The Journal of the

BRITISH INSTITUTION OF RADIO ENGINEERS

FOUNDED 1925 INCORPORATED BY ROYAL CHARTER 1961

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

VOLUME 22 NOVEMBER 1961 NUMBER 5

ACTIVITY IN EDUCATION

THE aims, methods and administration of formal education in Britain are currently under scrutiny. Commissions and committees are considering the future of education at all levels, with a willingness—indeed, an eagerness—to examine new ideas and to see beyond the stockades of established tradition.

With the inception of the Diploma in Technology, technical education has had early warning of the winds of change, but only the most smug complacency could assume that all the cobwebs have thereby been swept away. New thinking about higher education cannot but involve a re-appraisal of higher technological education, if only because of the new context and environment in which scientific and engineering studies may have to exist.

Within the Institution one would have to look far to find complacency in regard to higher education. Rather is there a sense of urgency. A pressing awareness of immediate needs is tempered by a desire to look critically at the whole pattern of technological education in our own country and abroad. The recent decision of the Council of the Institution to inaugurate an Education Group is an expression of this sense of urgency.

The Education Group now takes its place with the five existing groups (Medical and Biological Electronics, Computer, Radar and Navigational Aids, Electro-Acoustics, Television). As was the case with these other groups, the formal constitution of the Group amounts to a recognition of existing activity in the particular field. Indeed there may be many who, aware of the part which the Institution has already played in preparing Reports and providing a forum for the discussion of education and training, may be surprised to find that the group structure had not already been extended to Education.

That it has now been extended is evidence of a resolve to respond to the increasing pressure from our members for redoubled activity. This pressure, which comes not only from our members in the teaching profession, may surprise those not familiar with our activities. To them we may quote the following words from the official History of the Institution: "The promotion of suitably specialized advanced studies is the first conscious purpose of (professional) institutions. It remains always one of their prime functions. In this, they must be leaders, since no other body is either so well qualified or so urgently motivated."

While the winds of change are deciding their magnitude and directions, the Education Committee and the Examinations Committee of the Institution continue their diligent surveillance of the syllabus of the Institution's examination. It is a truism that an examination syllabus in an expanding specialized technical subject is bound to be out of date—if only because due notice of a change of syllabus has to be given to the teaching establishments. Two factors, however, assist the Institution in keeping the syllabus as up-to-date as possible. The first of these is a close liaison with the teaching establishments. The second factor is a realistic recognition that Radio and Electronic Engineering must be studied as a subject in its own right.

E. W.

INSTITUTION NOTICES

The Canadian Proceedings of the Brit.I.R.E.

A new Institution publication has just been issued: *The Canadian Proceedings of the Brit.I.R.E.* has the aim of providing members in Canada and the U.S.A. with information on Institution affairs within those countries. The contents include reports on local meetings and abstracts of papers by Canadian and American authors. The *Canadian Proceedings* will be published quarterly and sent free of charge to members in North America; members in other parts of the world may obtain a copy from the Institution price 5s.

"New Techniques in Non-Destructive Testing"

The West Midland Section Symposium on the above subject will be held on Wednesday, 6th December, at Wolverhampton and Staffordshire College of Technology and will start at 10 a.m., lasting until approximately 5.30 p.m.

The following papers will be presented:

"A Study of the Physical Factors Affecting the Reliability of Ultrasonic Non-Destructive Testing". L. Kay, B.Sc.

"Automatic Charting of Ultrasonically Detected Flaws in Bar". M. D. Chattaway.

"The Ultra-Sound Image Camera". C. N. Smyth, M.A., B.Sc.(Eng.), B.M., B.Ch.

"Detecting Flaws in Steel Tube or Bar with a Rotating Coil in a Magnetic Saturating Field". W. H. Baker (*Associate*).

"The X-Ray Image Intensifier as an Inspection Tool and its Application to Stroboscopic Examination". C. E. Paine.

"Automatic Evaluation of Defect Severity by Shape and Size". D. R. Aldridge-Cox.

An introductory address will be given by J. A. Sargrove (*Member*) and recent films on automatic inspection will also be shown by Mr. Sargrove.

A detailed programme, including synopses of the papers, may be obtained from the Local Honorary Secretary, Major C. W. T. Weech, M.Brit.I.R.E., 5 Shelton Fields, The Mount, Shrewsbury, or from the Institution's Offices. Advance registration will be necessary (members £3 3s.; non-members £5 5s.).

Colour Television Symposium

Under the sponsorship of the Television Group Committee, a symposium on "Constant Luminance Colour Television" will be held on *Thursday, 14th December 1961* at the London School of Hygiene and Tropical Medicine, starting at 6 p.m. (Please note the change of date of this meeting from that allotted to the Television Group in the programme card.)

The symposium will consist of a main paper which proposes an interesting modification to the N.T.S.C. colour television system and four shorter contributions which comment critically on some aspects of these proposals.

The papers are as follows:

"A Constant Luminance Colour Television System". I. J. P. James, B.Sc. (*Member*) and W. A. Karwowski, B.A.

"A Colorimetric Study of the Modified N.T.S.C. System". W. N. Sproson, M.A.

"Fluctuation Noise in Two Colour Television Systems". A. V. Lord, B.Sc.

"The Relative Visibility of Random Noise over the Grey Scale". K. Hacking, B.Sc.

"Some Aspects of V.S.B. Transmission of Colour Television with Envelope Detection". G. F. Newell.

Summaries of the papers will shortly be sent to members of the Television Group; others may obtain details on application to the Secretary of the Television Group Committee at 9 Bedford Square. Members wishing to bring non-members to this meeting are asked to apply to the Secretary of the Group Committee for tickets for their guests.

National Council for Quality and Reliability

As a result of preliminary meetings organized by the British Productivity Council to discuss quality control and reliability methods, a National Council for Quality and Reliability has been formed. The Institution is a member body of the Council and its nominated representative is Mr. F. G. Diver, M.B.E. (*Member*), chairman of the Technical Committee. Mr. Diver has represented the Institution in the preliminary discussions referred to in the April 1961 *Journal* (page 290).

The main objects of the Council are:—

(a) to promote appreciation of quality and reliability and foster the procedures for attaining these;
(b) to co-ordinate the activities of corporate bodies working in the field of quality and reliability;
(c) to act as a central source of information for quality and reliability;
(d) to advise on training courses in quality and reliability procedures and to promote further developments in the field of education in these procedures;
(e) to collect and adminster funds for the purpose of carrying out the above objects.

A number of committees have been established to carry out this work. The Institution's representative has been appointed chairman of the Government/Industry Committee.

U.D.C. 621.19

Satellite Launching Possibilities

By

D. J. LYONS, B.Sc. (Eng.)†

Presented at the Convention on "Radio Techniques and Space Research" in Oxford on 5th–8th July 1961.

Summary: The subject is introduced by a discussion on the characteristics of varying orbits and the changes in performance and payloads associated with attaining these orbits. Methods of injection are then discussed; reference being made to part-coasting, continuous thrusting and part-aeroplane-borne trajectories. Control and guidance methods are surveyed and environmental problems imposed on payloads during injection outlined. Finally the description is given of the practical immediate possibilities offered by the *Blue Streak* satellite launching vehicle together with its possible future developments.

1. Introduction

This paper is a discussion of the practical aspects of launching satellites into terrestrial and circum-lunar orbits. It comprises two parts: the first dealing with the problems and possibilities in a general form and the second, showing what launching capability will be available for European use assuming that the joint European/Commonwealth plan to develop the *Blue Streak* satellite launching vehicle goes ahead.

2. Orbits and Payloads

2.1. *General Limitations on Orbits Possible*

With any given satellite launching system, that is with given stages, a variety of orbital missions can be accomplished. This can be done in two main ways.

(a) The payload can be altered up or down in weight. This will cause a corresponding decrease or increase in the final energy/lb of payload plus spent last rocket stage and the possible orbits will be changed. The take-off acceleration of the rocket will be a little changed by the increase or decrease in weight, but with more or less conventional rocket launchers this will not be appreciable. There may, however, be limitations to this process arising from stressing considerations of the upper stages, and the optimization of weight distribution through the stages might be upset.

(b) The payload can be altered up or down in weight accompanied by an equal decrease or increase in weight of the rocket stages, usually in fuel loading. The fuel weight alteration can be distributed through the stages in an attempt to

(i) keep the first stage flight path constant in order to keep aerodynamic heating and stressing constant and within design limitations during exit from the earth's atmosphere and

(ii) to give reasonably optimum weight sharing between the stages.

Of the two methods (a) and (b) the second is the one that will be more usually applied.

Now there are some obvious limitations that will apply to the process of changing orbit; when the payload is reduced to a very small amount a particular launching vehicle will have reached its maximum capability in terms of energy of the orbit, that is in the total energy/lb of payload (potential plus kinetic). This possible maximum is quite different for different types of launching vehicle. There are, however, a number of less obvious restrictions.

The first of these is the limitation on maximum orbit energy due to restrictions in direction of launch. This arises from the fact that at the moment of launch the rocket possesses a space velocity relative to the centre of the earth equal to the earth's surface velocity (or an aeroplane launcher's velocity relative to the earth's centre) at the point of launch. Since the orbital characteristics are measured in space axes, the fixed (for one configuration) additional velocity of the rocket launcher has to be added vectorially to the initial velocity; so greater energy orbits can be obtained when firing eastwards and taking advantage of earth rotation than when firing North or South or firing westwards.

Secondly, there will be, in general, problems of range safety that will limit possible direction of fire, and possibly effect limitation to the optimization between the various stage weights. It is quite clear that the reliability of rockets is not 100%, so that steps have to be taken to prevent damage to people and property in the event of a failure of any one of the stages to perform correctly. A part of the precautions taken is the provision of a good radar tracking system

† Head of Guided Weapons Department, Royal Aircraft Establishment, Farnborough, Hants.

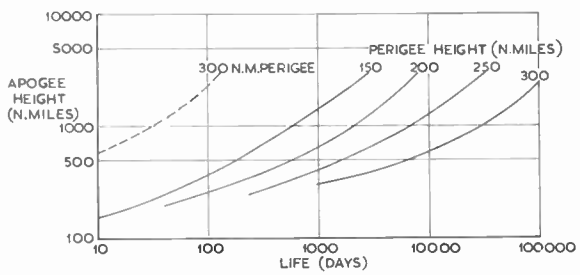


Fig. 1. Satellite life as a function of initial apogee and perigee.

——— $\frac{SC_D}{m}$ = 0·1 ft²/lb (typical satellite)

- - - - $\frac{SC_D}{m}$ = 100 ft²/lb (balloon)

S = cross-section area. C_D = drag co-efficient.

m = mass.

to follow the vehicle as it accelerates, and to watch for any large deviation from the expected path, so that a command signal can be sent to the missile either to destroy the missile or stop its engines so that no further deviation from the ballistic path can take place. So, for any given line of fire a range safety funnel must exist on the ground within which vehicles may land which have misbehaved. This funnel ideally must be clear of persons and property or sufficiently sparsely occupied so that the estimated chance of damage is acceptable to the nation which owns the property. If the funnel or part of it is over the sea, the sea area should ideally be clear of ships. Further as each stage burns out and despatches the next stage on its normal flight, the spent stages (certainly the earlier ones) will have insufficient velocity to orbit and they will fall back to the earth. The expected points of impact will have to be in planned impact areas, and the direction of launch from a given launching site will be limited so that acceptable impact areas can be provided. Some of the later stages may, of course, travel a very large distance (thousands of miles) before they impact and this could make the choice of impact area more difficult; this is offset, however, by the fact that much if not all of the spent stage will be reduced to metallic equivalent of matchwood under the combined action of the aerodynamic forces, and the heating effects during re-entry. These pieces or dust may have very low terminal impact velocities and may not incommode property or persons in the area of their fall.

There are further limitations which may be imposed by the type of guidance system and propulsion system; but these will be dealt with in more detail in Section 3 where methods of injection are described.

One consideration of orbital characteristics which may be very important in choosing a desirable orbit is the lifetime of a satellite in a particular orbit, or the rate of change of orbit shape, arising from the loss of energy due to atmospheric drag. The lifetime is both a function of the drag/weight ratio of the satellite vehicle in the orbit considered and also the shape of the starting orbit. Figure 1 illustrates the very large variations which occur. As examples, for a normal dense satellite, say an active radio satellite, a lifetime of three years would require a circular orbit of greater than 300 miles altitude or an apogee height up to 1000 or 2000 miles with a perigee height of 180 miles. A balloon type satellite would only last one day in a 300 miles circular orbit.

Summarizing, therefore, a variety of orbits may be obtained with one basic launching vehicle with a limit on the maximum orbital energy. Depending on the launch site there may be limits on direction of orbit, e.g. polar orbits may not be possible from some sites and equatorial from others purely on the score of safety.

2.2. *Payload-performance Exchange*

If a rocket combination of one or several stages is fired in vacuum and well away from any heavenly body (i.e. in a zero gravity field) the rocket will reach a given velocity relative to its starting condition; if the same rocket is fired still in vacuum but in a gravity field of any strength and along any path curved or straight, an integrating accelerometer fixed in body axes on the longitudinal axis of the body will read the

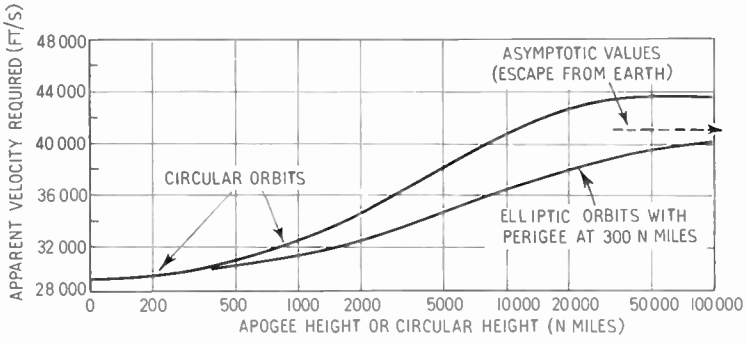


Fig. 2.
Apogee height or circular height (n. miles).

same velocity as that which would be reached in zero gravity field. This in rocket technology is known as the apparent velocity. The real velocity of the end stage of the rocket is of course less, when travelling away from the earth, say, as some of the thrust is being used to cancel out partly or wholly the gravitational attraction. This loss is called the "gravity loss"; in addition the aerodynamic forces during the flight through the planetary atmosphere cause further losses of energy. This loss will cause a decrease also in the apparent velocity as measured by the missile fixed integrating accelerometer. These losses could vary considerably with the flight path, the thrust over mass of the rocket combination and with the staging, but if optimum values are chosen for the parameters such as take-off acceleration, flight path against time and so on, it appears that there is generally a fairly constant relationship between the apparent velocity required and the energy of the orbit chosen for a whole range of possible rocket combinations. Thus assuming a certain orbit is required, a first estimate can be made of the performance of the rocket required without recourse to detailed trajectory calculation. This approximation is particularly good when estimating the change in performance required between a rocket launcher giving one orbit to the same launcher modified to give another orbit. Figure 2 shows the variation of "apparent velocity" capability required from the launcher for a variety of earth equatorial orbits, both circular and elliptic (1500 ft/s would have to be added for polar orbits to all these figures). Figure 3 is a companion picture to Fig. 2 and shows the "apparent velocity" requirements for Moon circular orbit missions starting from earth; included in these figures are the retro-rocket velocity requirements to stabilize the satellite in the Moon orbit. It is clearly demonstrated that the rocket performance required for Moon orbits is, at the least only slightly greater (3000 ft/s) than highly elliptical orbits to the altitude of the Moon's orbit, but the low transit times may require the provision of a much better performance rocket.

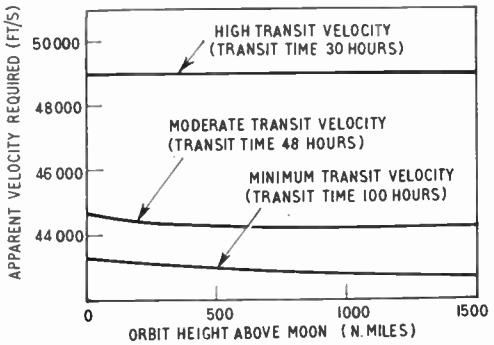


Fig. 3. Booster apparent velocity requirement for lunar circular orbit missions (equatorial launch.

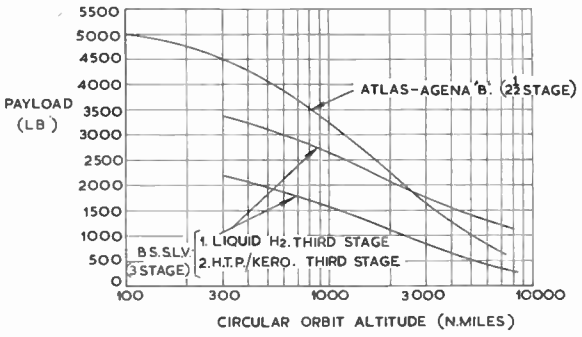


Fig. 4. Effect of orbit mission of payloads obtainable from typical boosters (circular polar orbits.

Now it is a fairly simple process for the rocket designer to compute the change in payload for a given change in orbit; curves of the type given in Figs. 2 and 3 give the change in velocity required and the designer can interpret this change in terms of the change in propellent required and payload alteration. It must be realized, however, that the variation in payload weight with orbit is not a simple, say, percentage change. Figure 4 illustrates this point; there it is shown that for typical launching vehicles one may be more capable than the other of putting a larger payload into a low energy orbit but the reverse may apply for higher energy orbits. This situation may be simply attributed to the different all-burnt weights of the end stage without the inclusion of the payload. This dead weight also has to be put into orbit with the payload, and so, if it is heavy, most of any extra fuel energy may go in accelerating this weight leaving a less and less proportion for the payload; if the dead weight is relatively light in comparison with the payload in the low energy orbit a larger proportion of the extra fuel energy goes in producing extra payload energy and so a greater payload can be achieved. Since an increase in the number of stages results in lower dead-weights of the last stage, it is in general an easier task to provide useful payloads in the higher energy orbits with these than with say 2-stage vehicles. In the cases illustrated the 3-stage vehicle shows a better capability at higher altitude orbits than the $2\frac{1}{2}$-stage (which is very little different to a 2-stage in the satellite launching role). Increase of percentage structural or dead weight, and decrease of specific impulse will also cause further degradation in payload as the orbit energy is increased.

2.3. *Staging*

Obviously, for a given number of stages, with given energy fuels and given structural efficiency, the maximum possible velocity is limited and even an infinite increase in size will lead asymptotically to a limiting apparent velocity. Figure 5 illustrates this

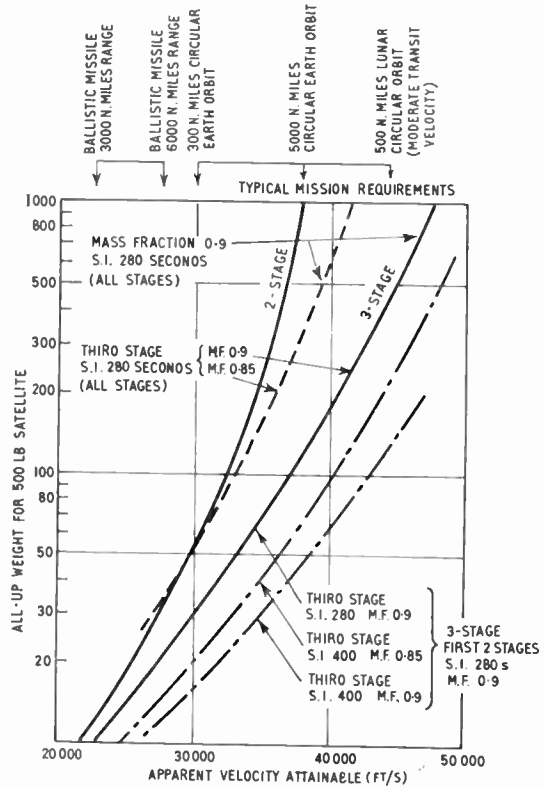


Fig. 5. Apparent velocities attainable as a function of payload and stage performance.

situation, by plotting the all-up weight at take-off necessary to put a 500 lb payload up to a series of given apparent velocities. Assuming a mean specific impulse for the propellants of 280, and a mass fraction $\left(\text{i.e. } \frac{\text{fuel weight}}{\text{total weight} - \text{total carried load of the stage}}\right)$ of 0·9 for each stage, 3 stages rather than 2 are needed if apparent velocities in excess of 36 or 37,000 ft/s are required. The sensitivity of the performance to changes in the mass fraction from 0·9 to 0·85 and to changes in the specific impulse of fuel in the first stage is also demonstrated for typical examples. It can be shown by simple theory that the performance of a rocket is a maximum for an infinite number of stages; however, inclusion of a number of practical considerations, such as separation devices, increased structural loads, systems for tracking, instrumentation and control which do not scale with size of the stage, and all of which increase the dead-weight proportion of the smaller stages, bring the optimum number of stages down to a reasonably small figure. Graham† suggests this number is 4 to 5, but this must depend on the particular case being examined. For cases where the payload is a very large percentage of the take-off weight a much smaller number of stages is more efficient than when the payload is a very small percentage of the all-up weight. However, there are other many practical factors which will tend to favour the rocket with the smaller number of stages; complexity is less, reliability is greater, development cost is smaller (due to the smaller numbers of items to be developed as well as fewer failures to be paid for) and manufacturing cost is smaller as the number of stages is reduced for a given maximum weight rocket.

Of course, in reality the large cost of booster vehicles and their development, means that the greatest possible use of existing vehicles will be made and the rocket designer's problem is usually one of making the optimum use of existing vehicles, or those planned for a variety of tasks, rather than designing the optimum vehicle for each application. Another related and important factor is the inclusion of some "growth factor"; this depends on many factors inherent in the design of particular rockets but some general considerations apply. The greatest growth potential for the least cost will generally be available when the launching vehicle is originally planned to have a smaller number of stages and/or the end stages have a relatively low efficiency: then rapid improvement in performance can be obtained by either increasing the number of stages or by improving the efficiency of the upper stages, typically by redesigning for higher energy propellents.

2.4. *Circular Orbits for Radio Use*

A number of references have been made to circular orbits because of the increasing importance of these orbits particularly for use in radio applications. There are a number of reasons for this:

(i) Earth-stabilized satellites on circular orbits possess constant angular rotational speeds so that orientation of satellite aerial systems to make use of maximum gain aerials becomes generally much easier.

(ii) The motion of the satellite round the orbit, except for earth rotation under the orbit, gives reasonably consistent conditions at tracking/transmitting stations and for higher circular orbits relatively constant range. This eases the design of both ground and satellite borne systems. In particular equatorial circular orbits where earth rotation effects are more or less absent, give almost ideally constant tracking conditions.

(iii) Perturbations of circular orbits are much smaller so that prediction and control of tracking systems is much easier.

(iv) If the orbit time divides exactly into 24 hours, this further eases ground aerial programming.

† B. Graham, "Communal Applications of Satellite Boosters" I.A.S. Paper No. 61–58.

3. Methods of Injection

3.1. *All-Rocket Systems*

Practically all known satellite launching attempts have been made with all-rocket systems. In such a system the initial launching is vertical, as maximum efficiency is achieved with very low initial acceleration (sometimes as low as $\frac{1}{4}$ g), but as the orbital velocity is required in a more or less horizontal direction the flight path is curved towards the horizontal as soon as is possible in order to start acquiring horizontal velocity. The rate of turn is limited by requirements to keep the speed down whilst the booster is still in the atmosphere in order to keep aerodynamic heating and loads within reasonable proportions. As the stages burn out they are jettisoned and the succeeding stages ignited just before, during or after separation. In the most usual system all but one of the stages are burnt in rapid succession and have fairly high mean accelerations, firstly so as to derive the maximum possible energy increase for given momentum increases and secondly to minimize the gravity losses. Since, however, this results in burn out at a fairly low altitude and the launching vehicle must enter the required orbit under power, a coast period is usually inserted between the penultimate stage and the end stage while altitude is being gained, and the final stage is ignited near orbital altitude to increase the energy to that needed to remain in the desired orbit. This might be termed the thrust-coast-thrust type of injection. It has the added attraction that on a simple mathematical approach it has the virtue of being the most efficient method of injection from the point of cessation of thrust of the first stages; its disadvantages are also clear: orientation systems have to be carried to hold spatial axis references for controlling direction of the end stage thrust, and there may be additional problems caused by the necessity to start the end stage engines from conditions of free fall when liquids will not be nicely placed in their required position in the "bottom" of the tanks.

