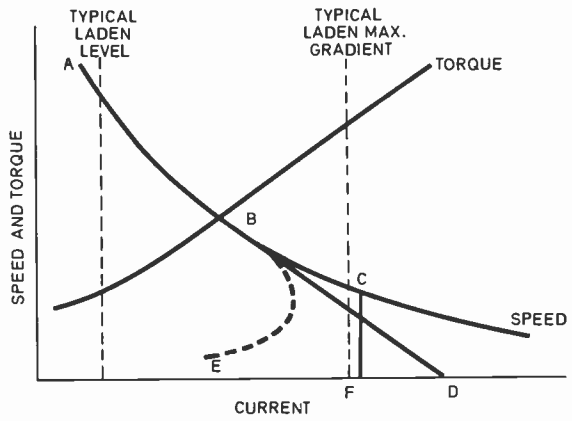


Fig. 15. Current control characteristics.

(3) Thermal protection can be provided over a wide range of temperatures. A temperature sensing element is fitted to the main thyristor heat sink and connected in a feedback circuit so that if the vehicle is being worked beyond its rating, the current limit line is gradually depressed to limit the power delivered by the controller so that the safe maximum ratings are never exceeded.

(4) Gives smooth safe transition into shunt contactor mode where the vehicle design requires it. For instance, a low voltage vehicle with a fairly resistive motor may require full stall current at full voltage to accelerate on a gradient. This is achieved by ensuring that the shunt contactor can only be operated when the input demand signal is maximum (i.e. full pedal depression) and the controller has been in current limit for a predetermined time. The circuit is arranged so that this time interval is proportional to motor current, and therefore torque, so that the time before the shunt contactor closes is longer for higher and shorter for lower values of torque.

The current limit line can be selected to operate over any part of the motor characteristic and, as the slope of the line increases, the maximum average battery current approaches motor current as shown in Fig. 16.

4.2. *Commutation*

As discussed earlier, it is common practice to use short gating pulses to fire the thyristors and consequently the commutation thyristor turns off automatically when

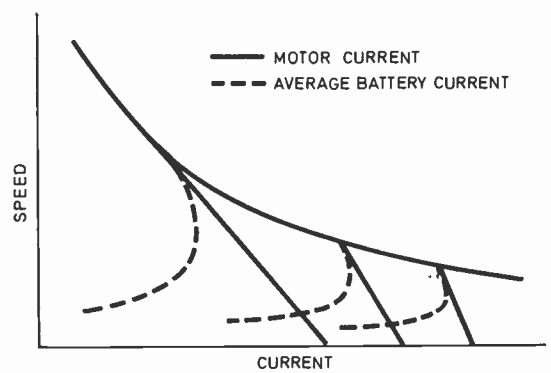


Fig. 16. Effect of current limit slope.

†Registered Trade Mark of Cableform Limited. Protected by British and Foreign Patents; their description here implies no right of licence.

the capacitor has charged. As we have seen, this permits the capacitor to discharge to a value less than battery volts unless a series diode or third thyristor is added as in Fig. 14. The Pulsomatic system does not employ short pulse gating but applies the gate voltage during the whole period that conduction is required. When the voltage across the commutating thyristor becomes positive it conducts again and the voltage on the capacitor remains at battery volts and the energy from the source inductance is transferred to the commutation choke L. A small bleed resistor R ensures that battery volts are maintained on the capacitor when the switch stops pulsing on full conduction. The circuit is shown in Fig. 17.

This technique achieves a simple reliable commutation system which does not attempt to store the enhanced voltage produced by 'ringing' with the source inductance and yields the following advantages:

(a) Eliminates the need for series diodes or third thyristors.

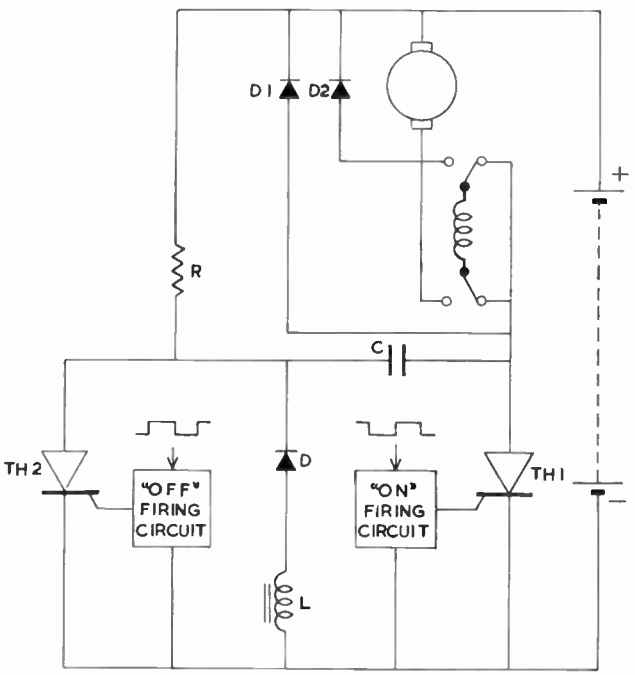


Fig. 17. Pulsomatic switching circuit.