An alternative method which has been evaluated in U.K. has been the elimination of the coast period and the substitution of a continuous but very low thrust end stage so that cessation of this thrust and attainment of initial orbit conditions are met at the same time. Figure 6 illustrates the general features of the two methods for typical optimized launching rockets, assuming constant first and second stages. The advantages of this latter method are numerous:

(i) The low thrust lowers the engine weight very considerably which in turn permits increase of nozzle size, so that greater propellent efficiency (higher specific impulse) can be obtained.

(ii) The continuous thrust motor can and must be used to orient the stage so that a separate orientation system is unnecessary.

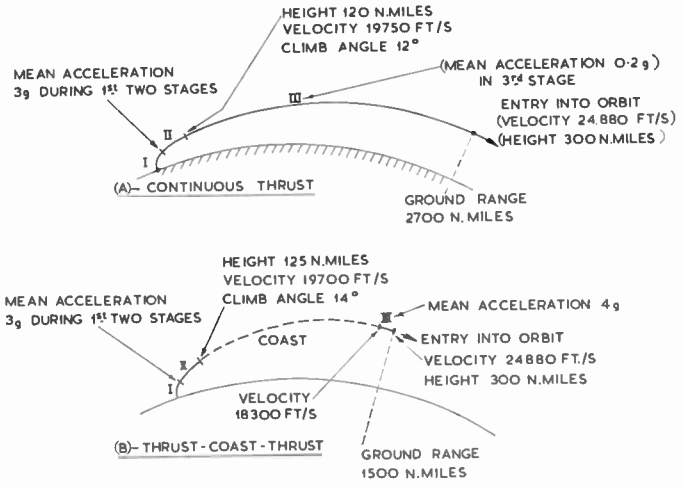


Fig. 6. Diagram showing difference between "thrust-coast-thrust" and "continuous thrust" methods of injection into orbit (300 n. miles circular orbit. Three stage booster).

(iii) For reasons of (i) and (ii) the dead weight is reduced and the resulting increase in performance offsets the simply mathematically calculated loss in performance referred to above.

(iv) Light-up problems can be eliminated if desired.

(v) Loss of cryogenic propellents due to solar heating, heat leakage, etc., can be avoided or minimized during the continuous thrust programme (this is not completely possible during coast).

(vi) If radar guidance is used, the very low acceleration permits very long smoothing times both during climb and cut-off so that radar guidance becomes very accurate and early correction of orbit possible.

3.2. *Effect of High Perigee Requirement*

If a high perigee is required for the orbit for either the thrust-coast-thrust system or the continuous thrust system, it is clear that during the injection phase the whole of the perigee altitude must be gained. Since for the sake of efficiency the climb angle must be small this inevitably means that:

(a) the end stage or stages must possess orbital or very near orbital velocity at the beginning of the climb and

(b) that the point of final injection must be a long way round the earth from the point of launch.

This general effect is demonstrated by Fig. 7 in which the thrust-coast-thrust system is assumed (though the same general results apply to continuous thrust trajectories) in which the variation in apparent velocity required with ground range during injection is presented. For low orbits, 300 n. miles, it is seen that down to 3000 miles range there is little change in rocket performance required, but for 5000 n. miles circular orbits a sharp rise in rocket performance is required if the coasting phase is reduced below one-half of a revolution round the earth.

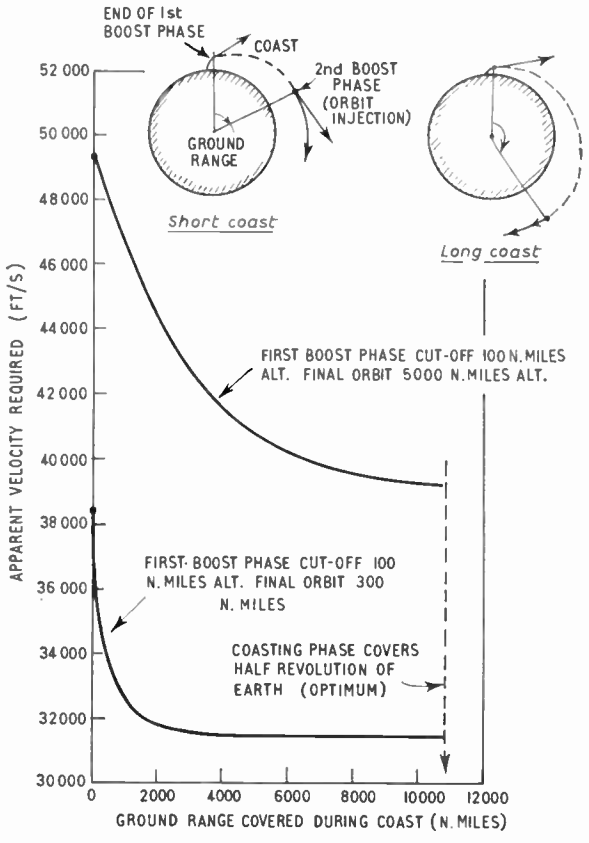


Fig. 7. Loss of efficiency due to reduction in total length of boost trajectory.

This has two major effects; any guidance or tracking stations must cover a much larger flight path if no uncovered periods are to be left, and final cut-off has to be observed from an area remote from launch; and secondly, if a continuous thrust system is used, extremely low values of motor thrust are necessary. The use of a very low thrust in the continuous thrust system has a very attractive feature in that it permits an even longer radar smoothing time than for injection into low orbits; thus even though the radar range is very much greater for the high perigee orbits than the low, the required angular accuracy of the radar used is virtually the same in order to achieve roughly constant velocity accuracy for the two classes of orbits.

3.3. *Aeroplane Launch*

Though the major portion of the speed needed for injecting a satellite into orbit must be given outside the atmosphere, in order to avoid too high aerodynamic heating effects, there are some potential advantages in using an aeroplane-like device for the first stage. The most obvious of these is the possibility of easier recovery and re-use of the first stage by making the first stage an aeroplane which can land in the normal way. Employment of present day aircraft is possible in this role though the performance gains are not striking; an aircraft say with a high subsonic speed of $M = 0{\cdot}8$, will only reduce the apparent velocity requirement of the rocket stages by 1500–2500 ft/s. This reduction is larger than the actual velocity of the aeroplane stage by virtue of the altitude of the aeroplane at rocket launch (taken to be 40 000 ft) with the consequential increase in specific impulse and the avoidance of much of the aerodynamic drag losses in the rocket stage. The reduction in weight of the rocket stages required to be carried by the aeroplane as compared with an all-rocket vehicle depends a lot on the number of stages assumed and to some extent on orbit desired: if the aeroplane is counted as one stage, for the subsonic aircraft quoted above the total *rocket* weight may be about the same or even increased, but a 30–50% rocket weight decrease may be obtained by keeping the same number of rocket stages as before (see Fig. 5). An increase in the aircraft speed to $M = 5$ will materially improve matters, the weight reduction for an equal number of rocket stages being of the order of 65–75%. The economies involved in the use of aircraft first stages need a lot more study.

3.4. *Guidance and Control*

3.4.1. Control

The principle means of control used in satellite launchers is thrust vector control: that is the thrust line, which for steady flight must pass through the centre of gravity of the vehicle, is temporarily deflected

so that an angular acceleration is applied to the vehicle. When sufficient heading angle change is obtained the thrust line is re-centred and the subsequent speed increase is applied in a new direction, so permitting control of the flight path. On many vehicles thrust vector control is used throughout flight, but it has proved beneficial on some specialized vehicles, for instance some with high acceleration first stages or aircraft first stages, to use aerodynamic control for the first stage.

The control system usually holds the missile to a steady flight path by an autopilot deriving its space reference from some form of gyro, very often with the addition of a missile borne automatic programme of heading changes to provide the predicted flight profile.

3.4.2. Guidance

The task of guidance system is to monitor the behaviour of the launching vehicle under autopilot control and to apply superimposed command heading changes designed to modify the injection path in the best possible way, subject to certain boundary conditions, into the required orbit. There are at present two main ways of performing the navigation function, by radar means and by internal inertial navigation, though there are various combinations of the two, and, in fact, the autopilot and heading programme can be regarded as a form of perhaps inaccurate inertial navigation. It should not be forgotten, however, that some form of accurate radio tracking means is needed even if an all-inertial guidance system is used, to provide information to ground control.

It might be thought that an ideal flight path could be pre-calculated before flight and the vehicle fired to follow this path; there are, however, a number of perturbations which in general do not permit the achievement of this state. Wind disturbances, thrust variations if a variable thrust device is not fitted, variations in the mass flow of propellent, all cause alterations in the flight path, so that a computer is necessary to work out the best compromise course to steer when it is supplied with measurements of the deviation from the flight path. For an inertial system this computer is in the launching vehicle, for a radar monitored system it will be on the ground. The advantages of the radar system are that the most complex part of the system is on the ground and the vehicle borne parts of the system can be fairly simple and light, beacons or transponders; the disadvantages are that complex and usually expensive ground equipments are necessary, and that fairly strong limitations are placed on directions of fire, in order to avoid the installation of multiple guidance systems. The advantages of the inertial system are that the system is self-contained in the vehicle and there is virtually no limitation on direction of fire: the disadvantages are that it is relatively heavy, is fairly expensive and is lost at each firing. It is difficult to predict the accuracy potential of both systems but since for most existing injection requirements both radar and inertial systems can be made sufficiently accurate, accuracy is not usually a criterion for the choice between the two systems. This is illustrated by Table 1 where the sensitivity of various orbit parameters to velocity errors are tabulated; a few tens of

Table 1

Errors in Space Mission due to Velocity Errors at Injection.

I CIRCULAR ORBITS OF THE EARTH.

CIRCULAR ORBIT HEIGHT (N MILES)	APOGEE OR PERIGEE HEIGHT. ERROR DUE TO HORIZONTAL VELOCITY ERROR (N. MILES PER FT/S)	APOGEE AND PERIGEE HEIGHT. ERROR DUE TO VERTICAL VELOCITY ERROR (N. MILES PER FT/S)	ORBIT TIME ERROR. DUE TO HORIZONTAL VELOCITY ERROR (SECS PER FT/S)
300	0·60	0·15	0·69
1,000	0·78	0·20	0·98
5,000	2·04	0·51	3·52

II ELLIPTIC ORBIT OF THE EARTH.

PERIGEE HEIGHT. 300 N. MILES
APOGEE HEIGHT. 100,000 N. MILES
} CUT-OFF AT PERIGEE.

ERROR IN APOGEE HEIGHT DUE TO ERROR IN HORIZONTAL VELOCITY:- 180 N. MILES PER FT/S

ERROR IN APOGEE AND PERIGEE HEIGHTS DUE TO ERROR IN VERTICAL VELOCITY IS VERY SMALL.

III LUNAR IMPACT TRAJECTORIES.

CUT-OFF AT 300 N. MILES ALTITUDE ABOVE EARTH :-

TRANSIT TIME (HOURS)	PERMISSIBLE HORIZONTAL VELOCITY ERROR TO GRAZE MOON * (FT/S)	PERMISSIBLE VERTICAL VELOCITY ERROR TO GRAZE MOON * (FT/S)
100 (MINIMUM TRANSIT VELOCITY)	± 12	± 900
48 (MODERATE TRANSIT VELOCITY)	± 180	± 150
30 (HIGH TRANSIT VELOCITY)	± 50	± 90

* MOON RADIUS IS 940 N. MILES

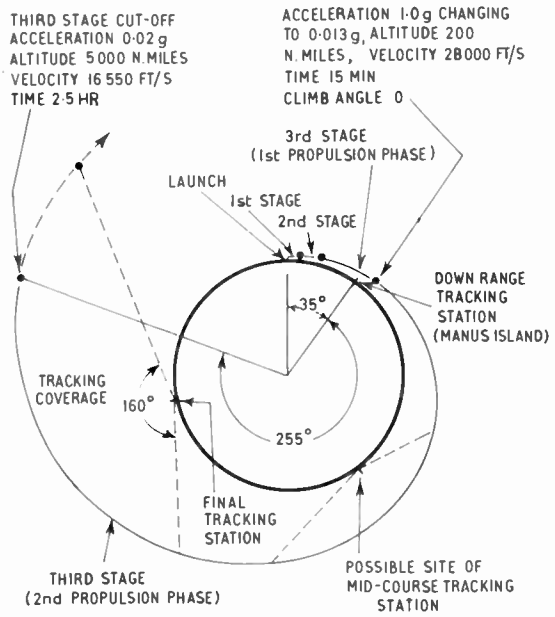


Fig. 8. Launch trajectory of vehicle for high (5000 n. miles) altitude circular orbit. Liquid hydrogen/liquid oxygen third stage.

feet per second are usually adequate for most missions. An exception must be made for station keeping satellites which may need extremely accurate initial velocities; a form of radio or optical tracking would seem essential for this task, and then the objective is only likely to be achieved by orbit adjustment rather than by accurate inertial injection and the problem is mostly a satellite problem rather than entirely a launching problem. It is of interest to draw special attention to the guidance of high circular or any high perigee launchings; as pointed out in Section 3.2 the high orbit is obtained most economically by an initial launch into a low orbit followed by an orbit transfer to the high orbit (coast or continuous thrust). This orbit transfer when done economically will be roughly a half revolution round the earth for the coast system (Hohmann transfer) and rather more for a continuous thrust system. A coasting system using radar guidance will probably need a minimum of two guidance stations in the launch area (one in the immediate vicinity of launch, one at transfer orbit injection) and one at final orbit injection: a conventional inertial guidance system could be used. A continuous thrust system may need an additional radar guidance station for mid-course guidance (see Fig. 8) which can be omitted if the thrust programme is accurate enough (within $\pm 1\%$ in the end stage); in an inertial guidance system a special low acceleration system is needed during the low thrust phase, though there are probably no fundamental obstacles in providing this.

3.5. *Launching Environment*

It must be of fundamental interest to designers of satellites to know the environment provided by the launching vehicle during the boost phase and the differences between this environment and that the satellite will have in orbit. The most important differences are in acceleration and vibration. The maximum longitudinal accelerations depend a lot on booster design and may vary between about 5 and 12 g; lateral accelerations are usually fairly low, under ± 1 g, except perhaps in aircraft launch, when this figure may rise by a factor of 2 or more. The vibration environment again is fairly dependent on vehicle design, but a typical design vibration envelope which is given in Fig. 9 shows that it is believed that the level of vibration in the area of the satellite is unlikely

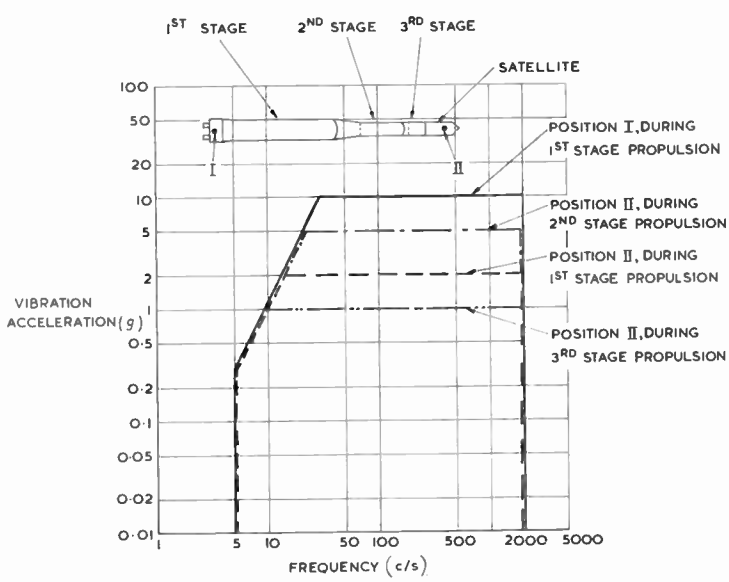


Fig. 9.
Typical vibration envelopes for various positions during boost.

to rise above an effective level of ± 1 to 2 g over the normal spectral band. These vibration levels bearing in mind the impressed d.c. accelerations, should cause few design problems.

The aerodynamic heating during exit through the earth's atmosphere does provide a serious problem. Figure 10 shows typical temperature variations with

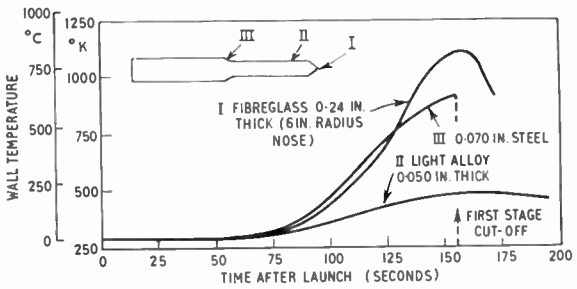


Fig. 10. Typical temperatures resulting from kinetic heating during boost.

time for the skin at varying portions along the missile, with varying material also included. It is clear that at the forward end high temperature skins are practically unavoidable and the satellite has to be protected from both convection and radiation heating. Fortunately the heating stops on typical trajectories at a time of 3 minutes or so, so that the heat protecting devices, which are heavy, can be discarded if desired: the *Blue Streak* satellite launching vehicle proposals include provision for jettisoning satellite heat protecting covers at about $3\frac{1}{2}$ minutes from launch.

4. The Blue Streak Satellite Launching Vehicle

4.1. *Description*

This section gives an outline of the type of launching vehicle and its capabilities which will be available if the Anglo-French proposals for a joint European/Commonwealth multi-stage satellite launching go ahead. The present proposals are based firmly on the use of the British rocket *Blue Streak* as the first stage and the use of a French rocket, as yet unnamed, for the second stage.

No firm joint European proposals have been made regarding the design and construction of a third, end stage for final injection into orbit, but in order to demonstrate the capabilities, two propellent combinations have been considered in assessment of possible third stages. The first, hydrogen peroxide (h.t.p.)/kerosene, a conventional, relatively low energy propellent combination, was considered for its comparative ease of development. The second, liquid hydrogen/liquid oxygen, yielding a greatly improved performance in terms of payload, especially in distant orbits, demands more extensive development. Only the final stage need be varied to meet a widely varied series of orbit requirements without significant departure from optimum staging conditions. Figure 11 shows the overall configuration. At the payload position, for illustration, is shown the outline of jettisonable fairings covering a typical low density solar cell powered satellite for communications work.

The first stage, *Blue Streak*, stripped of the equipment associated with its military guidance and warhead, requires minor modifications of the separation bay to carry the second stage. It is propelled by two pump-fed liquid oxygen/kerosene engines, each rated at 137,000 lb thrust at sea level, and in course of being uprated to 150,000 lb thrust. The mass fraction of 0·93 of *Blue Streak* which includes an allowance for unburnt propellent has been achieved by the use of very thin stainless steel tank walls, and reliance upon internal pressurization for structural stiffness.

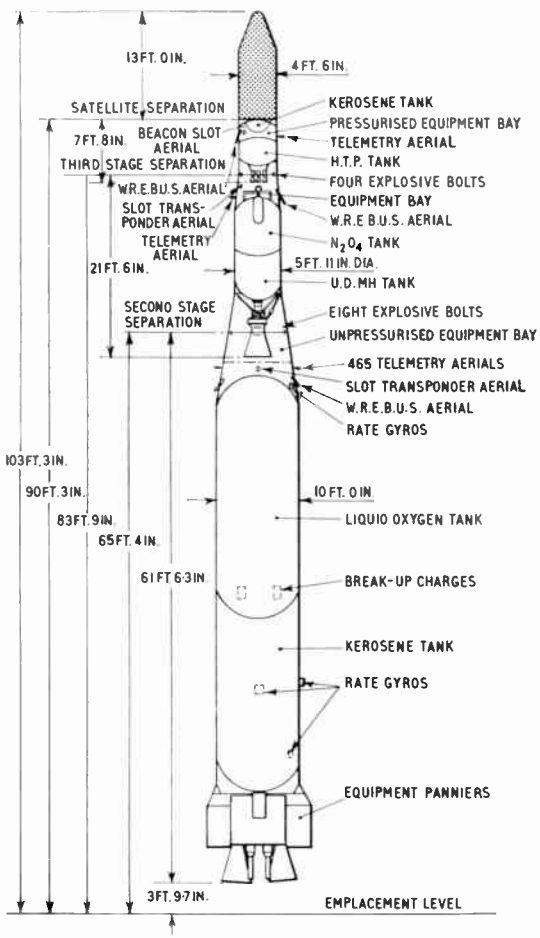


Fig. 11. Satellite launching vehicle.

The main structural sections are:

(i) Separation bay (originally the guidance bay).
(ii) Tanks, oxidant and fuel.
(iii) Propulsion bay.
(iv) Equipment fairings.

The propulsion bay is of aluminium alloy.

In flight the oxidant tank is pressurized by gasification of liquid oxygen, the fuel tank by gasification of liquid nitrogen. These cryogenic liquids are converted to the gaseous state by passage through heat exchangers located in the turbine exhausts.

The RZ2 engine, of which there are two, consists basically of an integral power pack comprising gas generator, propellent pumps, gear box and turbine together with the thrust chamber assembly. Liquid oxygen and kerosene in a fuel-rich mixture, are fed to the gas generator to produce relatively low temperature gas for the turbine drive. This turbine operates both fuel and oxidant combustion chamber injection pumps.

Each chamber is gimballed in two planes to provide control of the vehicle in pitch, yaw and roll. The chambers are moved by hydraulic actuators, high pressure oil being obtained from pumps on the main propellent supply turbine gearboxes.

The proposed French second stage will be propelled by a liquid propellent engine using nitrogen tetroxide (N_2O_4) and unsymmetrical dimethyl hydrazine (U.D.M.H.) with a sea level thrust of 25 tons. The propellents are pressure-fed by the single combustion chamber by means of a solid charge gas generator. The thrust chamber is gimbal-mounted for control in pitch and yaw, roll control being achieved by auxiliary jets.

In addition to the propulsion system, control system actuators and electronics, the second stage contains equipment for the initiation of engine light

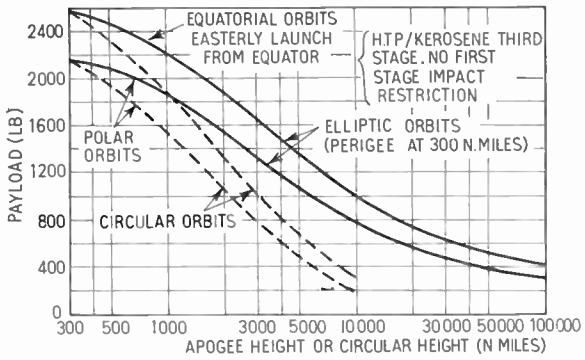


Fig. 12. Nominal payloads for circular and elliptic orbits.

up on separation from the first stage, equipment for initiation of second/third stage separation, and rate gyros for autopilot stabilization during the period of second stage operation.

A third stage using the continuous thrust principles would need an engine working at a relatively low thrust/level of between 1000 lb and 3000 lb (for the low perigee orbits), which would be started during separation from the second stage.

The engine proposed for an h.t.p./kerosene third stage is a four-chamber design, each chamber pivoted about one axis for steering in pitch, yaw and roll as in *Black Knight*. With low thrust and four chambers, very high nozzle expansion ratios, of 1000/1, are possible without undue chamber size, length and weight.

To minimize overall stage weight, by reducing propellent tank weight, pump feeding of propellents is to be preferred in this particular design. The propellent pumps would be driven by a hydrogen peroxide steam turbine.

It is possible to meet the varying orbital requirements by exchanging payload weight for propellent

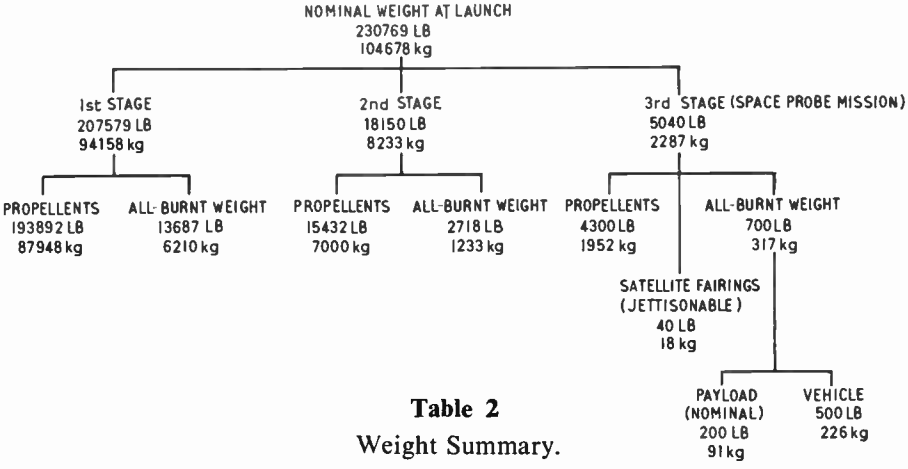


Table 2
Weight Summary.

weight in the third stage whilst maintaining constant the overall weight of the third stage plus satellite payload at some 5000 lb, that is, the third stage incremental velocity can be increased at the expense of payload. The tank volume, only, is altered to suit the orbital mission, allowing the remainder of the third stage, including the engine and all equipment, to remain sensibly unchanged. A weight breakdown of the whole launching vehicle is given in Table 2.

4.2. *Performance*

Figure 12 shows the estimated payload capability of the launcher with the illustrative h.t.p./kerosene end stage. These are based upon weight breakdowns as in Table 2, and a sea level take-off thrust of 300 000 lb (i.e. $2 \times 150\,000$ lb) and with exchange between payload and propellent weight in the third stage with mission, as mentioned earlier. As will be seen, the payloads vary from 1 ton actual payload in a 300 n. mile circle to 500 lb in a 6–7000 mile circle or in an elliptical orbit out to about 50 000 n. miles from a 300 mile perigee. In obtaining these results, stage weights and flight programmes have been optimized as far as possible. In general the stage weight/payload curves are fairly flat around the optimum point, and it has been found advantageous to choose end weights slightly off optimum on the "light" side for a number of reasons, such as, for example, the location of the impact point of the discarded first stage in an acceptable area.

The "payload into orbit" figures are considerably improved, as would be expected, using a liquid hydrogen/liquid oxygen third stage, as will be seen from Fig. 13. This is especially marked in distant orbits, where the third stage effective incremental velocity becomes a much larger proportion of the whole mission characteristic velocity. A payload of about $1\frac{1}{2}$ tons can be put into a near 300 n. mile orbit but $\frac{1}{2}$ ton can be placed in a 7000 mile circle.

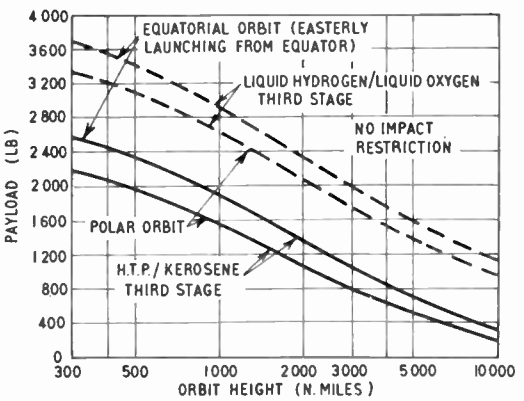


Fig. 13. Effect of introducing liquid hydrogen/liquid oxygen 3rd stage on the payloads into circular orbits.

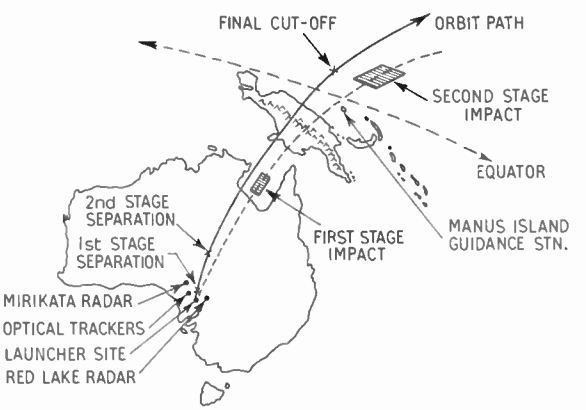


Fig. 14. Guidance stations and impact areas.

As expected, the optimum third stage weight with liquid hydrogen/oxygen motor increases (due to the fact that this stage is now more efficient and should therefore be used to provide a larger fraction of the total velocity increment). The optimum third stage weight (including payload) will be in the region of 7000 to 8000 lb.

Due to the increased third stage weight it has been found that the third stage thrusts should also be increased somewhat, compared with the low energy third stage case.

It is clear that the capabilities of the *Blue Streak* satellite launching vehicle as shown in Figs. 12 and 13 will cover an enormously wide variety of possible radio satellite requirements but if it is found necessary, still further development potential exists by converting the second stage to liquid hydrogen, first by itself and later with two stages above *Blue Streak*.

4.3. *Facilities*

The major facilities for *Blue Streak* already exist and a large capital expenditure has been invested to obtain this situation.

There are missile and engine testing stations in use at the Spadeadam Research Establishment in Cumberland in the North of England. At the proposed launching site at Woomera in Australia the major proportion of the work has been done.

Figure 14 shows a proposed layout of the launching facilities in Australia to be used for a polar launch into a low orbit: many of the necessary facilities exist in the launching area such as the guidance radars: the down range equipment such as the down range guidance station are as yet only projected.

Manuscript received by the Institution on 22nd June 1961. (Paper No. 677.)

APPLICANTS FOR ELECTION AND TRANSFER

As a result of its meeting on 26th October the Membership Committee recommended to the Council the following elections and transfers. The names of 89 students registered at this meeting will be published later.

In accordance with a resolution of Council, and in the absence of any objections, the election and transfer of the candidates to the class indicated will be confirmed fourteen days after the date of circulation of this list. Any objections or communications concerning these elections should be addressed to the General Secretary for submission to the Council.

Direct Election to Member

MARINER, Peter Frederick. *Radlett, Hertfordshire.*

Transfer from Associate Member to Member

MORTON, Alexander Hugh, B.Sc. *Great Malvern, Worcestershire.*
MOTHERSOLE, Peter Leonard. *Reigate, Surrey.*
PUDNER, Anthony Serle, M.B.E. *Epsom, Surrey.*

Direct Election to Associate Member

ADAMS, Roy Bertram. *Hemel Hempstead, Hertfordshire.*
COCKBAINE, David Robinson. *Swanage, Dorset.*
DANE, Edward Raymond Peter. *Faversham, Kent.*
DEARNLEY, John Dean, B.Sc. *Portsmouth, Hampshire.*
DEXTER, John Neville. *Smallfield, Surrey.*
HAUNG, Han Chao, Ph.D., B.Sc. *Kuala Lumpur.*
HUSSAIN, Major Mohammed Sardar, B.A. Pak. Sigs. *Rawalpindi.*
*KYNASTON, John Alfred Charles. *Cardiff, Glamorgan.*
NAIR, Lt. Bala Krishna Sreekumar, B.Sc., I.N. *Poona, India.*
WALKER, Major Ronald Cyril, R. Sigs. *Little Malvern, Worcestershire.*

Transfer from Associate to Associate Member

HALE, Leslie. *Stevenage, Hertfordshire.*
MAGUIRE, Donald Walter. *Enfield, Middlesex.*
WELLER, William Frank Edward. *Kenton, Middlesex.*

Transfer from Graduate to Associate Member

BEAVER, Captain Malcolm J., B.A.(Cantab.), R. Sigs. *Catterick Camp.*
BOND, Martin Edward. *Croydon, Surrey.*
CHEW BAK KHOON. *Singapore.*
DAVIES, Stanley James. *Farnborough, Kent.*
FIDLER, David Morris, B.Sc. *New Milton, Hampshire.*
FILES, William Henry. *Penketh, Lancashire.*
HAUGHEY, Peter. *Wirral, Cheshire.*
JAGER, William Frederick. *Thames Ditton, Surrey.*
JONES, David Lloyd. *Hythe, Hampshire.*
ROBSON, Alan. *Wells, Somerset.*
SAPSFORD, George. *Basingstoke, Hampshire.*
WOODS, John Frank. *Bromley, Kent.*

Transfer from Student to Associate Member

BENTOB, Meyer Toby. *Montreal, Canada.*
BLAKE, Frank George. *Billericay, Essex.*
KEEBLE, Ronald Sidney. *Holmer Green, Bucks.*
MILLWARD, Hubert Riley. *St. Annes-on-Sea, Lancashire.*
MITCHELL, Robert. *Stevenage, Hertfordshire.*
SANDERS, Frank Bundick. *Leigh-on-Sea, Essex.*
SAUNDERS, George Brian. *Nuneaton, Warwickshire.*

Direct Election to Associate

CARTMELL, Peter Airdrie. *Seascale, Cumberland.*
HALE, Peter Edward. *Romford, Essex.*

Transfer from Student to Associate

GENT, Leslie Frederick. *Stevenage, Hertfordshire.*
HOOKER, Spencer William George. *Eastleigh, Hampshire.*
KEANE, James. *Montreal, Canada.*

Direct Election to Graduate

AHMAD, Lieut. Masood, B.Sc.(Eng.). *Dacca Cantt, East Pakistan.*
ALEXANDER, Simon. *Bristol.*
BAKALL, Michael David, B.Sc. *Carshalton, Surrey.*
BALLARD, Thomas John. *Ashford, Middlesex.*
BANKS, Peter Henry Thomas. *Hillingdon, Middlesex.*
BARLOW Andree Kurt. *Didcot, Berkshire.*
BARTLETT, Victor Naylor. *Bournemouth, Hampshire.*
BASHFORD, Dennis John. *Erith, Kent.*
BATES, Brian. *Surbiton, Surrey.*
BEABEY, Robert John. *Morden, Surrey.*
BENNETT, John Robert. *Twyford, Berkshire.*
BESTER, Dennis Roy. *Coventry.*
BLAND, Randall Kenneth. *Hayes, Middlesex.*
BLOWERS, Derek Horatio Francis. *Colchester, Essex.*
BLUMIRE, Michael Alan. *Slough, Buckinghamshire.*
BOOTH, Frank. *Bolton, Lancashire.*
BORG, Lewis. *Gillingham, Kent.*
BOWER, John Edwin. *Chelmsford, Essex.*
BROOKS, John Richard. *Sutton Courtney, Berkshire.*
BROWN, Edwin George. *Bath, Somerset.*
CHIANG, Philip. *Southsea, Hampshire.*
COLE, Harry. *London, W.14.*
CONROY, Christopher. *Stevenage, Hertfordshire.*
COOPER, Reginald Thomas. *Blackburn, Lancashire.*
CRIMES, John Albert. *Wirral, Cheshire.*

Direct Election to Graduate (*continued*)

DOWSON, Peter. *London, S.W.16.*
ECCLESTON, Stanley. *Waterlooville, Hampshire.*
FIELDING, Warwick James. *Gosport, Hampshire.*
GIBSON, Peter John. *Crawley, Sussex.*
GRIFFITHS, Kenneth Edward. *Liverpool.*
GRUNDY, Brian. *Ashton-in-Makerfield, Lancashire.*
HAMER, Peter. *West Ewell, Surrey.*
HANDY, William Alfred. *Seaham, Co. Durham.*
HARTOG, Waldo Bastille. *Ilford, Essex.*
HARVEY, Arthur Soutar. *West Drayton, Middlesex.*
HAYES, Derek William. *Crawley, Sussex.*
HILL, Edwin George. *Reading, Berkshire.*
HOGAN, Daniel Joseph. *Liverpool.*
HUTCHINGS, Maurice John. *London, N.W.2.*
JONES, Alum Meredydd. *Rugby, Warwickshire.*
JONES, Kenneth Charles. *Woodford Green Essex.*
KEMPSON, David John. *Hayes End, Middlesex.*
KING, Michael Howard. *Southend-on-Sea, Essex.*
LARDER, Dennis Alexander. *Ilford, Essex.*
LAZENBY, Gordon Philip. *Leeds.*
LING, Frank Austin. *London, S.E.9.*
LLOYD, Gareth Treharne. *Kidnelly, Carmarthenshire.*
MARSHALL, Roy William. *Cheltenham, Gloucestershire.*
MAUNDER, Kenneth Henry. *Chelmsford, Essex.*
MEHTA, Vishwa. *Eastleigh, Hampshire.*
MUTCH, Norman Russell. *Liverpool.*
NORRIS, Bryan Leslie. *Nottingham.*
ODUAH, Michael Ikemefuna. *Birmingham.*
O'NEILL, Patrick John, B.Sc. *Winchester, Hampshire.*
OWEN, Elfed Griffith. *Chelmsford, Essex.*
PEACH, Peter John. *London, S.E.12.*
PYM, Roy Samuel. *Portsmouth, Hampshire.*
ROBERTS, Malcolm Edward Charles. *Crawley, Sussex.*
SEAR, Barry Lionel. *London, W.3.*
SIMMONS, Derek Brian. *Enfield, Middlesex.*
STILL, Peter Richard Morris. *Farnborough, Hampshire.*
STOCKDALE, Paul Keith. *Hitchin, Hertfordshire.*
SULLEY, Robert Alfred. *Copthorne, Sussex.*
SWEETMAN, Michael Victor. *Birkenhead, Cheshire.*
THOMAS, John Arthur. *Potters Bar, Middlesex.*
THURSTON, Alan Frederick. *Feltham, Middlesex.*
TOVEY, Nicholas Francis. *Chelmsford, Essex.*
TUFFIN, Alan Graham. *Southend-on-Sea, Essex.*
TURNER, Charles. *Liverpool.*
TUSTIN, Royston. *Cheltenham, Gloucestershire.*
UPTON, William James. *Birkenhead, Cheshire.*
WELLS, Alan Thomas. *Southall, Middlesex.*
WILLIAMS, Edwin Albert. *London, N.13.*
WITSEY, Kenneth. *Didcot, Berkshire.*
WOOLSEY, Jeremy John. *Tadworth, Surrey.*

Transfer from Associate to Graduate

PARRIS, Charles Deighton. *Kingston, Jamaica.*

Transfer from Student to Graduate

AHMED, Adnan. *Harlow, Essex.*
BRAVINSKY, Gary. *London, N.W.3.*
BRUTON, David. *Barnet, Hertfordshire.*
CLARKE, Eric. *Guildford, Surrey.*
CORNISH, Frederick Douglas. *Plymouth, Devon.*
COULSON, James. *Newcastle-on-Tyne.*
CROMARTY, David Denis. *Bracknell, Berkshire.*
CUTLER, William Alan. *Didcot, Berkshire.*
da SILVA, Eduardo Filipe. *Coventry.*
DAVIS, Michael John. *Edgware, Middlesex.*
DAWSON, John Sydney. *Wallsend, Northumberland.*
FALL, James Michael. *Ilkley, Yorkshire.*
GIBSON, Alfred. *Belfast.*
HARIHARAN, Varadarajan, B.Sc. *Tripunithura, India.*
HAWKINS, John Patrick. *Reading, Berkshire.*
JOSEPH, Michael. *Shenfield, Essex.*
KUMAR, Ramesh. *Weybridge, Surrey.*
LOCKHART, Reginald Frank. *Bath, Somerset.*
McQUIRE, Adrian Frank. *Romford, Essex.*
MABIN, Brian. *Bristol.*
MORTON, Brian William John. *London, S.W.11.*
NOBLE, David Cecil. *Totton, Hampshire.*
OGBU, Christian Okonkwo. *London, N.5.*
PARRIS, Charles Deighton. *Kingston, Jamaica.*
PENNY, Raymond William, B.Sc. *Carshalton, Surrey.*
RICHARDS, John Charles Thorpe. *Plymouth, Devon.*
SKINNER, John. *Bristol.*
TAT LIM ONG. *Kuala Lumpur, Malaya.*
TAYLOR, John Livesey. *Bolton, Lancashire.*
TILL, Peter Gordon. *Cowes, Isle of Wight.*
WASHINGTON, Derek. *Redhill, Surrey.*
WIFFILL, Edward John James. *Bristol.*
WOOD, Peter Ian McAskell. *Southampton, Hampshire.*

* Reinstatement

I.E.C. — Interlaken, 1961

This year the 26th General Meeting of the International Electrotechnical Commission took place in Interlaken, Switzerland, from 18th to 30th June; it was the largest meeting so far held by the I.E.C. and was attended by a total number of 950 delegates from 28 countries. The size of delegations ranged from 135 delegates from the United Kingdom, 112 from Germany and 111 from France down to 6 (Australia), 3 (Canada and Roumania), 2 (South Africa and Israel). Russia was represented by 22 delegates and China 13, while Japan sent 37 delegates.

Radio subjects were dealt with by various Technical Committees and sub-committees as follows:

Radio Communication	TC12
Electronic Valves and Tubes	TC39
Capacitors and Resistors	TC40
Cables, Wires and Waveguides for Telecommunications	TC46
(R.F. Cables TC46A; Waveguides TC46B; Cables TC46C.)	
Semi-conductor Devices	TC47
Ferromagnetic Materials	TC51

The writer attended, as a member of the United Kingdom delegation, the meetings of TC39 which took place during the first week of the sessions. The work of this committee deals with Electronic Valves and Tubes under two broad headings, namely

(a) Mechanical, and (b) Electrical Measurements.

In the United Kingdom these are dealt with at B.S.I. by the Technical Committees TLE5 and TLE5/2 respectively, on which the writer has served as Brit.I.R.E. representative since 1953.

U.S.–U.K. Liaison on Valves and Semi-conductors

For several years at I.E.C. General Meetings, before the formal meetings of TC39 start, a liaison meeting between the United States and United Kingdom delegations has been held. At these liaison meetings, which are informal, any differences of opinion or view can be ventilated.

This year there were a number of items which were freely and usefully discussed on valve and tube matters, while a similar liaison meeting took place on semi-conductor devices. These informal meetings were followed in the evening by a reception given by the U.S. delegation for the U.K. delegates to both TC39 and TC47. At this reception the writer had the good fortune to have a discussion with Mr. Acheson, the leader of the U.S. delegation to TC39, and was able to get some impression of the American attitude to the much wider use of I.E.C. material. It is the desire of the Americans to include as much as possible of the fundamental and theoretical data for the benefit of engineers and students of the present and future.

In this connection it may not be generally realized that, at the present time, the U.S. is preparing a number of papers for submission to and eventual issue by I.E.C. which deal with the fundamentals of some of the more important matters connected with the measurement of valve characteristics, such as, for instance, valve noise, cathode interface resistance, gas tubes, etc. Most of the data used for these fundamental documents are being taken from existing data prepared and used as standards within the American Institute of Radio Engineers and have only to be changed in form of presentation to make the documents suitable for I.E.C. use.

Such papers as have already been seen and read by the industrial organizations in this country have been well received and the data provided have been regarded as fairly complete. These papers would undoubtedly prove directly useful to engineers and students and should be given the widest publicity as sources of fundamental information.†

TC39 Meetings

The whole of Committee TC39 has to deal with all the material, and the chairman therefore proposed that the mechanical matters should be dealt with first to allow the maximum time for the electrical measurements. This was agreed and all the outstanding mechanical documents were dealt with on the first day.

On the second day before proceeding to consider the individual electrical measurement documents, a considerable amount of time was spent on the item, "Philosophy on Measuring Methods". This item had been raised indirectly by an I.E.C. Secretariat document (No. 102). The essence of this document was that if close agreement is to be obtained in measurements made with equipments conforming to I.E.C. recommendations, considerable detail may have to be given about circuit arrangements and measuring conditions. The Secretariat gave three possible approaches to this matter:

(1) Recommend one well-defined measuring condition from which good correlation may be expected in results from different units.
(2) Define the measuring conditions and give information on one circuit only. This would be expected to give a lower degree of correlation than (1).
(3) Define the measuring conditions and give alternative circuit arrangements which may give poorer correlation than (2).

The U.S. view was that measurement requirements differed for different stages, particularly regarding

† Distribution for I.E.C. is undertaken in U.K. by B.S.I.

accuracy. A very high degree of accuracy is required for laboratory work, rather less accuracy is sufficient on production line testing and considerably lower standards, such as are found in simple testers, are good enough for many users.

The German delegation thought that only those aspects likely to be a source of conflict between valve maker and user should be covered. Both Germany and France recommended only one circuit to be used as a reference. The U.K. and U.S.A. pointed out that the high-precision reference type of equipment was seldom employed by users and that its accuracy depended upon the definition of many circuit and even constructional details.

Considerable discussion ensued, particularly on the idea of having only one recommended method (favoured by France and Germany and the Netherlands) compared with having alternative methods (favoured by the U.K. and U.S.A.).

The U.S.A. suggested that four factors were involved in defining measurement methods:

(1) Definitions and terms
(2) Theory of the objectives of the measurement
(3) Basic circuits
(4) Measuring conditions

and proposed that this should be the pattern for all measurement documents, with allowance for deviations when necessary or desirable. This proposal was accepted, but it was recognized that the amount of theory required would depend on the complexity of the measurement and the amount of common knowledge involved.

The U.S. referred to alternative circuits in terms of their degree of "sophistication", a word that worried some of the European delegates and was later replaced by "refinement". The main factors in circuit refinement were accuracy, complexity, cost and correlation, but complexity was probably the more important item. The U.K. pointed out that the presence of all these factors made it difficult to decide which was the "best" method, hence the U.K. favoured alternatives. As a result of the decision to include theory in the documents, many of those before the meeting will have to be revised to cover this before they are progressed to I.E.C. Secretariat.

At last year's meeting in London it was proposed that documents covering fundamentals of complex characteristics might eventually be issued under I.E.C. and this was formally agreed to at this meeting.

With the ever-increased interest and use of microwave types of valves and tubes, magnetrons, klystrons, gas discharge devices, travelling-wave tubes, backward wave oscillators, etc., it is natural that methods of electrical measurement of the various parameters of these devices should be actively considered. Preliminary documents have been tabled covering magnetrons, gas discharge devices and klystrons. From these it was realized that there are many common features throughout the microwave measurement field.

At the meeting, it was felt that a specialist working group should be formed to handle this matter. The U.K. offered and it was agreed, to prepare a document covering the common features of microwave measurements as early as possible in 1962. All other National Committees were requested to supply any material for this within the next six weeks.

The meeting's decision to rationalize microwave valve measurements, by first producing a general section covering definitions and common factors applicable to this class of device, affords an opportunity for the Institution to offer its own recommendations regarding the contents of this document.

International Acceptance of Standards

There is evidence that the I.E.C. documents (and B.S.I. standards, where the former do not exist) are being used to an increasing extent by the new and developing countries for their own national organizations. This use will fairly certainly grow since year by year the membership of I.E.C. increases irrespective of political outlook. Also, I.E.C. data are being increasingly used in the formulation of U.S. and U.K. military specifications.

In this country the Services and Trade Associations are giving active support by the preparation of documents for submission to I.E.C. through the B.S.I. Technical Committees, certainly on matters relating to valves and semi-conductor devices.

It is clear from the foregoing that there are several aspects of this work in which the Institution can play its part. Two of these are:

(1) Publication at an early date of the material contained in the various individual documents.
(2) Prepare its own recommendations for individual documents on radio, radar, electronic and allied matters for presentation to B.S.I. and I.E.C.†

The former (1) may be a considerable assistance to students and engineers in making sure that the methods and background theory of the measurement of the various parameters are well known.

In the other case (2) the preparation by the Institution of recommendations on methods of measurement of particular parameters could influence the finally-approved procedure.

The next meeting of the I.E.C. is expected to take place in Paris in September 1962. G. R. JESSOP.

† This has already taken place in the case of the Technical Committee's "Recommended Methods of Expressing Electronic Measuring Instrument Characteristics: 1. A.M. or F.M. Signal Generators," *J. Brit.I.R.E.*, **18**, p. 7, January 1958.

U.D.C. 621.394.4

The Application of a Hamming Error Correcting Code to a Standard Teletype Equipment

By

R. W. LEVELL *(Graduate)*†

Summary: The paper describes apparatus which enables a standard single-channel teletype equipment to be converted for use with a Hamming error correcting code. The added apparatus is of small size and involves little interference with existing equipment. The electronic portion of this apparatus has been fully transistorized and a unit construction system, using printed circuit boards, has been employed.

1. Introduction

Communication engineers have always striven to improve the accuracy with which information may be transmitted over communication systems and one result of this effort has been the widespread adoption of telegraph working, using the Baudot code for transmitting literals and numerals.

With the invention of error-detecting and error-correcting codes considerable improvement in accuracy has been made possible over channels subject to noise and other interference, although this can only be obtained at the expense of signalling speed or by increasing the bandwidth of the channel.

This paper describes apparatus which has been added to a standard telegraph equipment to allow for the automatic detection and correction of errors in the signals received over a teleprinter link. For design purposes a Hamming code was chosen and a maximum of one digit error assumed to occur within each block of two consecutive teleprinter characters.

The signal is derived from a standard paper tape reader, and for convenience in applying the code the characters are read off in blocks of two; for each block the start and stop signals normally associated with each character are removed and the remaining digits transmitted to line together with a synchronizing digit and four parity check digits. Thus a group of fifteen digits is transmitted for each pair of teletype characters. The parity digits are generated by the coder according to a scheme devised by Hamming,[1] and they are added to each block of information digits in such a way that, when the decoder is fully synchronized, it is able to indicate the existence and location of an error and can therefore correct it. After correction, each group of the signal has the parity and synchronizing digits removed and the start and stop signals re-inserted; the output of the decoder is then suitable for applying to a serial teleprinter.

The additional transmitter equipment takes the form of a coding drum which is mechanically coupled to the tape reader (Fig. 1), and electronic circuits which operate from voltage pulses derived from this drum. Receiving equipment is required to establish digit and frame synchronism, to detect and correct digit errors and to convert the signals to a form suitable for driving a teleprinter.