(b) The d.c. gating gives a low impedance system virtually immune from noise interference and avoids sensitive timing requirements.

(c) Permits full conduction operation without complex circuitry.

(d) Allows safe operation of thyristors in parallel for high current requirements.[2]

4.3. *Pulse Modulation*

There are two basic methods of applying the variable mark/space pulse waveform to drive the high-speed switch.

(1) Variable frequency modulation in which the 'on' pulse width is fixed and the repetition frequency is variable.

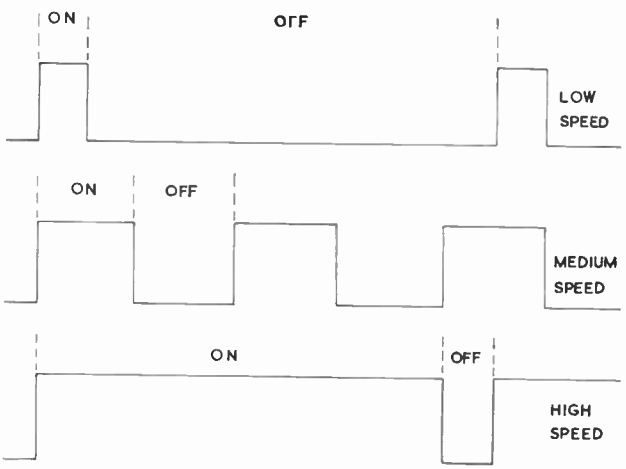


Fig. 18. Pulsomatic modulation waveforms.

(2) Fixed frequency modulation in which the frequency is fixed and the 'on' pulse width is varied.

The Pulsomatic system employs a hybrid system in which both frequency and pulse width are varied over the range. At low speeds the 'on' pulse width is 1 ms and the mark/space ratio is controlled by varying the 'off' period. At equal mark/space ratio the pulse width is 2 ms. Further increases in the mark/space ratio are achieved by increasing the 'on' time and the 'off' period is gradually reduced back to 1 ms and then at maximum speed a permanent 'on' signal is produced and the system is in full conduction. These waveforms are shown in Fig. 18.

The advantages of this system of pulse modulation are:

(1) Load current ripple is practically constant for all torques and speeds. Figure 19 shows a comparison of ripple content for all three methods.

(2) Gives a wide range of mark/space ratio to allow smooth speed control and braking.

(3) Allows simple smooth run out into full conduction.

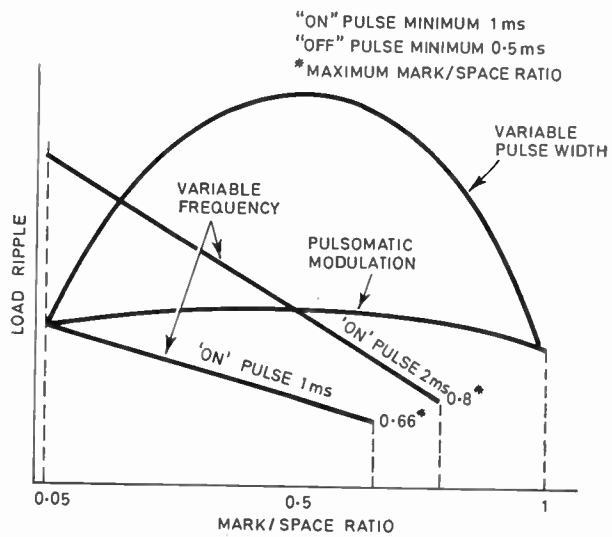


Fig. 19. Ripple variation with modulation system.