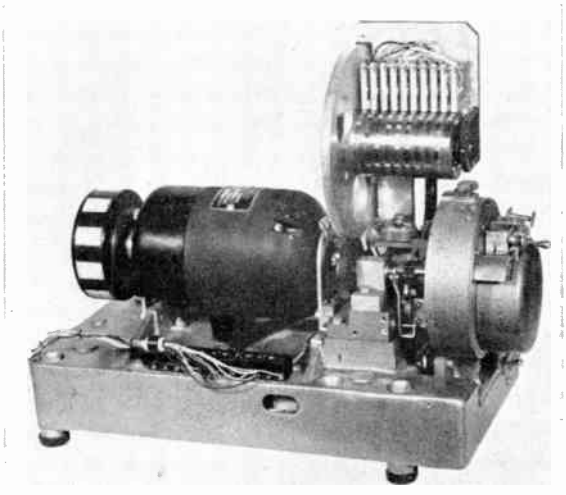


Fig. 1. The coding drum attachment shown mounted on the standard tape reader.

The particular Hamming code used in the design was chosen as a convenient compromise between the complexity of the electronic equipment involved, the degree of error cover provided, and the ease of adaptation of standard telegraph equipment.

2. The Detection and Correction of Errors using a Hamming Code

Assume that there are four digits (Pc1 to Pc4) available for correction purposes, then since each digit can have either of two values, namely a 0 or a 1 then there are $2^4 = 16$ possible arrangements of these digits. All of these arrangements are tabulated below, the first arrangement being reserved to indicate error-free transmission, the remaining 15 of them being made to correspond to different error patterns.

† Ministry of Aviation, Signals Research and Development Establishment, Christchurch, Hampshire.

Pc1	.	x	.	x	.	x	.	x	.	x	.	x	.	x	.	x
Pc2	.	.	x	x	.	.	x	x	.	.	x	x	.	.	x	x
Pc3	.	.	.	.	x	x	x	x	.	.	.	.	x	x	x	x
Pc4	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	x
		Pc1	Pc2	1	Pc3	2	3	4	Pc4	5	6	7	8	9	10	11

Those four arrangements containing only a single mark x are of course the check digits themselves; the remaining arrangements are labelled 1 to 11. By summing the digits marked x, with the exception of the term labelled 11 (since there are only ten information carrying digits in two teleprinter characters) from each line of the above table according to the rules of modulus-two addition (i.e. $0 \oplus 0 = 0$, $0 \oplus 1$ and $1 \oplus 0 = 1$, and $1 \oplus 1 = 0$), the following expressions result:

$$\text{Pc1} \oplus 1 \oplus 2 \oplus 4 \oplus 5 \oplus 7 \oplus 9$$
$$\text{Pc2} \oplus 1 \oplus 3 \oplus 4 \oplus 6 \oplus 7 \oplus 10$$
$$\text{Pc3} \oplus 2 \oplus 3 \oplus 4 \oplus 8 \oplus 9 \oplus 10$$
$$\text{Pc4} \oplus 5 \oplus 6 \oplus 7 \oplus 8 \oplus 9 \oplus 10$$

where $\oplus$ indicates addition modulus two.

Each of these expressions contains seven terms and the result, since each of the terms can have either the value 0 or 1, can only be 0 or 1.

The polarities of the check digits are chosen so that the four sums always give an even (= 0) result, this will be obtained if the check or parity digits are made equal to the sum of the other six digits, since both $0 \oplus 0$ and $1 \oplus 1 = 0$. Thus

$$\text{Pc1} = 1 \oplus 2 \oplus 4 \oplus 5 \oplus 7 \oplus 9$$
$$\text{Pc2} = 1 \oplus 3 \oplus 4 \oplus 6 \oplus 7 \oplus 10$$
$$\text{Pc3} = 2 \oplus 3 \oplus 4 \oplus 8 \oplus 9 \oplus 10$$
$$\text{Pc4} = 5 \oplus 6 \oplus 7 \oplus 8 \oplus 9 \oplus 10$$

Assume that a group of ten information digits are as shown in the second line of the table below; then the parity digits, generated by the coder in accordance with the above parity rules will be as shown in column Pc.

The decoder adds the terms of the four expressions previously given, to check for parity (= 0).

Thus

$$\text{Pd1} = \text{Pc1} \oplus 1 \oplus 2 \oplus 4 \oplus 5 \oplus 7 \oplus 9$$
$$\text{Pd2} = \text{Pc2} \oplus 1 \oplus 3 \oplus 4 \oplus 6 \oplus 7 \oplus 10$$
$$\text{Pd3} = \text{Pc3} \oplus 2 \oplus 3 \oplus 4 \oplus 8 \oplus 9 \oplus 10$$
$$\text{Pd4} = \text{Pc4} \oplus 5 \oplus 6 \oplus 7 \oplus 8 \oplus 9 \oplus 10$$

Normally, if there is no error in the received group of digits the result of carrying out these checks at the decoder will be 0 as shown in column Pd; these summations are always zero if the digits are correctly received.

If a digit becomes reversed in sign during transmission, then the decoder check equations will no longer all yield 0. For example, suppose digit no. 5 becomes a 1 due to error in transmission then the decoder check sums become

$$\text{Pd1} = 1 \oplus 0 = 1$$
$$\text{Pd2} = 0 \oplus 0 = 0 \quad \text{remains unchanged}$$
$$\text{Pd3} = 1 \oplus 1 = 0 \quad \text{remains unchanged}$$
$$\text{Pd4} = 0 \oplus 1 = 1$$

The final column is obtained by summing the coder parity digits and the selected character digits, including the effects of the inserted error, since this is the action of the decoder.

Now there is only one digit which can affect equations 1 and 4 alone and that is digit no. 5. Thus the decoder can decide which digit is in error and can therefore correct it. To carry out the correcting operation it is first necessary to store the incoming character digits in a shift register and then to change the state of the stage which holds the incorrect digit.

Digit No.	1	2	3	4	5	6	7	8	9	10	Coder Check	Decoder Check
Digits	1	0	0	1	0	1	1	0	0	0	Sum Pc	Sum Pd
Pc1 = Σ	1	0		1	0		1		0		1	0
Pc2 = Σ	1		0	1		1	1			0	0	0
Pc3 = Σ		0	0	1				0	0	0	1	0
Pc4 = Σ					0	1	1	0	0	0	0	0

Four check digits, ten character digits and one further digit for synchronizing purposes total fifteen, which is also the number of digits in three teleprinter characters (discounting the start and stop digits normally associated with each character). Each of the digits of two characters may thus be covered by the free arrangements of the four parity check digits. Two characters and their appropriate checks may therefore be transmitted, with a synchronizing digit, in the same time as that normally required for transmitting three characters.

These facts greatly simplify the engineering of an experimental error correction system for use with telegraph equipment. Two characters may be read from a paper tape, the tape may then be stopped for the period of one character, before reading a further pair, and the interval may be used for inserting check and synchronizing digits.

The code described is designed to cope only with single errors within the group and the presence of more than one error will cause the correction system to fail. In some cases this means the insertion of a further error, e.g. suppose that both digits 5 and 8 were received in error, then the decoder would decide that check equations 1 and 3 were unsatisfied and that therefore digit 2 was in error. The decoder would attempt to "correct" this digit and would thereby insert a further error.

The code so far described is a particular code using four check digits to cover ten information digits against single errors occurring within the group. Other codes can be designed to deal with n binary digits representing information using m check digits in each group.

The number of ways in which a single error can occur is ${}^{m+n}C_1$; the number of ways in which double errors can occur is ${}^{m+n}C_2$ and in general the number of ways in which r errors can occur is ${}^{m+n}C_r$.

There are 2^m possible arrangements of m check digits, one of which will correspond to error free transmission; in principle all other arrangements can be made to correspond to (2^m-1) different error patterns. Thus if it is required to correct all single errors, it is necessary to ensure that the number of such patterns is not greater than the available number of check digit arrangements, so that

$$ {}^{m+n}C_1 \leqslant 2^m - 1 $$

If it is required to correct all single and double errors then

$$ {}^{m+n}C_1 + {}^{m+n}C_2 \leqslant 2^m - 1 $$

and in general if it is required to correct all single errors, double errors and so on up to r errors in each block then

$$ \sum_{s=0}^{r} {}^{m+n}C_s \leqslant 2^m $$

This equation gives the basic requirements of a realizable system.

If it is assumed that all digits are equally likely to be in error (with probability p) and that there is no correlation between errors, one may readily calculate the probability Q_r that a block of information and its check digits can be recovered free from error.

$$ Q_r = \sum_{s=0}^{r} {}^{m+n}C_s \, p^s (1-p)^{n+m-s} $$

The probability of some error, $P_r = 1 - Q_r$, can thus be found for a variety of conditions.

It is desirable that the redundancy be kept as low as possible so that the maximum information rate can be achieved for a given bandwidth. The amount of redundancy will be determined from a consideration of the prevailing error rate. In general, errors tend not to be uniformly distributed but to occur in clusters and a practical system must be designed to cope with the error rate prevailing during these clusters.

Other factors influence the design of an error-correcting system which is supplementary to teletype equipment, for example implementation must be possible with little interference to existing apparatus. The system design here described was chosen as a convenient compromise between the error cover provided and the mechanical and electronic complexity.

3. Description of the Basic System

Briefly the task of the transmitter and coder is to:

(1) Read two characters sequentially from the paper tape.

(2) Remove from these characters the associated start and stop signals.

(3) Send the remaining ten digits (five forming each character) to line at a constant band rate.

(4) Generate the four appropriate check digits in accordance with the coder parity rules and to send them to line.

(5) Generate a fifteenth digit which alternates in polarity from frame to frame (this digit is used to achieve frame synchronization at the decoder) and to send to line.

Correspondingly the task of the decoder is to:

(1) Achieve digit and frame synchronization.

(2) Carry out the decoder parity check and if an error is discovered to correct the digit found to be in error.

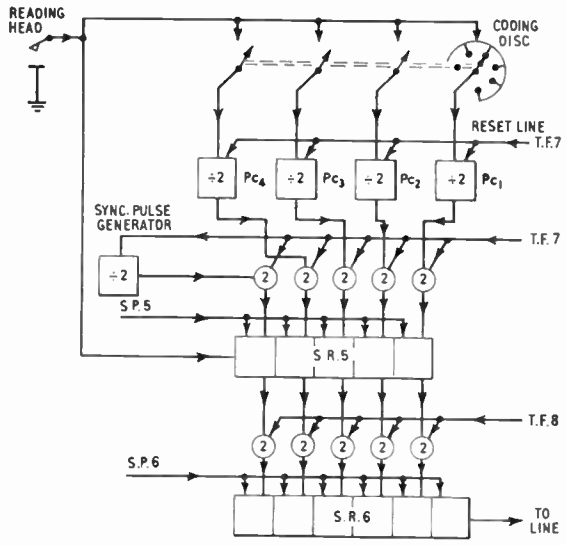


Fig. 2. Block diagram of coder.

(3) Remove the check digits and re-insert the start and stop signals necessary to operate a serial teleprinter.

(4) Send to the printer the characters, corrected if necessary, at a rate appropriate to the printer.

3.1. *Operation of the Coder*

A block diagram of the coder is shown in Fig. 2 and the action is as follows.

The tape reader is driven from a constant-speed motor the drive of which can be mechanically connected to the tape reading head by energizing a relay. A coding drum is geared to the motor shaft and revolves at one-third of the driving speed to the reading head.

This drum has ten circumferential grooves cut in its surface, dividing it into ten conducting rings, nine of which have notches cut parallel to the axis of the drum. The grooves and notches are filled with a non-conducting resin, and ten carbon brushes are applied to the rings, one of which acts as an earth return. The drum, brushes and gears are assembled as a unit which is mounted on the tape reader in a convenient position. Figure 3 shows a disc with a number of rings having marks in different angular positions, the notches on the circumference of the coding drum corresponding to these marks.

If the brushes were to be connected, through resistances, to a suitable voltage supply then, as the drum revolves, voltage pulses would be obtained at the brushes. The short black lines on the right-hand side of Fig. 3 show the ideal positions, in time, at which these pulses occur.

Ring 9 was intended to operate the relay on the reading head twice during each revolution of the drum, so that two characters could be read from the tape for each revolution. It has been found necessary in practice, because of the slow speed of operation of this relay, to lengthen the insulated portion of the ring to approximately half of the circumference and to re-position it relative to the other rings.

Rings 1 to 8 produce the gating and driving pulses required for the operation of the coder circuits. It was found in practice that commutator noise was superimposed on these pulses and, as an alternative to redesigning the commutator and brush assembly, it was decided to "clean up" the pulses electronically. This meant increasing the amount of electronic equipment involved in the coder and further work needs to be directed towards obtaining a more satisfactory mechanical system. (For example, the use of a glass disc and phototransistors may prove more satisfactory.)

The circuits used at present for cleaning up the voltage pulses from the coding drum consist of simple integrating networks followed by transistor slicing circuits. The outputs from these slicers are fed to standard pulse generators which drive the shift register and counter stages forming the remainder of the coder. The circuit of the pulse generator

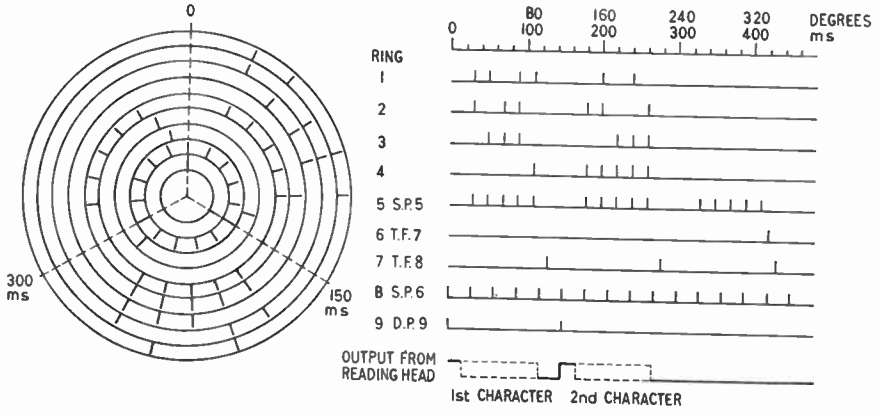
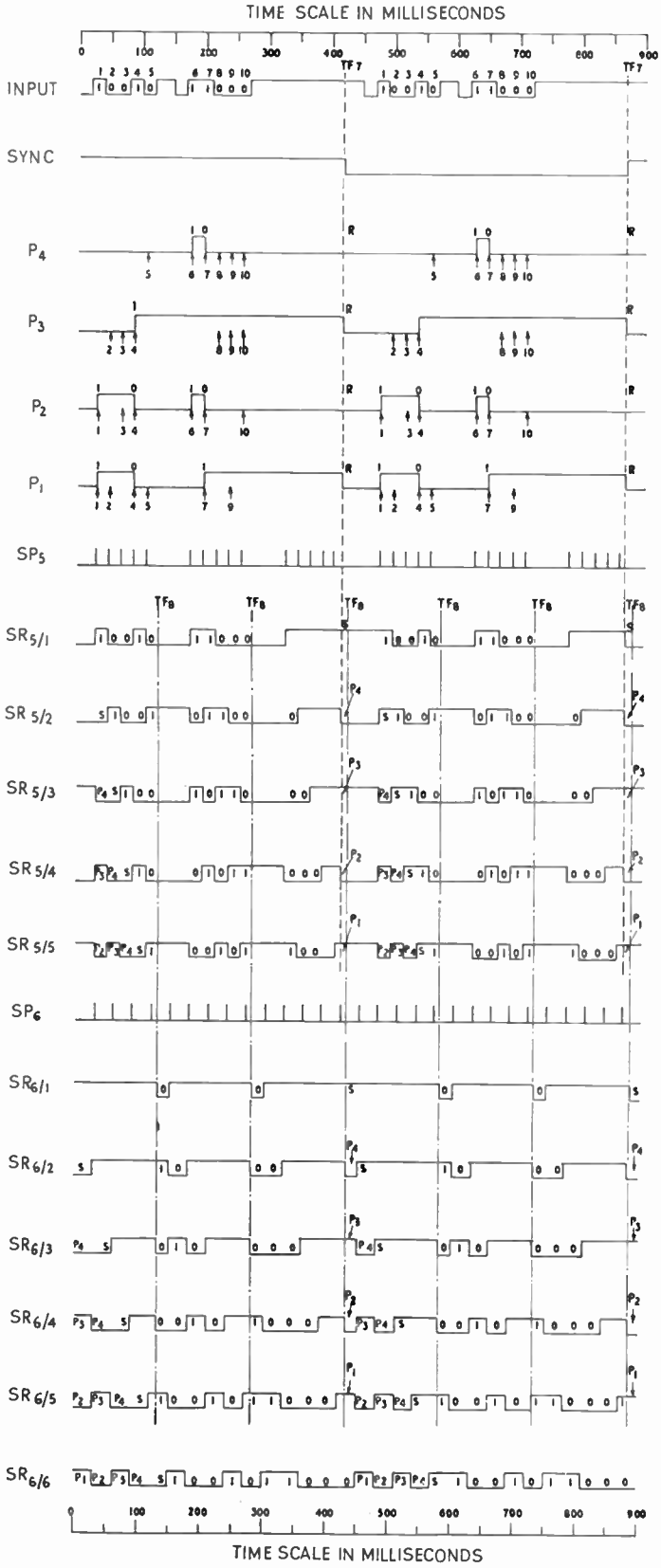


Fig. 3. The coding disc.



is based on a design by D. J. Hamilton[2] and it has the advantages that ease of triggering is not obtained at the expense of temperature stability and that a large output current is available during the period of the pulse. The shift register and counter stages are formed from standard units, each of which consists of a printed circuit board holding a transistor bistable circuit of the Eccles-Jordan type, two sets of diode gates and temporary stores. Contacts at the base of the board enable the unit to be wired externally to perform any one of a range of operations. (Further details are given in the Appendix.)

Shift register S.R.5 is driven by the S.P.5 pulses derived from ring 5 on the drum. These pulses are timed so that immediately after the fifth S.P.5 pulse the shift register contains the five digits of the first character (see Fig. 4). During the time that these digits are being "run" into the register they are also gated by rings 1, 2, 3 and 4 to the appropriate divide-by-two counters. These circuits therefore perform the necessary modulus two additions. In order that the character may be passed to line at a constant digit rate, the stages of shift register S.R.6 are set to the same states as the stages of S.R.5 by T.F.8, the "transfer" pulse, which is timed to occur immediately after the character has completely filled S.R.5. The digits are then passed to line at a constant rate by the pulse stream S.P.6.

This process is repeated during the time of the second group of S.P.5 pulses, the divide-by-two circuits now registering the appropriate check digits and the shift register S.R.5 being filled with the digits of the second character. After the second T.F.8 pulse, S.R.6 will also contain these digits, the first character having passed to line.

During the third group of S.P.5 pulses, S.R.5 will be emptied and the divide-by-two circuits remain unchanged since no character is read from the tape since the clutch mechanism connecting the reading head to the motor shaft is disengaged at this time.

T.F.7 now causes the check digits and the synchronizing digit to be set into S.R.5, resets the divide-by-two circuits and changes the state of the stage providing the synchronizing digit.

Fig. 4. Waveforms associated with the coder. (R denotes "reset")

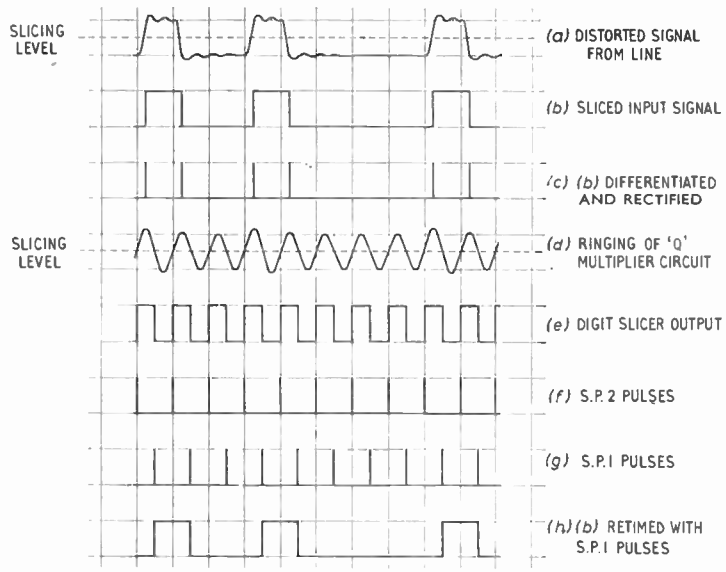


Fig. 5.
Digit recovery.

The third T.F.8 pulse occurs immediately after T.F.7 and sets S.R.6 to the same state as S.R.5. This completes the cycle, the contents of S.R.5 being run out during the next group of S.P.5 pulses and the contents of S.R.6 being passed to line as before.

Figure 4 illustrates the operations described above for two particular characters which are transmitted over a period of two frames. The progress of these characters may be traced in this diagram from the time they are read off the tape to the time they are transmitted to the line. The timing of the operative pulses, the generation of the parity check digits and the states of all the stages of both the shift registers are shown throughout the process.

3.2. *Operation of the Decoder*

3.2.1. Digit recovery

The incoming signal is normally somewhat distorted due to the attenuation and phase characteristics of the transmission path. Figure 5 shows typical waveforms of the digit recovery circuits. The circuits themselves are illustrated in Fig. 6. The signal is applied to a voltage-slicing circuit which operates at the half-amplitude level of the signal. The output of this slicer is differentiated and rectified to produce pulses corresponding to the changeovers. These pulses are applied to a "Q-multiplier" circuit which is resonant to the digit frequency of 33 c/s. The circuit rings at each impulse and continues ringing in the

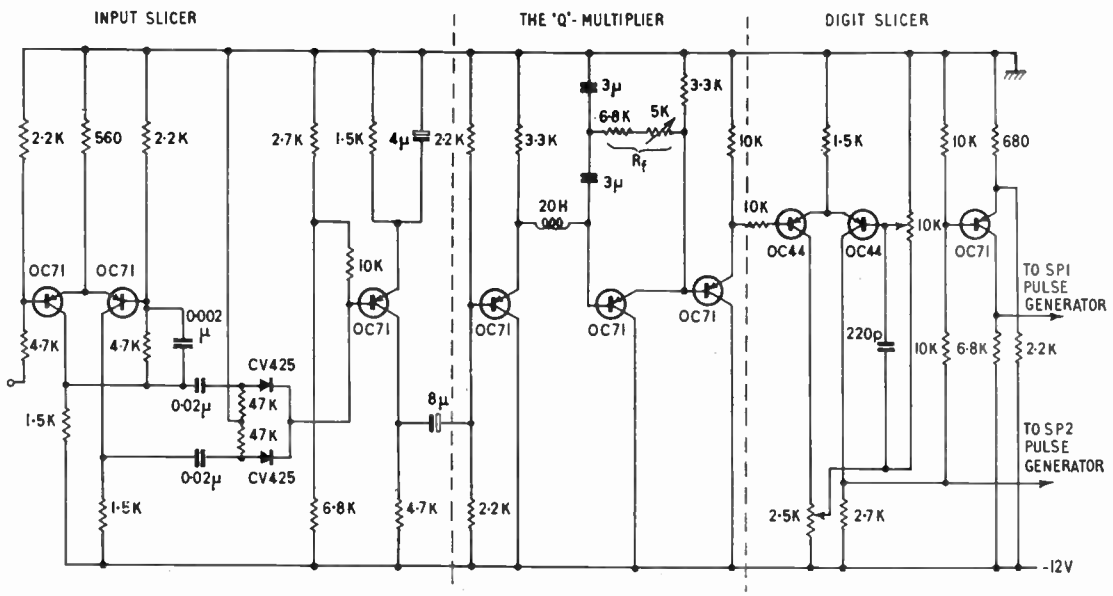


Fig. 6.
Digit recovery circuit.

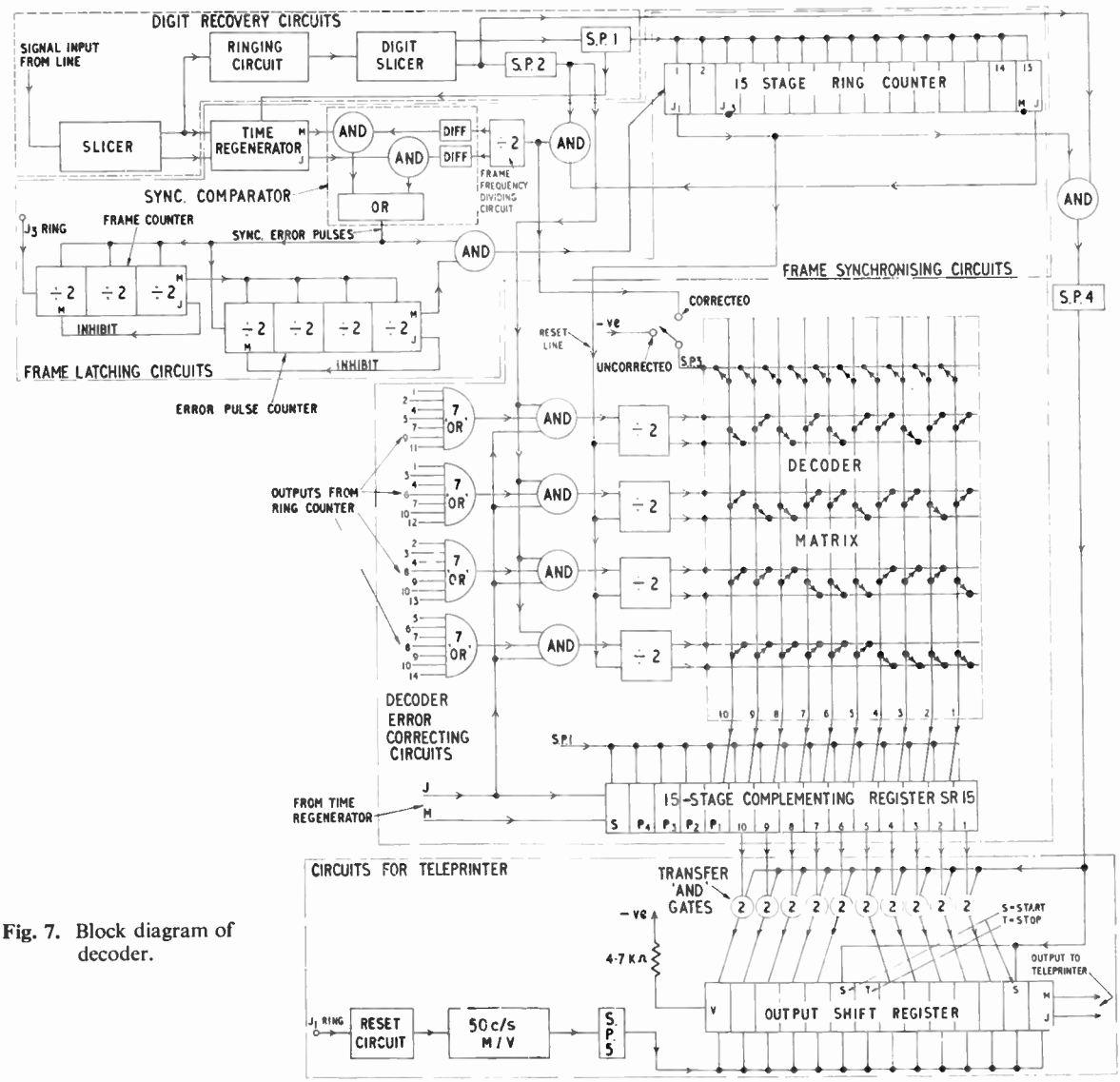


Fig. 7. Block diagram of decoder.

absence of further pulses for several frames. A second slicing circuit connected to this multiplier provides a two-phase square-wave output, and is adjusted by means of the two potentiometers to give a zero slicing condition—analogous to the zero-grid-base valve slicer. It has been found necessary to include a further inverting amplifier after this slicer since the waveform at the first collector was found to have an insufficiently short rise-time to trigger one of the standard pulse generators. The two phases of the square-wave, one from the second collector of the slicer and the other from the invertor, are used to drive standard pulse generators which thus produce the pulse streams S.P.1 and S.P.2.

The Q-multiplier is based on a circuit by G. B. Miller[3] the signal being injected in series with the tuned circuit from a low-impedance source. The d.c. conditions are arranged to allow the maximum output voltage swing, the feedback being taken from an emitter follower through a suitable resistor R_f to the junction of two series tuning capacitors at its base. Variation of this feedback resistor enables the effective circuit Q to be changed, any value less than critical causing the circuit to be self-oscillatory. With a coil Q factor of about 12 this circuit has been found capable of giving a stable multiplication factor of some 150; thus an effective Q of some 1800 is obtainable. At 33 c/s this enables the digit frequency to be maintained over 600 digit periods, i.e. 40 frames. Since there is an alternating synchronizing pulse in the transmitted digits then this circuit receives at least two impulses during two frames; an adequate "flywheel" can thus be obtained with comparative ease.

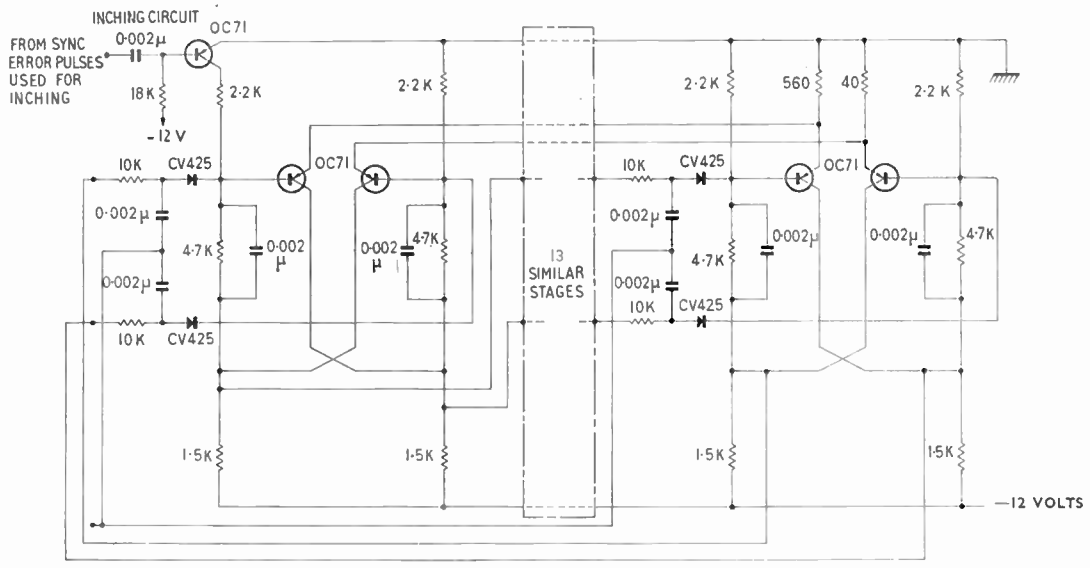


Fig. 8. The ring counter.

In addition to passing to the Q-multiplier the sliced input signal is also passed to a regenerating unit, and is re-timed by S.P.1, it is then used to establish frame synchronization and to feed the decoder circuits (Fig. 7).

3.2.2. Frame synchronizing and noise guard circuits

Frame synchronism is established by inspecting the incoming signal from line for the alternating synchronizing digit since that digit is the only digit in the frame which regularly alternates in polarity. The 15-stage ring counter is the principal device used in the search for this digit. It is driven by S.P.1 pulses and since there are 15 digits in the frame, then a single 1 held in the ring counter circulates at the frame rate. The AND gate, fed from the 15th stage of this counter and the S.P.2 pulse stream, produces a single S.P.2 pulse whenever the 1 is present in that stage of the ring counter. This pulse is fed to a divide-by-two circuit which thus changes state on alternate frames. This circuit is called the frame frequency dividing circuit for further reference.

The ring counter consists of a 15-stage shift register whose last stage output is connected to the input of the first stage so that any digits present in the register are constantly circulated at a rate determined by the applied shift pulses (Fig. 8). The method of ensuring that only a single 1 circulates[4] is to connect one of the

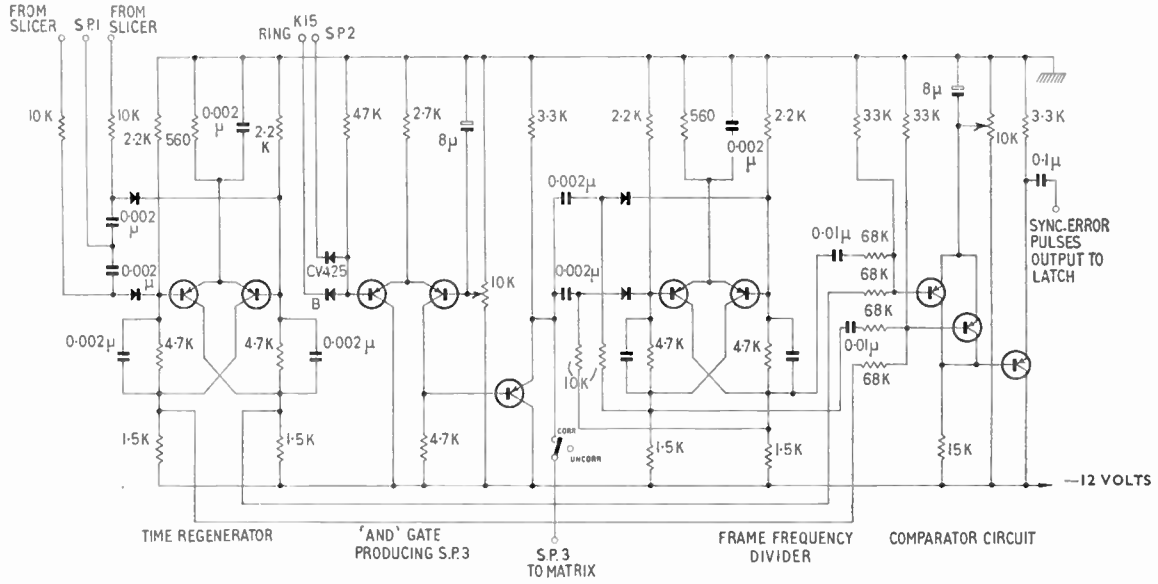


Fig. 9. Sync. error pulse generator.

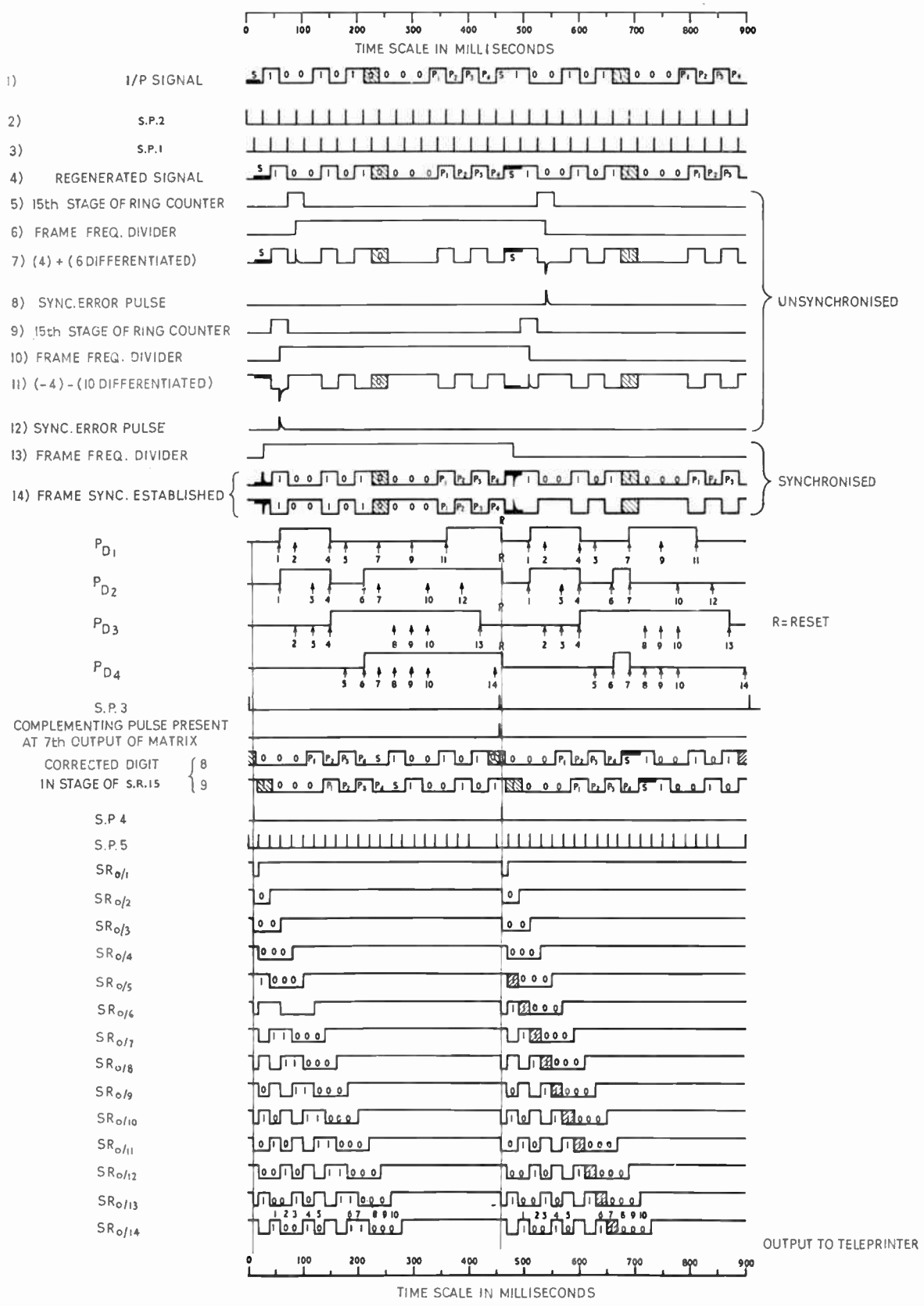


Fig. 10. Waveforms associated with the decoder.

emitters of each stage to a common line and to return this line to earth through a resistance R; the other emitters being connected to a second resistance of $R/(n-1)$ where n is the number of stages in the ring. For the standard circuit used R is 560 ohms so that $R/(n-1)$ is 40 ohms. The circuit is only stable if the two emitter voltages are at about the same value and this can only be so when one transistor connected to R and 14 of those connected to $R/(n-1)$ are conducting. If the single conducting state is called a 1 then this 1 will be circulated around the ring on the application of shifting pulses, the gating between successive stages determining which stage is the next to conduct. Ring counters of this type can be started from any desired stage by forcing that stage to the 1 state at the time required, the previously conducting stage immediately reverting to the 0 state. This is the process which is used to "inch" the ring counter with respect to the input signal when the receiver is endeavouring to establish synchronization.

Synchronizing error pulses are produced whenever the receiver is out of synchronism by a comparator which consists of the differentiating networks and the AND and OR gates shown (Figs. 7 and 9). These pulses are derived from the sum of the regenerated input signal and the differentiated output waveform of the frame frequency dividing circuit. Both phases of the input signal and of the dividing circuit are used to generate them, the reason being that error pulses come from either one AND gate or the other (see Fig. 10). After slicing, the output of these gates is in the form of positive pulses if the receiver is unsynchronized, no pulses being present when the receiver is correctly synchronized.

These synchronizing error pulses, one of which occurs in each frame, are used to force the first stage of the ring counter into the 1 state, thus causing the circulating 1 to be advanced by one digit. The next test for synchronization is made with the frame frequency divider (driven from the 15th stage of the ring counter), fourteen digit periods later. The test for synchronization is made again during the next frame and if the receiver is still out of synchronism then the 1 is advanced a further digit position. This process continues until frame synchronization is established.

To prevent incorrect synchronization and to maintain synchronization during disturbances, further circuits have been added to ensure that (*a*) there are at least four consecutive frames during which no synchronizing error pulses occur before the receiver decides that it is in synchronism, and (*b*) that there are eight consecutive frames during which synchronizing error pulses are produced before the receiver decides that it is out of synchronism.

To achieve these effects, the passage of the synchronizing error pulses to the ring counter is controlled by the synchronizing guard circuits which make the decision "in" or "out" of synchronism in accordance with the above criteria. The decision "in synchronism" prevents the flow of error pulses through the AND gate and thus leaves the ring counter undisturbed. The "out of synchronism" signal on the other hand, enables the pulses to pass through to the ring counter and so permits the receiver to search for the synchronizing digit and thus establish synchronism.

The operation of these guard circuits is best described by first assuming the receiver to be out of synchronism.

The frame frequency divider is compared with the incoming signal and error pulses indicating the unsynchronized condition are produced by the comparator. These error pulses are counted by the error pulse counter and are also used to reset the frame counter. After eight consecutive error pulses the last stage of the error pulse counter changes state and latches the counter, thus preventing further error pulses from being counted. This last stage also allows pulses to be applied through the AND gate to the ring counter. Thus the ninth error pulse "inches" the ring counter one digit. A further test is made for synchronism, and if still out of synchronism, a further error pulse is produced which in turn passes through the AND gate and again "inches" the ring counter. This process continues until error pulses are no longer produced by the comparator. The frame counter then begins to count the number of frames, it being no longer reset by the error pulses.

After the fourth frame pulse from the ring counter the last stage of the frame counter changes state, causes the counter to latch, and also resets the error pulse counter. This latter operation causes the AND gate to close and so prevents the flow of chance error pulses to the ring counter.

Suppose that a noise pulse produces a synchronizing error pulse; then this resets and unlatches the frame counter, and is counted by the error pulse counter, since this had previously been reset by the frame counter. If there is a gap of more than four frames between one error pulse and the next, the frame

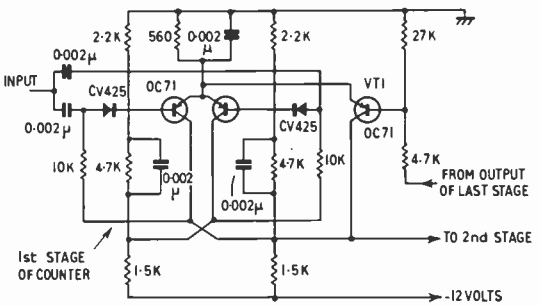


Fig. 11. Latching of counters.

counter will latch, the error pulse counter will be reset, and the ring counter will not be disturbed. A minimum of eight error pulses, with gaps between them of less than four frames, are thus required for the receiver to decide that it is out of synchronism.

The latching operation of these two counters is carried out by shorting the output of the first stage of the counter with a transistor VT1 (See Fig. 11). The output of the last stage of the counter is applied to this transistor so that when this stage changes state, due to the normal counting operation, then VT1 conducts and appears as a short circuit across the first stage thus preventing further input pulses from being counted. The counter will remain latched until VT1 is made non-conducting by resetting the counter to zero.

3.2.3. The decoding circuits and the drive to the teleprinter

After digit and frame synchronism have been established it is necessary to check the ten information bearing digits together with the four parity digits to determine if any of them have been received incorrectly. This is achieved by automatically carrying out the operations described in Section 2.

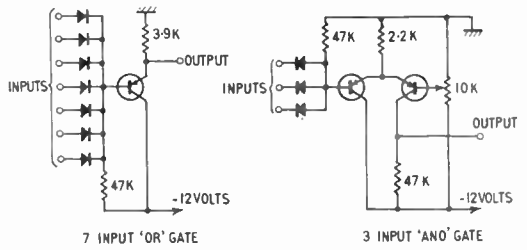


Fig. 12. Gating circuits.

Note: All gates take the same form unless specifically stated otherwise, using the appropriate number of diodes.

The regenerated signal is passed to each of four AND gates which are also fed with S.P.2 pulses and gating waveforms obtained from the 15-stage ring counter. Figure 12 shows the circuits of the seven input OR gates and also that of the three input AND gates; they consist of simple diode gates, followed by an emitter-follower and a slicer respectively. The slicer is necessary in the case of the AND gates since the input signals to these gates are of different voltage levels. Figure 10 shows the inputs and outputs of these gates for the same signal as that used to illustrate the action of the coder. Those S.P.2 pulses which pass through the gates are fed to the divide-by-two circuits which thus carry out the decoder check summations.

The outputs of these divide-by-two circuits are connected to a diode matrix which can present a pulse (S.P.3) on any one of its output wires. The matrix converts the two phase signals from the decoder divide-by-two circuits to the complementing pulses which are used to correct the information bearing digits held in the complementing register. Figure 10 shows waveforms illustrating the operations. The matrix consists essentially of ten diode AND gates, each having five inputs, four for the check digit inputs and one for S.P.3; emitter-followers are used on the outputs of the matrix for impedance conversion purposes. If an error has occurred during the transmission of a frame of digits then an S.P.3 pulse will be present on the output wire associated with the appropriate

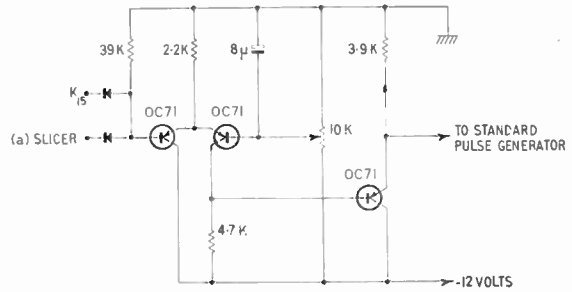


Fig. 13. Transfer pulse generation.

character digit. Note that S.P.3 is obtained by gating the S.P.2 pulse stream with the 15th stage of the ring counter in a two-input, diode AND gate, the output of the gate being sliced and fed through an emitter-follower to the matrix (Fig. 9). There is no need to indicate the presence of an error in either the parity digits or the synchronizing digit since it is assumed that no more than one error occurs within a frame. Thus an error in these digits implies that the character digits have been received correctly.

The regenerated input signal is passed to the fifteen-stage complementing shift-register S.R.15, the last ten stages of which are wired, in addition to the normal shift-register connection, as divide-by-two circuits by using the second pair of gates provided on each of the printed circuit boards composing this register. If an error is present an S.P.3 pulse will appear at gates of the appropriate shift-register stage and cause that stage to change state, i.e. S.P.3 acts as a complementing pulse drive to the stage concerned.

The register then contains the ten character digits; any single error having been corrected; these are then shifted one place further along the register to ensure that all digits are of the same duration and are then transferred in parallel by means of the transfer pulse S.P.4 to the output register as shown in the decoder block diagram. The transfer pulse is produced by gating one of the ring counter stages with the output of the digit slicer (Fig. 13) and feeding the resulting

waveform to one of the standard pulse generators. This pulse is thereby made to occur immediately after the complementing pulse S.P.3. Transference to the output register is necessary so that the start and stop digits, required for sequential operation of the teleprinter, may be reinserted, and also permits the digits to be sent to the printer at the required speed. If a parallel-operated printer had been available then the above operation would not have been necessary and

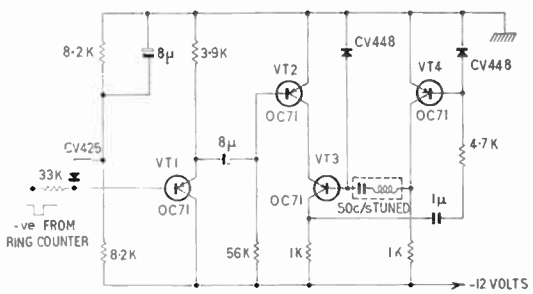


Fig. 14. The output multivibrator.

equipment would then have been saved. By suitable biasing, a 1 is always present at the input of the output register, the "stop" signals thus automatically appearing in the output. The "start" signals are then inserted by applying the transfer pulse to the 6th and 13th stages in the same way as in the complementing register described above. A further stage is used to make the first digit coming from the output the same duration as the others, since at the time when it is inserted by the transfer pulse it is shortened in duration.

The oscillation of the 50 c/s multivibrator,[5] which controls the rate at which digits are fed into the printer, is itself controlled by the tuned circuit connected between transistors VT3 and VT4. This multivibrator (Fig. 14) must be brought into synchronism each frame and this is ensured by using one of the ring counter pulses to define the starting point of the oscillations. The CV425 diode, in conjunction with the 33 kΩ resistor, has been included to improve the shape of this pulse and to prevent any unwanted voltage spikes from affecting the oscillator. Transistor VT2 is normally conducting, but is switched off during the period of the counter pulse to force the oscillations to commence at the same time after its occurrence. The pulse generator connected to its output produces suitable shifting pulses for feeding the digits into the printer. The magnet of the teleprinter is normally driven by a telephone relay (Type 3SE1) which is energized by two OC72 transistors operated in push-pull. The required supply and biasing voltages for these drive circuits are provided by a separate power source.

4. Performance of the Equipment

Duplication of transmission conditions with a controllable content of white noise was impracticable in the time available; some measurements were therefore made with errors inserted by mixing the coder output with noise derived from a diode noise source.