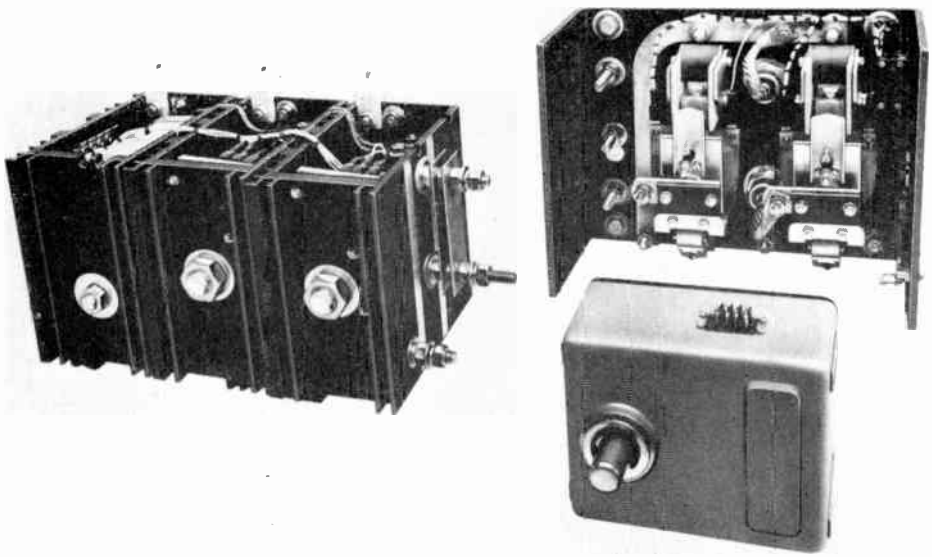


Fig. 20.
Pulsomatic controller units.

4.4. *Controlled Dynamic Braking*

As discussed earlier, when the motor is used as a generator to provide electrical braking, the field current must be controlled at a much lower level to limit the generated e.m.f. and avoid field saturation. On many controllers this reduced field current is maintained at a constant level during braking by applying a fixed 'on' pulse at a fixed frequency and since the excitation is constant, braking torque $\propto N$, i.e. the braking torque is high initially and decreases as the vehicle slows down.

The Pulsomatic braking system is controlled by the pedal position by detecting when the residual field is reversed and connecting a variable low level pulse generator to the field.

Since the excitation can be varied over the whole range, braking torque $\propto N\Phi^2$, and the braking effort is under the driver's control.

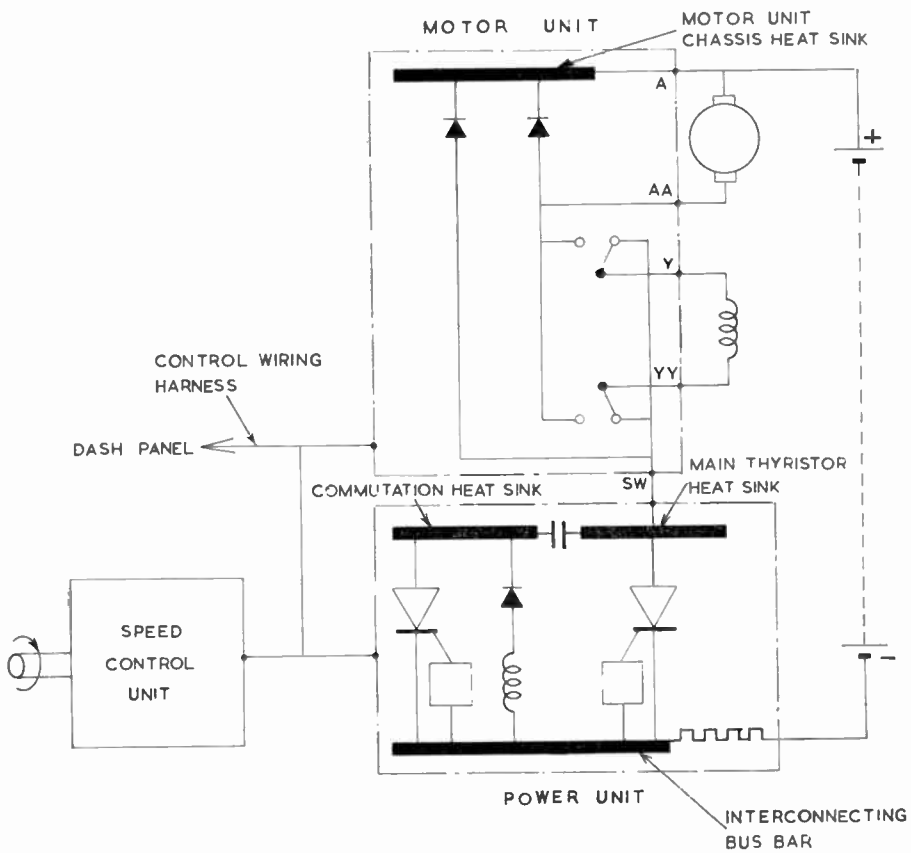


Fig. 21.
Unit interconnexions.