The presence of the complementing shift register in the decoder delays the digits by 15 digit periods; the synchronizing pulse in the coder output is therefore out of phase with the synchronizing pulse from the output of the decoder complementing register. A further fifteen stages of shift register are therefore required to match the outputs. This shift register has been inserted between the coder outputs and a 40 c/s low-pass filter (Fig. 15). The filter was included to introduce some phase and frequency distortion in a manner similar to that introduced by real transmission paths. The filtered signal was then mixed with 0–200 c/s noise and fed to the receiver input slicer. Errors were thus inserted depending upon the amplitude of the injected noise and the phasing of the regenerating pulses of the S.P.1 pulse stream.

Automatic correction of these deliberately inserted errors was avoided by preventing S.P.3 pulses from flowing to the decoder matrix, achieved by operating the correction switch in the appropriate line (see Fig. 7).

The uncorrected output of the complementing register S.R.15 was then fed to a modulus-two adder together with the output of the coder. The latter was suitably phase-adjusted by re-timing in the delay stage

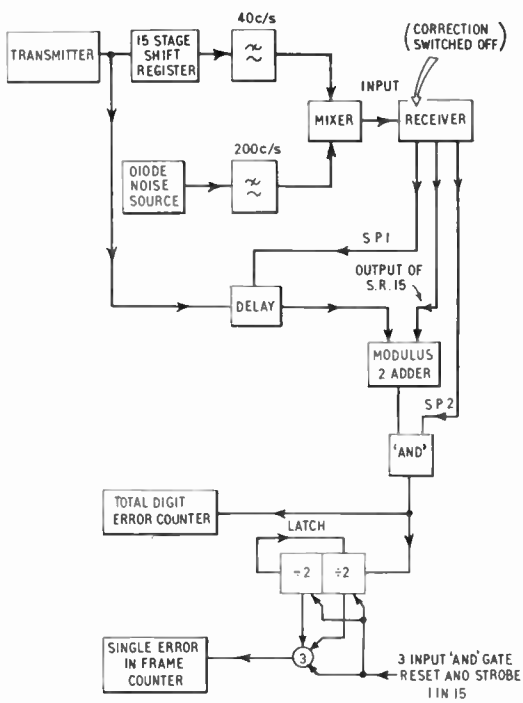


Fig. 15. Block diagram of error rate measuring apparatus.

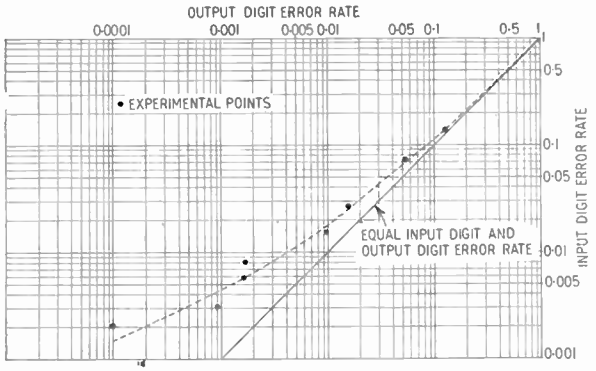


Fig. 16. Output digit error rate as a function of input digit error rate.

with S.P.1 pulses from the decoder; the modulus-two adder thus gave an output only when its inputs differed due to inserted errors. This output was converted to pulses by gating with the S.P.2 pulse stream and the number of these pulses counted by the total digit error counter. Single errors in the frame of 15 were counted by the second counter which only registered if the preceding latching counter counted a single pulse between resetting at each frame.

Readings of the total digit errors and the number of single errors in groups of 15 digits were noted from the counters over periods of 10 minutes and the output digit error rate was calculated and plotted against the input digit error rate.

$$\text{Input digit error rate} = \frac{\text{total digit errors}}{\text{total digits}}$$

$$\text{Output digit error rate} = \frac{\text{total digit errors–no. of single digit errors in group of 15}}{\text{total digits}}$$

The graph shown in Fig. 16 shows how the error correcting system gives a gain at low error rates. It does not, however, show the effect of the error correcting system at high error rates, since no measurements have been made of the extra errors inserted by the decoder under these conditions. This latter omission is not considered important since a coding system would not, in general, be used at error rates far in excess of those for which it was designed.

5. Conclusions

The method described for converting teletype equipment for use with an error-correcting code has the advantages that only minor modifications need be made, and that the presence of the additional apparatus does not prevent the use of the machine as a normal teletype whenever required. Further development work is needed to obtain a more satisfactory coding arrangement, since the present drum requires electronic circuits to improve the waveform of the coding signals.

The majority of the electronic equipment is built up from a number of identical printed circuit boards, each containing a transistor bistable circuit and diode gates. The board is made to be plugged into an 18-way socket and sockets can be wired to allow the circuit to perform a number of digital operations. The electronic equipment has therefore the feature that faults may be located and corrected more easily, since a suspect board may be replaced with a correctly operating one, the suspect being checked independently of the equipment with a special circuit testing device if necessary. Since the equipment is constructed from solid-state devices there are of course the attendant advantages of reliability and small size.

It must be emphasized that the whole of the apparatus described is of an experimental nature. A fully engineered version may thus differ considerably in matters of detail.

Since the preparation of this paper for publication, further work of importance in this field has been published.[6, 7] Recent work has tended to concentrate on codes designed for correcting bursts of errors.

A summary of some of the different types of error correcting codes can be found in reference 7, where they are approached from a new viewpoint involving polynomial equations.

6. Acknowledgments

This article is published by permission of The Controller, Her Majesty's Stationery Office.

The author is indebted for the encouragement and helpful criticism provided by Mr. P. H. Cutler, under whose direction the work described was carried out.

7. References

1. R. W. Hamming, "Error detecting and error correcting codes", *Bell System Tech. J.*, April 1950.
2. D. J. Hamilton, "A transistor pulse generator for digital systems", *Trans. Inst. Radio Engrs (Electronic Computers)*, **EC-7**, No. 3, pp. 24–9, September 1958.
3. G. B. Miller, "Transistor *Q* multiplier for audio frequencies", *Electronics*, **31**, p. 79, May 1958.
4. T. K. Sharpless, "High speed 'N' scale counters", *Electronics*, **21**, pp. 122–5, March 1948.
5. D. D. McLeod, "A multivibrator controlled sinusoidal oscillator", *Electronic Engineering*, **30**, pp. 724–5, December 1958.
6. W. W. Peterson and D. T. Brown, "Cyclic codes for error correction", *Proc. Inst. Radio Engrs*, **49**, pp. 228–35, January 1961.
7. S. H. Rieger, "Codes for the correction of clustered errors", *Trans. Inst. Radio Engrs (Information Theory)*, **IT-6**, pp. 16–21, March 1960.

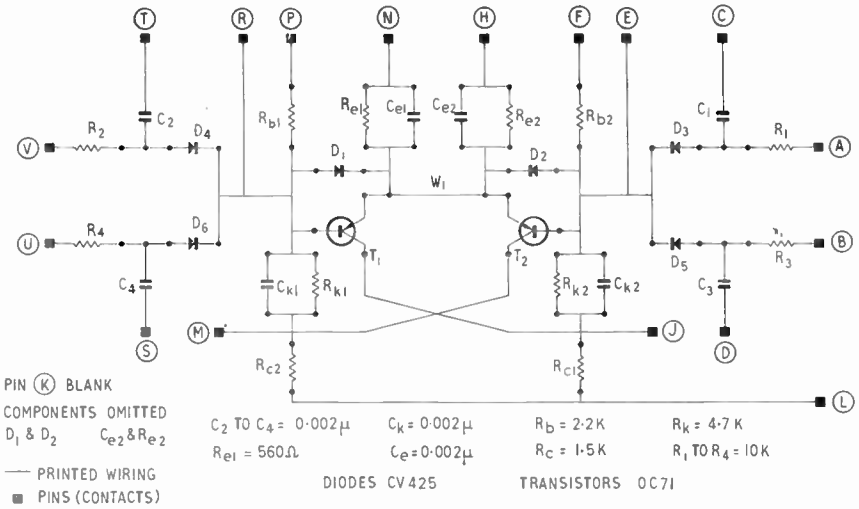


Fig. 18. The standard unit.

8. Appendix

The Printed Circuit Boards used in the Construction of the Coding and Decoding Equipment

The electronic portion of the equipment described in this paper contains a comparatively large number of basic circuit elements such as shift register stages and scale-of-two counters. To facilitate assembly of the equipment and to ease its development it was decided to adopt a "unit" method of construction, the units being transistor circuits mounted on printed wiring boards (Fig. 17). The size and shape of these boards was standardized and contact made with the printed wiring by means of edge connecting sockets. Containers were developed for mounting convenient numbers of these boards and sockets. It is outside the scope of this paper to deal fully with the system of construction but some of the circuit details of the boards are of interest.

Two basic units were initially designed, a bistable unit and a pulse generator. Gating circuits of various types occur throughout the equipment but, in view of the difference in circuit detail and the small number of each type, their design as units was not undertaken.

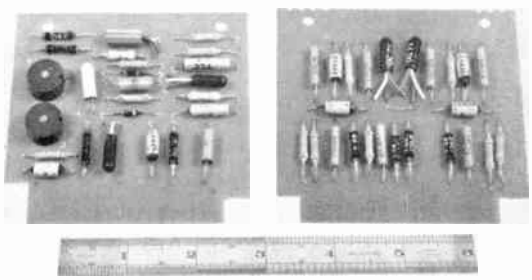


Fig. 17. Standard plug-in units.
(*Left*) Pulse generator. (*Right*) Standard bistable unit.

8.1. *The Bistable Unit*

The basic bistable circuit has a number of uses in the equipment: in shift registers, scale-of-two counters, ring counters and in complementing registers. It was decided to design the printed wiring board so that it could be used for any of the above purposes by selecting the components to be mounted on the board and by suitably wiring the contacts of the socket. Of the many ways in which the board may be used only the first three will be described here.

The basic unit consists of a printed wiring board on which are mounted the components of a transistor bistable circuit of the Eccles-Jordan type and two sets of diode gates with temporary memories (Fig. 18). Provision has been made in the design of the board for base-emitter voltage limiting diodes (D1 and D2) which prevent excessive back biasing of the bases and so increase the trigger sensitivity. Two emitter networks are possible, only one being used in the bistable circuit described. Access to different parts of the wiring has been provided to permit triggering, resetting and latching operations to be carried out.

8.1.1. Shift register stage

A shift register is a serial store which, on the instruction "shift", moves all of the digits held in the store by one digit in the direction of flow. A stage of shift register is shown in Fig. 19, together with the block symbol used to denote it.

The outputs from the preceding stage are used to determine which of the two diode gates allows the passage of shifting pulses. To achieve this the leading and trailing edges of the incoming waveform are

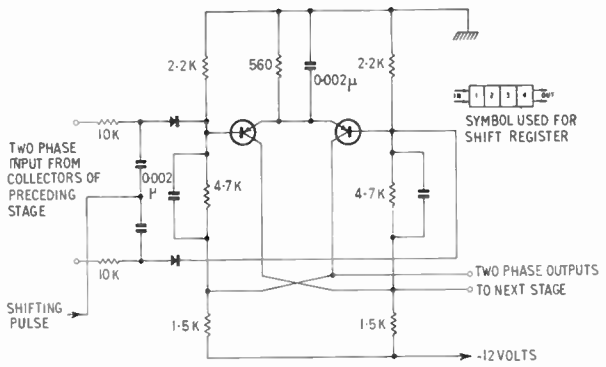


Fig. 19. Shifting register stage.

delayed by the resistance-capacitance network and shifting pulses superimposed on them through the capacitors. At the most positive excursion of the input, those shift pulses occurring *after* the full rise of the leading edge and *before* the fall of the trailing edge cause the diodes to conduct. The inputs to the bases of the bistable circuit are thus complementary positive pulse streams; these pulses cause the bistable circuit to change state in the same manner as the preceding stage but delayed by one digit period.

A stage of shift register may be formed from the standard unit by making the following connections to the socket:

Earth	pins P, N and F
−12 V supply	pin L
Input (i)	pins V and A
Input (ii)	pins U and B
Outputs	pins J and M
Shifting pulse	pins C and T for input (i)
	pins S and D for input (ii)

Those components shown in Fig. 18 are included on the board with the exception of diodes D1 and D2 and emitter network R_{e2} and C_{e2}.

8.1.2. Scale-of-two counter stage

The circuit is illustrated in Fig. 20; the counter is effectively a stage of shift register whose output is connected to its input.

Assume that transistor VT1 is conducting so that its collector is the more positive; when a positive pulse, which it is required to count, is applied through the input capacitors, then the diode D1 rather than D2 will conduct since its anode is more positively biased. The collector of VT1 therefore moves negatively and a change of state occurs, leaving VT2 conducting and VT1 non-conducting.

The next input pulse will pass through diode D2 and will cause the circuit to revert to its original state. The operation repeats for each input pulse, the counter returning to its original state every second pulse.

To make scale-of-two counters of any desired length, it is only necessary to connect one of the collectors of the first stage to the capacitor inputs of the second. The input collector waveforms are differentiated by the input network and the positive edges passed on through the diodes. Thus two changes of state of the first stage only causes one of the second.

The circuit is formed from the standard bistable unit by wiring it as a stage of shift register, and in addition connecting:

pin V to pin J
and pin A to pin M } feedback connections

Input from pin M of previous stage to pins C and T.

Counter stages may be reset in either of two ways: (*a*) by applying pulses to the bases of the transistors via pin R or pin E, or (*b*) by inserting a transistor, which is normally conducting, between either pin F and earth or pin P and earth and switching off the current when desired.

8.1.3. Ring counter stage

A stage of ring counter is formed from the unit if the stage is wired as a shift register, wire links replace the emitter networks and the link W_1 is omitted. The corresponding emitters of all the stages of the ring counter are then connected together in the way described in Section 3.2.2.

8.2. *The Pulse Generator*

The circuit used is based on a design by D. J. Hamilton[2] and the principal advantages are (i) that ease of triggering is not obtained at the expense of temperature stability, and (ii) a large output current is available during the period of the pulse.

A circuit diagram is shown in Fig. 21; transistor VT1 is reverse biased 0·2 V by the voltage across D1 and VT2 is reverse biased by the difference of the voltages across D1 and D2, i.e. by 0·7 V − 0·2 V = 0·5 V. D1 is a germanium junction diode and D2 can

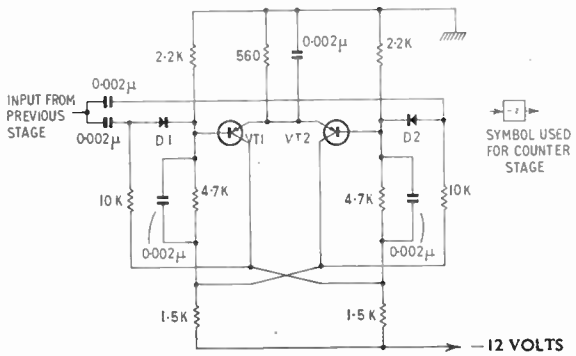


Fig. 20. Scale-of-two counter circuit.

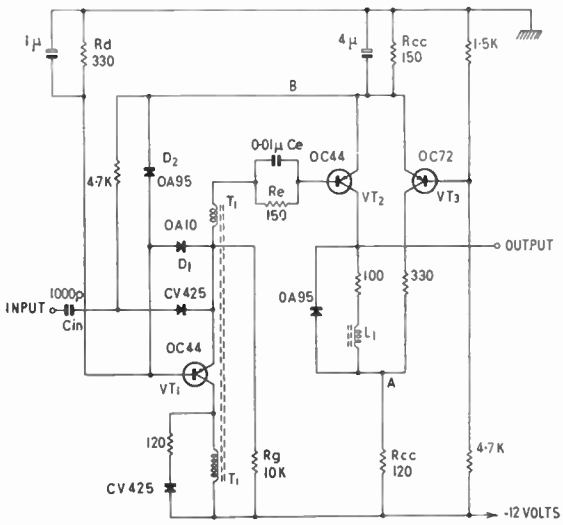


Fig. 21. Standard pulse generator.

T1 Ratio 4 : 1 Cores FX1011
Primary inductance ~ 2 mH
80 turns: 20 turns 38 s.w.g.
L1 ~ 7 mH
80 turns 38 s.w.g. Core FX1011

Note: OC41's should be used instead of OC44's from peak current considerations.

be either a silicon junction diode or a point-contact germanium diode.

A positive trigger pulse at the emitter of VT1 causes diode D1 to cease to conduct and the transistor to begin conducting; the transformer becomes clamped to the base of VT1 by the base-emitter diode and regeneration occurs, driving the base of VT2 negatively. The whole of the emitter current of VT1 flows through the base of VT2. The available output current during the pulse is βI_b and is therefore large.

Two modifications have been made to the original circuit;

(i) A low Q inductive load was used in place of the collector resistance of the output transistor to compensate for the capacitive loading of the boards.

(ii) A circuit has been added to provide the correct voltage levels to the generator when supplied from a 12 V source. Simple potentiometer methods have been found unsuitable because the average current consumption varies with frequency. For this reason also, it is impracticable to use presently available Zener diodes.

The point B is connected to the emitter of VT3 which acts as a constant-voltage source. The current flowing through the emitter resistors is therefore constant and divides between two paths, the line B and the transistor VT3. The majority of the varying current rejoins and flows through the common collector resistor R_{cc}, thus producing a voltage drop across it which is independent of either load or frequency.

The generator is capable of being driven from any of the bistable circuits and it has a pulse width of 6 μs. Its maximum operating frequency limit is well above the frequency limit of the bistable circuit previously described. The generator has been operated at an ambient temperature of 85° C and at a frequency of 62 kc/s, when it will drive a load equivalent to 40 counter or shift register stages without any indication of failure.

Manuscript received by the Institution on 16th January 1961 (Paper No. 678).

U.D.C. 621.375.2.029.64/5

Some Types of Low Noise Amplifier

By

R. HEARN, B.Sc.†

R. J. BENNETT, B.Sc.†

AND

B. A. WIND, B.Sc.†

Presented at the Convention on "Radio Techniques and Space Research" in Oxford on 5th–8th July 1961.

Summary: The need for amplifiers with optimum noise performance is discussed, with particular reference to the requirements of space communication. The paper then deals with the application of vacuum tubes, transistors, masers, parametric amplifiers and tunnel diodes to the problem of obtaining a low-noise receiver in the frequency range 100–1000 Mc/s. The advantages and disadvantages of these devices are discussed. Experimental results, with a brief outline of the underlying theory, are given for a varactor amplifier used as a preamplifier to a u.h.f. receiver. Results are also given for a transistor used as a self-oscillating mixer showing conversion gain and a low noise figure. Consideration is then given to the application of these devices to the various frequency ranges and means of optimizing parametric amplifier noise figures.

1. Introduction

As the range between a transmitter and receiver increases, the signal power at the receiver decreases while the noise power remains constant. Thus the received signal power after transmission over a sufficiently long distance becomes comparable with the noise power in the receiver and the signal may then be unintelligible. The situation can be relieved by an increase in the transmitted power or by a reduction of the inherent noise in the receiving system. Recent developments such as the maser, parametric amplifier and tunnel diode, have made feasible a considerable reduction in receiver noise.

In u.h.f. systems with conventional receivers, a large fraction of the total system noise is contributed by the receiver. In order to determine the system improvement possible by reducing receiver noise, the latter must be compared with noise arising from other sources.

Noise introduced in the transmission of radio or radar signals is from many different sources: sky noise, thermal noise from ground objects within the receiving aerial pattern, static, both natural and man-made, and jamming. Further, thermal noise is added by losses in the aerial, feeders or any other circuitry arrangement preceding the receiver proper.

Thus the input to the receiver consists of noise in addition to the required signal. As noise is a temperature-dependent phenomenon, it is convenient to represent the noise preceding the receiver by a quantity called the aerial noise temperature T_a, measured in absolute degrees. This is defined as the temperature at which a passive impedance, equal to the aerial impedance, would produce at the receiver input a thermal noise power equal to the observed noise power. Correspondingly, the receiver noise performance may be measured by an equivalent noise temperature T_e such that

$$T_e = \frac{\text{noise output from receiver}}{kB \times \text{receiver gain}}$$

provided that the aerial is disconnected and an equivalent impedance is connected across the receiver input terminals.

T_e may thus be directly compared with T_a, the total system noise then being proportional to $T_e + T_a$. Although this is a very convenient method of rating receivers, another expression is still widely used in which the noise figure F is given by

$$F = 1 + \frac{T_e}{T_0}$$

where T_0 is the standard reference temperature (290°K).

The noise figure of a receiver is a measure of the degradation of the system signal/noise ratio, by the receiver, when the aerial noise temperature is 290°K. For other aerial noise temperatures, the relation

$$F_{\text{eff}} = \frac{T_a}{T_0} + (F - 1)$$

must be used to give the effective noise figure F_{eff} in terms of the receiver noise figure F. Figure 1 shows this relationship for aerial noise temperatures from 1000°K to 50°K.

2. Survey of Low Noise Devices

2.1. *Vacuum Tubes*

2.1.1. Triodes

New construction techniques have made possible the development of triodes with very useful noise figures in the u.h.f. range. This has been accom-

† The Plessey Company Limited, Ilford, Essex.

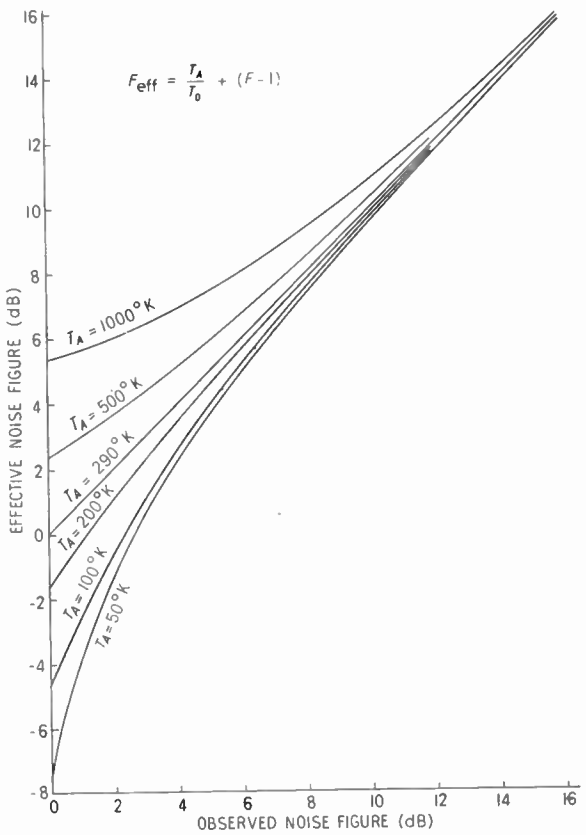


Fig. 1. Dependence of effective noise figure on receiver noise figure.

plished through the incorporation of frame grids and very close electrode spacings, both of which demand the highest quality precision engineering.

The Bell 416B planar triode has a noise figure of 3·0 dB at 500 Mc/s. This is the best triode performance to date and any further noise reduction using present techniques will be small in comparison with the increased manufacturing costs involved.

2.1.2. Travelling-wave tubes

A most valuable device for obtaining low noise amplification at high frequencies is the travelling-wave tube. The tube consists of an electron gun, a helix with input and output couplers and an electron collector as shown in Fig. 2 (*a*). When a signal is injected on the helix at the input end, it travels to the output end and, in doing so, velocity modulates the electron beam. The helix pitch is chosen to make the velocity of propagation slightly less than the beam velocity. The velocity modulation causes the formation of bunches which tend to collect in those parts of the field pattern of the propagation wave where they are decelerated. This enhanced bunching adds energy coherently to the wave on the helix producing the voltage increase shown in Fig. 2 (*b*).

Typical performance figures for travelling wave tubes are 2·5 dB noise figure at 3 kMc/s and 4·0 dB noise figure at 10 kMc/s with 10% bandwidth.

These devices are of major application to frequencies above 1 kMc/s since they become uneconomic below this figure.

2.1.3. The quadripole amplifier

Adler, Hrbek and Wade[1] have reported the development of a quadripole amplifier in which the amplification is produced by the parametric action of quadripole fields on the fast cyclotron wave of an electron beam.

An electron beam passes from a gun through an input coupler, an amplifying region and an output coupler, all the while flowing parallel to a uniform magnetic field. By arranging the cyclotron frequency to be nearly equal to the signal being amplified, the fast and slow waves may be widely separated and it becomes possible to couple the fast wave only. The Cuccia coupler to facilitate this comprises a pair of flat, parallel plates between which the beam passes and is modulated and demodulated respectively at the input and output. In the same way that the signal is stripped from the beam by the output coupler, any noise in the form of transverse fluctuations of the beam may be removed by the input coupler if it is correctly loaded.

The orbital radius of an electron is a measure of the amount of r.f. signal power it carries and so any increase in this radius corresponds to signal amplification. This amplification is achieved by passing the beam from the input coupler into a quadripole field, the strength of which is zero on the axis, increasing with displacement from it. This field is arranged to rotate with the electrons by driving the electrodes in pairs at twice the cyclotron frequency. Electrons in the beam fall generally into two categories, those in phase with the quadripole field and those in antiphase. The electrons of the first class will have their orbital radii

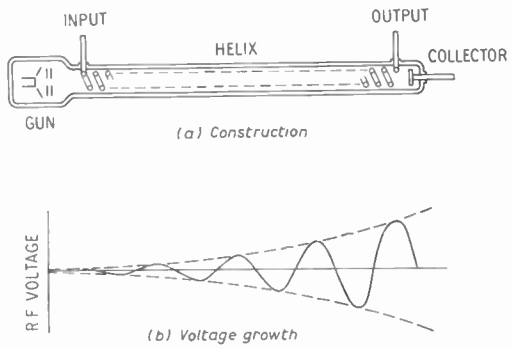


Fig. 2. Travelling wave tube.

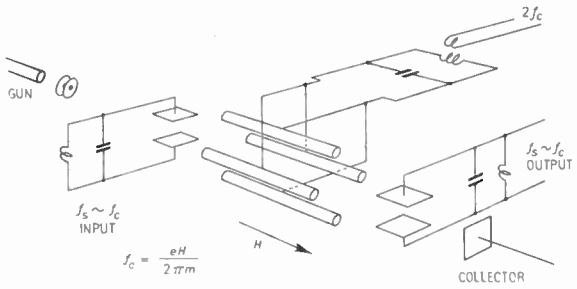


Fig. 3. Quadripole amplifier construction.

exponentially increased by the pump field while the orbital radii of the second class will experience an exponential decrease. On balance therefore there will be an overall increase of signal power in the beam. Since this is accomplished at the expense of r.f. pump power, there is a negative conductance action on the signal by the quadripole field. Figure 3 shows a schematic view of the quadripole amplifier.

The bandwidth of the device is limited only by the range over which efficient coupling can be effected and is at present restricted to 10% of the centre signal frequency. In the band 400–1000 Mc/s noise figures of 1·28 dB have been achieved while experimental tubes have produced 0·6 dB. At 4000 Mc/s, Bridges and Ashkin[2] have recorded a noise figure of 2·8 dB with a 60 Mc/s bandwidth.

These devices suffer from the finite lifetime which is characteristic of all thermionic devices, and lack of facility for tuning.

2.2. *The Transistor*

The noise performance of a transistor may be summarized by the equation relating noise figure to transistor parameters, constructed by Nielsen [3]:

$$F = 1 + \frac{r_b' + r_e/2}{R_g} + \frac{(r_b' + r_e + R_g)^2}{2\beta_0 r_e R_g}\left[1 + \left(\frac{f}{f_t\sqrt{1-\alpha_0}}\right)^2\right]$$

where R_g is the source resistance r_b' is the transistor equivalent base resistance, r_e its equivalent emitter resistance, α_0 and β_0 are the d.c. current gains in grounded base and grounded emitter respectively, f is the frequency of operation and f_t is the gain-bandwidth product of the transistor.

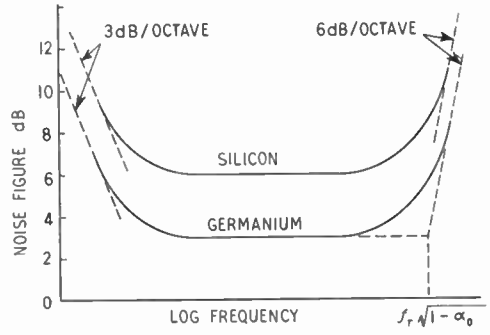


Fig. 4. Variation of transistor noise figure with frequency.

The equation is composed of a constant term, representing the resistive contribution of the input circuit, and a term representing the shot noise of the transistor. When $f_t \gg f$ the frequency dependent term drops out and the theoretical noise figure is a function of the source resistance, the intrinsic resistive elements of the transistor and β_0. Assuming a matched condition and that the noise contribution of the source resistance and the intrinsic transistor resistance are identical in nature, then the noise contribution of the source and transistor will be equal and the noise figure will be 3 dB. Figure 4 shows the variation of noise figure with frequency giving a flat region between 3 and 4 dB for germanium. The fall-off at the high frequency end is 6 dB per octave from frequency $f_t\sqrt{(1-\alpha_0)}$ while, at the low frequency end, the fall-off is at 3 dB per octave from a less well defined point. Silicon transistors follow the same shaped curve except that noise figure in the flat region is 6 to 7 dB, caused by the imperfect crystal structure of present day silicon transistors.

The Philco MADT L.5442 is a good example of the performance achievable with transistors. A typical noise figure is 3·8 dB at 200 Mc/s, which has been achieved by reducing the emitter area and decreasing the base width.

McCotter, Walker and Fortini[4] have reported the development of a coaxially packaged micro-alloy diffused-base transistor giving a noise figure of 9·9 dB at 1 kMc/s. This performance should be taken as a sign of things to come rather than an end in itself.

Much the same performances can be obtained with mesa construction but the noise performance may well be limited by the parasitics induced by the extremely small geometries involved in high frequency work.

It seems that present-day techniques of transistor production are coming near to their ultimate low-noise capabilities. Noise figures of 3 dB at 500 Mc/s and 6 dB at 1 kMc/s should, however, be achieved.

2.3. *The Maser*

Maser operation is based on the ordered rearrangement of electron spin energies in a crystal, such as ruby or rutile, doped with paramagnetic ions.

When a d.c. polarizing magnetic field is applied, the electron spin precesses around it with a characteristic angular velocity. If now a circularly polarized r.f. magnetic field is impressed on the spin, the rotation of the field being synchronous with the precession, the spin will absorb energy. Quantum mechanics shows that only certain spin energies are allowed, shown in

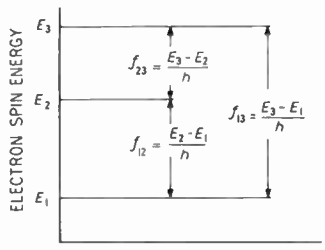


Fig. 5. Energy-level transitions in maser action.

Fig. 5, e.g. levels E_1, E_2 and E_3. To change the spin from a lower to higher energy requires an excitation at a frequency equal to the energy difference divided by Planck's constant. Similarly, radiation will be emitted at this frequency if the spin drops from one level to another.

To obtain amplification a crystal, doped with paramagnetic ions to increase spin density, should be pumped with a signal at frequency f_{13} to transfer spins from energy level 1 to level 3. Output power may then be taken at f_{23} or f_{12} by stimulating the required transitions, the frequency not used being called the idler. The required transitions may take place spontaneously in a random manner producing unwanted noise or when stimulated by an r.f. signal, adding energy coherently to that signal. This represents a conversion of r.f. pump power to signal power and thus is effectively a negative conductance action. The first unwanted effect may be reduced by cooling the crystal to liquid helium temperatures, thus giving a very low noise figure.

The maser may be operated at very high frequencies with extremely low noise-figures, but the expense and bulk of the associated cryostat and magnet make it impractical except for the highly specialized application.

Dr. J. V. Jelley of Harvard University reported the performance of a packaged 1420 Mc/s maser at the Institute of Physics "Solid State Microwave Amplifiers" Conference held in April, 1960. The noise figure of the maser was only 0·3 dB but it was found that the effect of connecting leads and various components and antenna spill-over raised the overall figure to 0·9 dB.

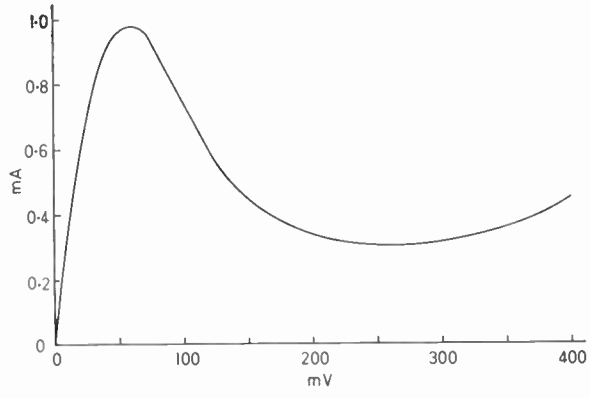


Fig. 6. Typical tunnel diode characteristic.

2.4. *Tunnel Diode*

Tunnel diode amplifiers, like masers and parametric amplifiers, use a negative conductance effect to provide amplification. A tunnel diode consists of a semiconductor p–n junction using highly doped material, giving it a low series resistance and a characteristic as shown in Fig. 6. In the region between 50 and 100 mV, the diode displays a negative conductance; the corresponding diode current is usually less than 10 mA.

In order to understand the reason for the device being a low noise amplifier, consider the circuit of Fig. 7 which is a schematic diagram of any negative conductance amplifier whether it be a maser, parametric amplifier or uses a tunnel diode. In the case of the maser, the negative conductance G is produced by a source of r.f. energy changing the population of

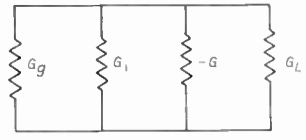


Fig. 7. Equivalent circuit of negative conductance amplifier.

energy levels of a paramagnetic salt. For the parametric amplifier, it is produced by the mixing of pump power into signal via the idler. The negative conductance of a tunnel diode is given by the slope of the diode characteristic at the operating bias point.

The noise figure of the amplifier is

$$F = 1 + \frac{T}{T_0}\left[\frac{G_1 + G_L + G_N}{G_g}\right]$$

where G_g = source, or generator conductance,

G_L = load conductance,

G_1 = loss conductance of the circuits associated with the amplifier,

G_N = inherent noise conductance of the amplifier,

T = amplifier temperature,

T_0 = standard reference temperature (290°K).

In the case of the maser, T is very small, so that the noise figure of the amplifier, using a circulator, is approximately that of the ideal amplifier. This is not usually the case for the parametric amplifier or tunnel diode so the minimum noise figure is determined by G_N. There is no shot noise contribution to

G_N for the parametric amplifier as the diode is in reverse bias. However, there is a contribution due to noise in the idler circuit given by

$$G_N = \frac{f_1}{f_2} \cdot G$$

where f_1 is the signal frequency,
f_2 is the idler frequency.

Thus, the noise figure for the parametric amplifier is

$$F = 1 + \frac{G_1}{G_g} + \frac{f_1}{f_2} \cdot \frac{G}{G_g}$$

when an ideal circulator is used.

For the tunnel diode there is no idler contribution due to G_N, but there is one from the shot noise so that

$$G_N = \frac{eI_0}{2kT}$$

where I_0 is the diode current. Thus, the tunnel diode will give low noise amplification by virtue of the low value of this current.

The performance of some tunnel diode amplifiers is shown in Table 1.

Table 1

Input frequency (Mc/s)	Output frequency (Mc/s)	Power gain (dB)	Noise figure (dB)
210	30	6	5·3 (ref. 5)
210	30	23	3 (ref. 5)
4000	4000	23	7 (ref. 6)

2.5. *The Parametric Amplifier using Semi-conductor Diodes*

When a semi-conductor diode is used in a parametric amplifier, its voltage sensitive capacitance is used to convert energy from a pump frequency f_p to a signal frequency f_s by interaction with another circuit tuned at an idler frequency $f_i = (f_p - f_s)$. The device presents a negative resistance input at the signal and idler frequencies and consequently can be used for amplification. It is a low-noise amplifier because, ideally, it is purely reactive and therefore no noise power is emitted. In practice, noise power is generated by the conductances that must be present in the signal and idler circuit.

In the simplest version, the pump frequency is chosen to be approximately twice the signal frequency so that the signal and idler frequencies are supported in one circuit. If the bandwidth of this tuned circuit is sufficiently large, the signal and idler frequencies need not be exactly the same so that the complications of the pumping at the correct phase do not arise. However, the noise figure of this amplifier will not be as good as for one in which the idler frequency is greater than the signal frequency. The former type of amplifier is sometimes referred to as a degenerate amplifier and the latter as a non-degenerate amplifier.

Negative resistance amplifiers are inherently unstable. One mode of operation of a parametric amplifier is up-converter operation in which the output frequency is the sum of the signal and pump frequencies. The power gain of this amplifier is the ratio of the output to the input frequency and the amplification is unconditionally stable. The disadvantage of this type of amplifier is that it requires a high pump frequency to achieve any useful gain.

One way of combining the high gain of a negative resistance amplifier with the stability of an up-converter is to use a negative resistance converter. In this case, the input is at the signal frequency f_s and the output at the idler frequency which is made much higher than the signal frequency. Some of the gain then arises from an up-converter action but it is combined with a negative resistance action. Thus there will be increased stability and a larger bandwidth as the negative resistance gain is decreased. It can also be shown that the noise figure of this type of amplifier, whilst not as good as that of an up-converter, is better than for a negative resistance amplifier. Seidel and his co-workers at Bell Telephone Laboratories,[16] using an input at 1000 Mc/s, a pump at 12 000 Mc/s and the output at 11 000 Mc/s, obtained an overall noise figure of 1·5 dB, a gain of 20 dB and a bandwidth of 20 Mc/s.

One basic shortcoming of the type of negative resistance amplifier which employs a single diode in a simple resonant cavity, is its lack of adequate bandwidth. Attempts to overcome this limitation have followed two separate paths, both of which have led to interesting and useful results.

The earliest approach uses a number of diodes distributed in some transmission line structure which has a broader bandwidth than a single resonant circuit. Signal and pump energy is fed down the line and, providing the phase velocities and group velocities at the signal and pump frequency obey certain conditions, amplification will result. It may be shown that these conditions lead to unidirectional gain which is an advantage over other types of parametric amplifier. Engelbrecht [13] has reported the performance of an amplifier of this type in which the input signal frequency was 600 Mc/s. The amplifier used 16 pairs of diodes and the results showed a bandwidth of 100 Mc/s, an average gain of 9 dB and a noise figure of 3·5 dB. The comparatively low gain of this amplifier should not be interpreted as a basic limitation of the device but rather on the particular values of input and

output mismatch between which the amplifier was required to operate.

An obvious and serious disadvantage of the above broadbanding technique is that it requires a number of high quality identical diodes. To circumvent this disadvantage, Seidel[14] has been experimenting with a technique in which a single diode interacts with suitable filter circuits. He obtained a useful bandwidth of about 100 Mc/s centred around 450 Mc/s.

3. An Experimental Parametric Amplifier

3.1. *Description of the Amplifier*

Experiments have been performed in the authors' laboratory with an amplifier of the negative resistance type with the input and output at the same frequency. Whilst this method does not use the inherent stability of the up-converter[7] or the increased bandwidth of the negative resistance converter,[8] this type was chosen because it can be used as a pre-amplifier for existing equipment.[9] The greater bandwidth is not a serious limitation because consideration of the role of the equipment shows that a bandwidth of 0·5 Mc/s with a noise figure of 3 dB is adequate.

The amplifier is shown diagrammatically in Fig. 8. It consists of a coaxial line cavity, short-circuited at one end, with the diode placed at the other. The cavity is resonant at the signal frequency in the quarter wavelength mode which is effectively shortened by the diode capacitance, whilst the idler resonance is provided by the three-quarter wavelength mode. Signal input and output is achieved by two inductive loops which may be matched into the cavity by an adjustable capacitance in parallel with the coupling loop.

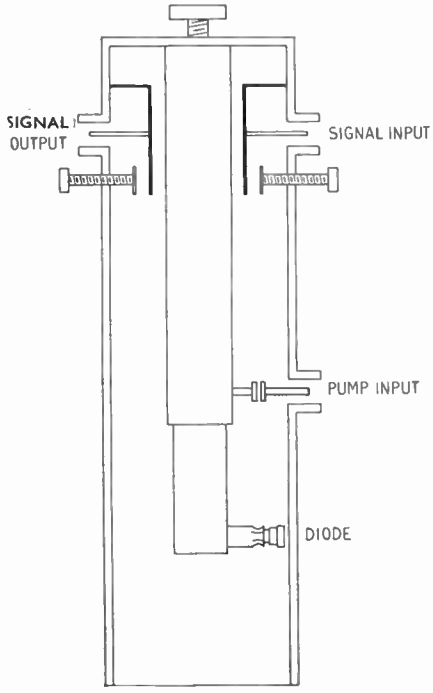


Fig. 8. Construction of experimental parametric amplifier.

The amplifier is tuned by altering the length of the inner conductor which is telescopic in construction and is varied by a screw adjustment. The frequency coverage is increased by the extra capacitance caused by the inner conductor approaching the cover at the open circuit end of the line.

The pump oscillator was a General Radio 1218A unit oscillator tuneable from 900–2000 Mc/s with an output power of 200 mW. The pump signal was fed into the cavity through a small capacitive probe placed near the diode. Whilst a circuit resonant at the pump frequency was not provided, it was found that the pump power available was sufficient to overcome any cavity losses at the pump frequency.

3.2. *Theoretical Considerations*

Expressions for the gain, bandwidth and noise figure for the negative resistance type amplifier have been obtained by Heffner and Wade.[10] Analysis shows that the power gain g is given by

$$g = \frac{4G_g G_L}{(G_1+G_g+G_L-G)^2} \qquad \text{.....(1)}$$

where G_g is the antenna or generator conductance,

G_L is the receiver or load conductance,

G_1 is the loss conductance of the cavity, due principally to the diode.

G is the negative conductance introduced into the signal circuit by the interaction of the idler circuit and the pump. It is proportional to the pump power and is given by

$$G = \frac{\omega_1 \omega_2 C_p^2}{4G_2}$$

where $\frac{\omega_1}{2\pi} = f_1$ is the signal frequency,

$\frac{\omega_2}{2\pi} = f_2$ is the idler frequency,

G_2 is the loss conductance of the idler circuit,

C_p is proportional to the pump voltage swing at the diode and is given by the time-varying diode capacitance which is approximated to

$$C = C_0 + C_p \sin \omega_p t$$

where $\frac{\omega_p}{2\pi} = f_p$ is the pump frequency.

One of the most important parameters for describing the amplifier is the noise figure F which is given by

$$F = 1 + \frac{G_L}{G_g} + \frac{G_1}{G_g} + \frac{\omega_1}{\omega_2} \cdot \frac{G}{G_g} \qquad \text{.....(2)}$$

the term G_L/G_g being introduced because the amplifier is a two-terminal device in which noise in the subsequent stages contributes to the noise figure. It is to be expected that the cavity design is such that $G_1 \ll G_g$ so that a low noise figure is obtained when $G_L < G_g$ and $\omega_1 < \omega_2$. The term G_L/G_g cannot be made zero, unless a non-reciprocal device is used, because the amplifier would have zero bandwidth as is obvious from examination of the expression for the gain bandwidth product which is

$$g^{\frac{1}{2}}.B = \frac{2f_2(G_L G_g)^{\frac{1}{2}}}{f_2 Q_1(G_g+G_L)+f_1 Q_2 G} \quad \ldots\ldots(3)$$

where B is the half-power fractional bandwidth,

Q_1 = quality factor of the signal cavity,

Q_2 = quality factor of the idler cavity.

Assuming that $G_1 < (G_L+G_g)$ and that operation is for a high gain, i.e. $G \simeq G_L+G_g$, eqn. (3) becomes

$$g^{\frac{1}{2}}.B = \frac{2\left(\frac{G_L}{G_g}\right)^{\frac{1}{2}}}{\left(1+\frac{G_L}{G_g}\right)\left(Q_1+\frac{f_1}{f_2}.Q_2\right)} \quad \ldots\ldots(4)$$

This is a maximum when $G_L = G_g$ but for a noise figure of 3 dB it is necessary to make $G_L < G_g$ and sacrifice bandwidth.

Figure 9 shows gain plotted against G/G_g for various values of G_L/G_g. As G_L/G_g decreases, the value of G/G_g necessary for instability is decreased as

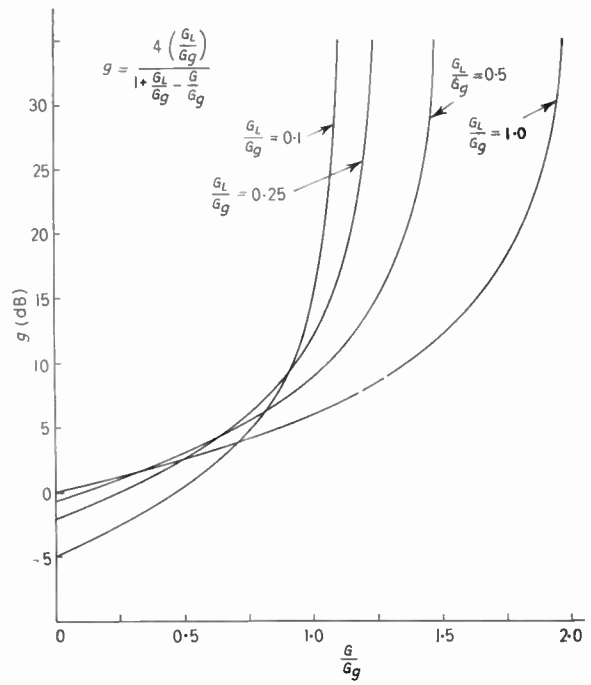


Fig. 9. Variation of gain with negative conductance for various amplifier loadings.

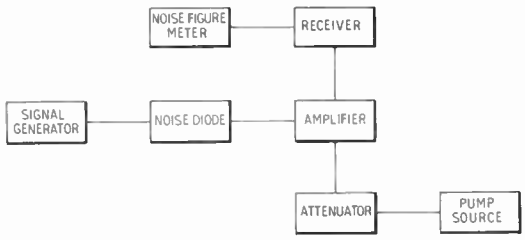


Fig. 10. Experimental arrangement for evaluation of the performance of the parametric amplifier.

the slope of the curve increases. Thus, less pump power is necessary but the pump source has to have greater stability.

3.3. *Design Considerations for the Amplifier*

For 20 dB gain, Fig. 9 shows that $G \simeq G_g+G_L$ if $G_1 < (G_L+G_g)$. Equation (2) then becomes

$$F = \left(1+\frac{\omega_1}{\omega_2}\right)\left(1+\frac{G_L}{G_g}\right)$$

The mode of operation requires that

$$\frac{\omega_1}{\omega_2} = \frac{1}{3}$$

Hence, for $F = 3$ dB, $G_L/G_g = \frac{1}{2}$.

Having specified the ratio G_L/G_g, the bandwidth of the amplifier is determined by eqn. (4).

The quality factors Q_1 and Q_2 are those of the diode at the respective frequencies. The diode used in the experiment (G.E.C. SVC21) has a zero bias capacitance of 4 pF and a series resistance of 4Ω. If the signal frequency is 400 Mc/s,

$$Q_1 = 25$$
$$Q_2 = 8$$

This leads to an expected bandwidth of approximately $1\frac{1}{4}$ Mc/s for $G_L/G_g = \frac{1}{2}$.

3.4. *Experimental Results*

The experimental arrangement is shown in Fig. 10. In order to obtain $G_L/G_g = \frac{1}{2}$, the output coupling loop was rotated so that its plane was at an angle of 45 deg to the axis of the inner conductor. Measurements were then made of parametric amplifier gain and noise figure for combination of amplifier and receiver for various settings of the pump attenuator, and the results are shown in Fig. 11. It can be seen that the noise figure suddenly increases for large gains. This is thought to be due to the pump oscillator swinging the diode into forward conduction thus introducing shot noise in the diode. It is seen from Fig. 12 that, for reduced G_L/G_g, the pump power necessary for oscillation is insufficient to drive the diode into forward conduction.

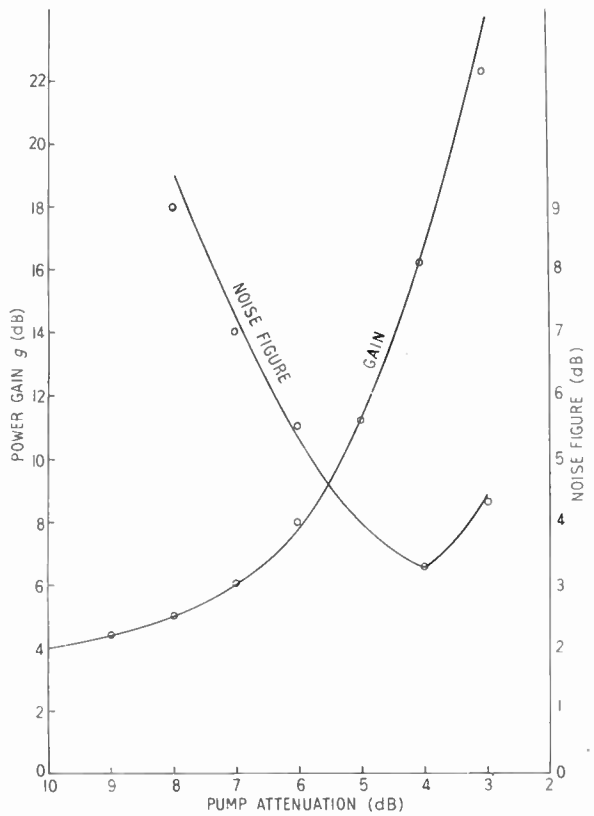


Fig. 11. Performance of parametric amplifier at 400 Mc/s for $G_L/G_g = 0{\cdot}5$.