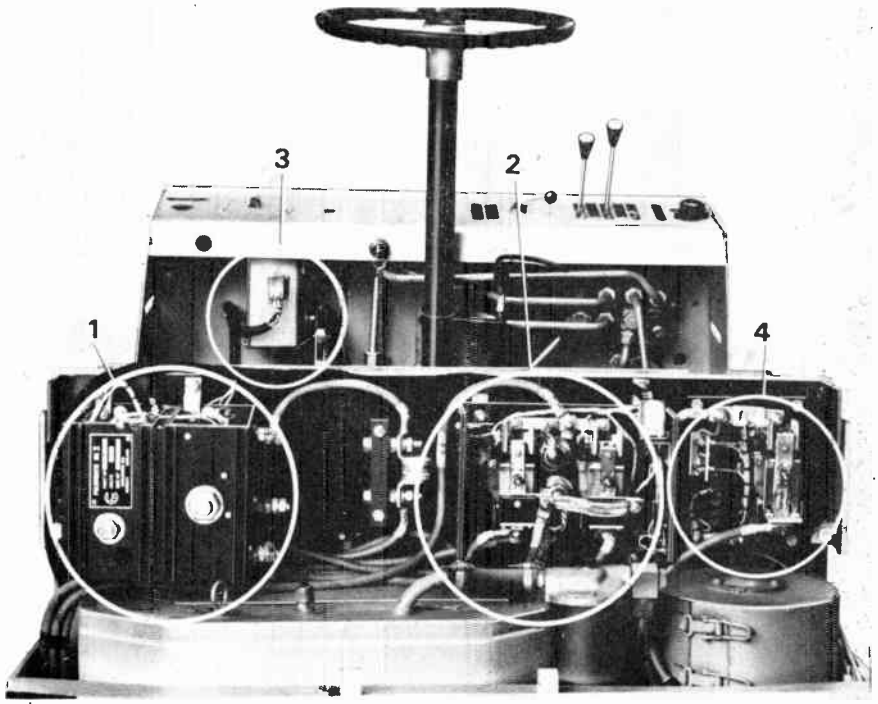


Fig. 22.
Typical pulse controller installation.

4.5. *Construction*

The cost and reliability of electronic equipment which is to be used under the arduous conditions of service demanded by industrial traction duties are greatly affected by the mechanical and thermal design of the units which make up the system.

The Pulsomatic system consists of three fully specified functional units:

(1) *Speed Control Unit.* This is a die-cast box with an output shaft which would be connected to the acceleration pedal or lever and contains all the control electronic circuits.

(2) *The Motor Unit.* This contains all the components associated with motor conditioning: the direction contactors, free wheel diode and braking diode. The chassis is of 6 mm aluminium which provides the diode heat sink.

(3) *The Power Unit.* This is the high-speed switch which controls the output to the motor. The construction is based on box-shaped extrusions to provide both heat sink capacity for the thyristors and an integrated robust construction. The centre bar which is used to fit the extrusions together is also a busbar connexion for one end of the battery supply.

Typical units are illustrated in Fig. 20.

This functional arrangement simplifies the interconnexions, reduces stray inductance and ensures it is constant on every unit and makes maximum utilization of the hardware. Figure 21 shows the unit breakdown in relation to the complete circuit and the simplification of inter-unit connexion and vehicle wiring is clearly seen.

This arrangement of fully specified fundamental units gives the vehicle designer complete flexibility and the high degree of standardization in the sub-assembly modules yields considerable cost savings and increases reliability.

Figure 22 shows a typical arrangement on a vehicle, 1 is the power unit, 2 is the motor unit, 3 is the speed control unit and 4 is the hydraulic pump control.

It has been convenient to refer to the traction motor in discussing control methods but of course these electronic pulse controllers can also be applied to the hydraulic pump motor on fork lift trucks and this can be an advantage particularly with high masts and fast lift speeds.

5. Acknowledgments

The author wishes to express his thanks to Cableform Limited for permission to publish this paper and also to acknowledge his gratitude to his colleagues, Messrs. Morton, Stevens and Thexton, who have been responsible for much of the original work described.

6. References

1. Mangan, M. F. and Griffith, J. T., 'Motors and Controllers for Electric Cars'. Electricity Council Research Centre Report No. M238—January 1970.
2. S.C.R. Manual—Section 6.2.4. General Electric—1967.

Manuscript received by the Institution on 9th October 1971. (Paper No. 1436/IC 60).