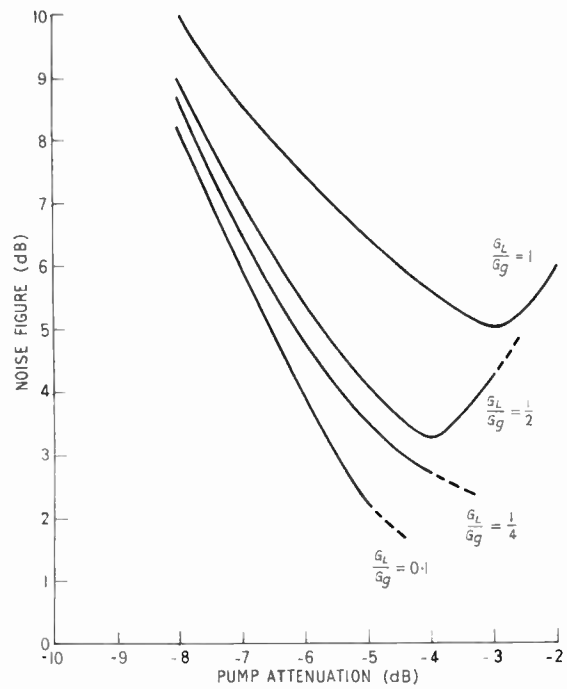


Fig. 12. Effect of loading on the noise figure at 400 Mc/s.

The experiments were then repeated for various frequencies between 225–400 Mc/s and the noise figures obtained plotted in Fig. 13. The results do not lie on a smooth curve but have a distribution between 3 and 4·5 dB. This is probably due to variations of the input impedance of the receiver which result in different contributions to the noise figure by the load at different signal frequencies.

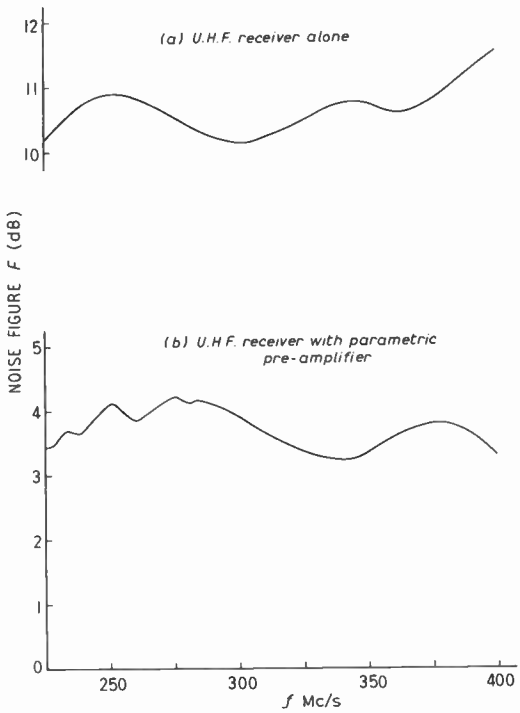


Fig. 13. Noise figure of a u.h.f. receiver (*a*) without and (*b*) with a parametric amplifier.

4. Transistor Self-oscillating Mixer

Vodicka and Zuleeg[11, 12] have shown that certain high frequency transistors (Philco 2N502 and Hughes GXG4) may be operated in a new mixing mode at u.h.f. to produce very large gains with low noise figures. The circuit which is shown in Fig. 14 is a self-oscillating mixer incorporating a "Hartley" type local oscillator with the transistor in grounded base connection.

The work reported here was carried out, in the main, using the Philco MADT 2N502, although similar results were also obtained subsequently with the Philco 2N1158 and T1832 transistors. The maximum oscillation frequency of the 2N502 is 500 Mc/s so the butterfly circuit maintaining the local oscillations was made tuneable over the range 200 Mc/s to 400 Mc/s. The output was taken at 1·85 Mc/s across a tuned circuit in the collector. The conversion

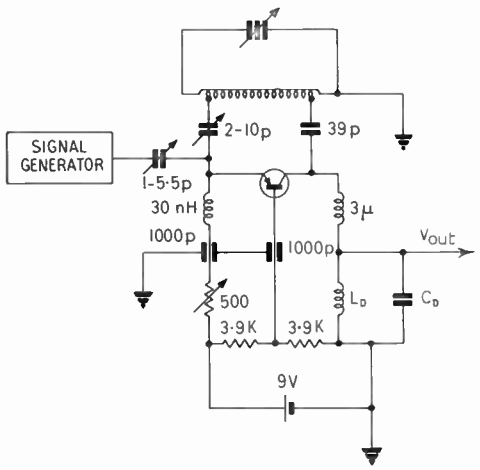


Fig. 14. Circuit of self-oscillating mixer.

gain (*C.G.*) is defined by

$$C.G. = 20 \log_{10} \frac{V_{out}}{V_{in}} \text{ decibels}$$

It should be noted that this is essentially a voltage ratio and should not be treated as a power gain.

It was found that the conversion gain increased with emitter current until it reached 60 dB; at this point the circuit oscillated. After a further increase of emitter current, the oscillation region was passed and stable operation was resumed at lower gain. The oscillation region coincides with the value of emitter current giving maximum conversion gain and it has a symmetrical response as shown in Fig. 15. Further

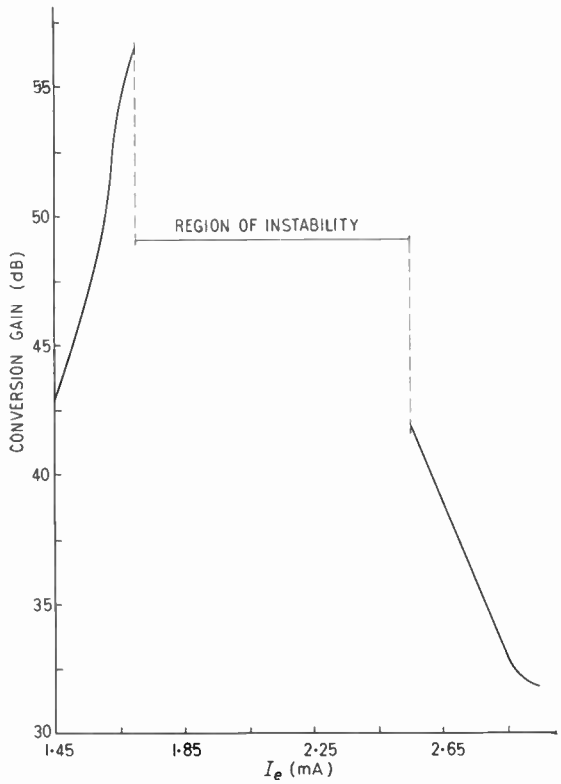


Fig. 15. Conversion gain versus emitter current at 307 Mc/s.

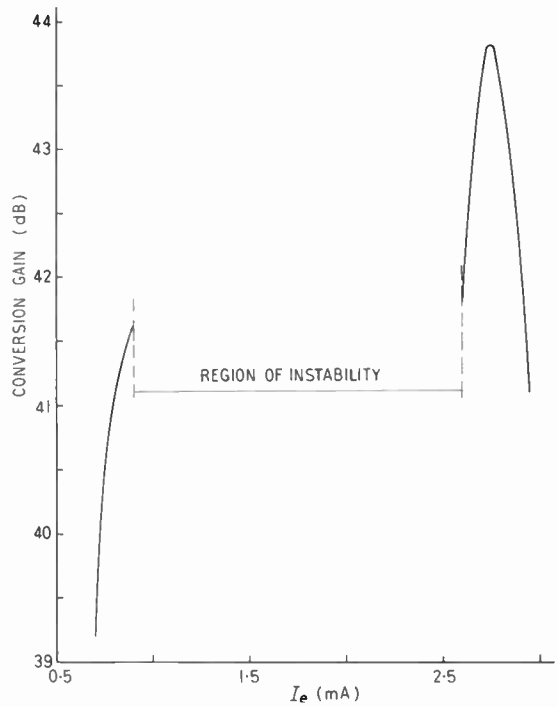


Fig. 16. Conversion gain versus emitter current at 225 Mc/s.

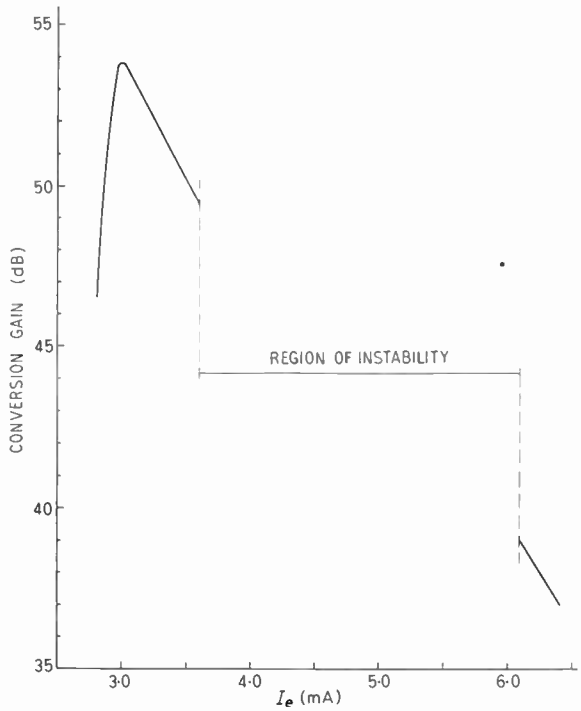


Fig. 17. Conversion gain versus emitter current at 320 Mc/s.

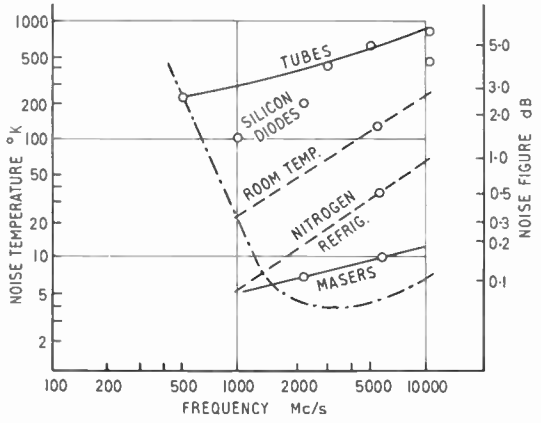


Fig. 18. Comparison of noise performance of masers, parametric amplifiers and vacuum tubes. The chain linked line represents the variation of sky noise with frequency.

work in this laboratory has shown that, at various signal frequencies, an asymmetrical response may be obtained in which the oscillation region occurs before or after the conversion gain maximum, as shown in Fig. 16 and 17. This behaviour cannot be explained by the theories presented so far and it may be caused by a positive feedback effect. This theory is supported by the observation of extremely narrow bandwidths which were only 0·05%.

The measured noise figure was 12 to 13 dB over the frequency range 200–400 Mc/s using a wide-band noise generator; figures of 7 dB at 450 Mc/s have been reported for this transistor.[15]

5. Conclusions

The performance of the three main types of low-noise amplifier, i.e. masers, parametric amplifiers and vacuum tubes, is compared in Fig. 18. A plane triode developed at Bell Telephone Laboratories has yielded a noise figure better than 3 dB at 500 Mc/s, while noise figures of 3·5 dB at 3000 Mc/s have been obtained with travelling-wave tubes by several workers.

Maser operation is indicated by the lowest curve in the diagram which represents results obtained with a ruby travelling-wave maser at Bell Telephone Laboratories.

The performance of parametric amplifiers lies generally in the region between masers and vacuum tubes and is represented by the two broken lines and the separate experimental points. The broken lines represent the performance to be expected for a parametric amplifier optimized either by using a very high pump frequency or by refrigerating the idler load. A gallium arsenide diode is assumed for the optimized amplifier and the performance at room temperature is shown in the upper curve and at liquid nitrogen temperature in the lower one.

The variation of sky noise with frequency is shown by the chain-linked line in the diagram, demonstrating a minimum in the range 1500 to 10 000 Mc/s. These figures correspond to an antenna angle of approximately 40 deg to the vertical.

It can be seen that, from noise considerations, the performance of vacuum tubes is best matched to the sky noise in the several hundred megacycle range, parametric amplifiers in the u.h.f. and low microwave regions, and masers in the mid-microwave range. The maser then is ideal for the reception of weak microwave signals which traverse the atmosphere at a steep angle, as in space and satellite communication and radio astronomy. Below 1000 Mc/s masers are not suitable because of the high sky temperature which increases rapidly as the antenna looks down toward the horizon. This leaves parametric amplifiers with their very low noise figure, and travelling-wave tubes with their relative simplicity and advanced state of development.

The possibility of a u.h.f. solid-state pump source in the near future for the parametric amplifier promises a low-drain, long-life device particularly suitable for compact systems.

6. Acknowledgment

The authors are indebted to the directors of the Plessey Company Ltd. for permission to publish this paper.

7. References

1. R. Adler, G. Hrbek and G. Wade, "The quadripole amplifier, a low noise parametric device", *Proc. Inst. Radio Engrs*, **47**, p. 1713–23, October 1959.
2. T. J. Bridges and A. Ashkin, "A microwave Adler tube", *Proc. Inst. Radio Engrs*, **48**, pp. 361–3, March 1960.
3. E. G. Nielsen, "Behaviour of noise figure in junction transistors", *Proc. Inst. Radio Engrs*, **45**, pp. 957–63, July 1957.
4. J. D. McCotter, M. J. Walker and M. M. Fortini, "Coaxially packaged madt for microwave applications", *Trans. Inst. Radio Engrs*, (*Electron Devices*), **ED-8**, p. 8, January 1961.
5. K. K. N. Chang, G. H. Heilmeier and H. J. Prager, "Low noise tunnel-diode down converter having conversion gain", *Proc. Inst. Radio Engrs*, **48**, pp. 854–8, May 1960.
6. R. Trambarulo and C. A. Barnes, "Esaki diode oscillators from 3 to 40 kmc", *Proc. Inst. Radio Engrs*, **48**, pp. 1776–7, October 1960.
7. D. Leenov, "Gain and noise figure of a variable capacitance up-converter", *Bell Syst. Tech. J.*, **37**, No. 4, pp. 989–1008, July 1958.
8. J. D. Pearson and J. E. Hallett, "Comparisons of gain, bandwidth and noise figure of variable reactance amplifiers and convertors", *Proc. Instn Elect. Engrs*, **107B**, pp. 305–10, May 1960.
9. J. G. Cottrell, "New u.h.f. air-to-ground communications for the British armed forces", *Brit. Commun. Electronics*, **6**, pp. 586–91, August 1959.

10. H. Heffner and G. Wade, "Gain, bandwidth and noise characteristics of the variable parameter amplifier", *J. Appl. Phys.*, **29**, pp. 1321–31, September 1958.

11. V. W. Vodicka and R. Zuleeg, "Transistor operation beyond cut-off frequency", *Electronics*, **33**, No. 35, pp. 56–60, 26th August 1960.

12. R. Zuleeg and V. W. Vodicka, "A new gain and power concept in transistors with circuits extending the frequency range into the microwave region", *I.R.E. International Convention Record* **9**, Part 4, pp. 191–201, 1961.

13. R. S. Engelbrecht, "Non-linear Reactance (Parametric) Travelling Wave Amplifier for U.H.F.", I.R.E. Solid State Circuits Conference Digest. February 1959.

14. H. Seidel and G. F. Herrmann, "Circuit aspects of parametric amplifiers", *I.R.E. Wescon Convention Record.* Part II, Circuit Theory, 1959.

15. V. W. Vodicka, private communication.

16. E. D. Reed, "Diode parametric amplifiers", *Semiconductor Products*, **4**, p. 35, February 1961.

Manuscript received by the Institution on 8th May 1961 (Paper No. 679)

DISCUSSION

Mr. C. J. Beanland: In Section 2.2 of your paper you quote incorrectly an equation of Nielsen:

$$F = 1 + \frac{r_b' + r_e/2}{R_g} + \frac{(r_b' + r_e + R_g)^2}{2\beta_0 r_e R_g}\left[1 + \left(\frac{f}{f_t\sqrt{1-\alpha_0}}\right)^2\right]$$

In the final term, you have used the function f_t (f_1), i.e. short circuit current gain, grounded emitter. However, Nielsen's paper (ref. 3) uses the function f_α (frequency where $\frac{\alpha}{\alpha_0} = \frac{1}{\sqrt{2}}$) not f_t.

I would also like to draw your attention to a comment on Nielsen's work which draws attention to the possible error of the corner frequency (h.f. end) at which the noise factor degrades at 6 dB/octave.† The letter also points out the significantly different results obtained if f_t (or f_1) is used instead of f_α in calculating the noise factor.

May I ask Mr. Bennett if he has had any experience in using a parametric up-converter at very much higher power levels than the conventional application of low noise input stages and, if so, up to what signal power level. I am considering the possible use as a frequency translating device from i.f. to microwave. Present day devices can provide reasonably high power very efficiently, as frequency multipliers. I suspect these properties may be applied to the use I have proposed.

The authors (*in reply*)**:**

Mr. Beanland is quite justified in correcting the reference to Nielsen's equation and we confirm that the original contained f_α and not f_t. The justification for effecting the substitution is that we have quoted the equation to demonstrate the manner in which the noise figure of a junction transistor depends on its various parameters. Since noise figure depends in the same way on both f_α and f_t the qualitative relation is still valid.

† H. F. Cooke, "Transistor upper noise corner frequency", *Proc. Inst. Radio Engrs*, **49**, p. 648, March 1961 (letter).

We note with interest Cooke's letter but would have more faith in the results quoted if the operating point had been specified when stating figures for f_t and f_α. The general conclusions in the letter, regarding Nielsen's neglect of the first three terms in the calculation of upper noise corner, are of great value.

We have no experience with up-converter operation but we understand‡ that a conversion gain is possible using this mode and that the power output is limited to that of the pump source. This may be verified by reference to the Manley-Rowe§ relationships—

$$\frac{w_p}{f_p} + \frac{w_o}{f_o} = 0 \qquad \ldots\ldots(1)$$

and

$$\frac{w_s}{f_s} + \frac{w_o}{f_o} = 0 \qquad \ldots\ldots(2)$$

where w_p, w_s represent the power at pump and input frequencies, f_p and f_s respectively, flowing into the non-linear element. w_o is the power out of the non-linear element at frequency $f_o = f_s + f_p$. Equation (2) is the normal gain equation of an up-converter but eqn. (1) gives

$$(-w_o) = \frac{f_o}{f_p} w_p \qquad \ldots\ldots(3)$$

We believe the frequencies f_p and f_o in Beanland's application are very close to each other ($f_o = 4{\cdot}00$ Gc/s and $f_p = 3{\cdot}93$ Gc/s) so that (3) would agree with Rowland's findings of output power being equal to pump power.

Beanland mentions conversion from i.f. to microwave by these devices and this is certainly an attractive means when the i.f. power contains information. If, however, it is purely carrier-wave conversion, then the advantage is undoubtedly with high-efficiency varactor harmonic generation.

‡ R. Rowland, private communication.

§ J. M. Manley and H. E. Rowe, "Some general properties of non-linear elements: Part I. General energy relations", *Proc. Inst. Radio Engrs*, **44**, No. 7, p. 903, July 1956.

News from the Sections . . .

South Wales

A large audience was at the first meeting of the session at Cardiff on 11th October, to hear Mr. G. B. Townsend present an excellent lecture on colour television in general and the SECAM system in particular. This was a very timely choice of subject and gave useful background knowledge to the serious attempts being made by the industry to introduce a colour receiver of an acceptable standard and at a lower price than hitherto.

Whilst the N.T.S.C. system employs continuous transmission of two colour signals, SECAM uses a line sequential transmission and line-to-line storage, a delay line in the receiver providing for simultaneous display. The much simplified colour information circuitry of the SECAM receiver was claimed to result in benefit to the user because the receiver would use perhaps six fewer valves than the N.T.S.C. receiver.

Mr. Townsend's large array of equipment included a flying spot scanner which gave an excellent demonstration of SECAM receivers. Vertical misregistration of colour could not be observed during the demonstration and Mr. Townsend aptly referred to this by pointing out that "What the eye doesn't see the television engineer doesn't grieve over!" There were many pairs of eagle eyes in the audience who had to agree with him, but it is only fair to add that the demonstration was on 625 lines.

One was left wondering just what colour television system will eventually be adopted in Britain.

C. T. L.

South Western

The first meeting of the Session on 11th October was held at the School of Management Studies, Bristol, the subject being "Inertial Navigation". The first speaker was Mr. R. Collinson who gave a "General Introduction to Inertial Navigation". He showed that it was essential to establish accurately a reference for a set of axes, to be followed by the use of precision accelerometers and finally a computer to carry out the calculations necessary. The fundamental requirement generally accepted is the setting up of "stable table" which can be held to its position to fractions of a second of arc. Gyros placed on this table are used as the most sensitive measuring accelerometers available. The concept of "Schuler tuning" of the stable table was also discussed; this concept, named after its originator, is based upon an 84-minute pendulum and enables the system to maintain the table aligned to the local horizon at any position above the earth's surface.

Mr. R. Bristow then spoke on "Components and Techniques used in Inertial Navigation", and developed the theme laid down by Mr. Collinson. He showed the role of the computer in carrying out the two special integrations needed to determine, from measured acceleration and the given inertial references, the velocity and distance travelled in the inertial axes. Both authors spoke of the need to incorporate modifying allowances for the correction of errors.

D. R. M.

North Western

The Section held its first meeting of the new session on 5th October in the Reynolds Hall, College of Science and Technology, Manchester, when Mr. G. H. Smethurst read a paper on "The Basic Principles of Digital Computers". The emphasis of the paper was on computer applications as opposed to the circuitry; particular reference was made to a commercial data processing system and a film of the system was used effectively to support the general description.

Mr. Smethurst pointed out that facilities for automatic coding were provided in the system and a common language coding system known as COBOL (Common Business Oriented Language) enabled the programmer to write in a restricted form of English, giving as little consideration as was practical to the computer being used. Subsequently, the computer, acting as a compiler, will translate the written instruction into machine language. A wide range of computer applications was then described.

F. J. G. P.

Southern

The first meeting of the Section was held at Farnborough Technical College on 26th September. In his opening address the new Chairman, Dr. W. A. Gambling, paid tribute to the enthusiastic leadership and support given to the Section by the retiring Chairman, Cdr. J. Brooks.

The subject of Dr. Gambling's Address was "Lightning—Facts and Fancies" and he began by outlining some of the early beliefs and superstitions and the considerable influence which storms and lightning had over the affairs of state. He then demonstrated with the aid of a van de Graaf generator some of the experiments performed by Benjamin Franklin which resulted in the first lightning conductors. A detailed analysis of the action of the discharge was then given and Dr. Gambling concluded by a discussion of the effects of electric storms on radio reception, and he described and demonstrated with the aid of tape recordings the phenomenon known as "whistlers".

J. M. P.

(*Editorial note:* This report has been shortened in view of the forthcoming publication in the *Journal* of the full text of Dr. Gambling's address.)

U.D.C. 537.525:551.510.5

The Use of Probing Electrodes in the Study of the Ionosphere

By

R. L. F. BOYD, Ph.D.†

Presented at the Convention on "Radio Techniques and Space Research" in Oxford on 5th–8th July 1961.

Summary: The Langmuir probe and its developments provide a means whereby such important ionospheric quantities as electron concentration and temperature, and ion mass spectrum, concentration and temperature may be measured. These quantities are of fundamental importance in any consideration of the ion and electron equilibrium in the ionosphere. The problems involved arise largely because the vehicle is isolated from Earth and because of the varying aspect of the vehicle. They thus require very careful study before any instrumentation is built and their solution involves the use of special techniques which are not needed in the laboratory. The ways in which the problems have been tackled will be discussed and special reference will be made to the instrumentation planned for the first Anglo-U.S. satellite.

1. Introduction

The Langmuir probe technique has long been a standard method for determining the electron and ion concentrations and the electron temperature in a discharge plasma. Several significant extensions to the theory and practice of the method have been made. Notably, of special importance in the study of the ionosphere, by the Druyvesteyn theory,[1] the introduction of a grid[2]; the use of double probes,[3] and the development of modulation methods.[4] The method is based on the fact that a charged electrode in a plasma produces around itself a space charge sheath whose thickness is given approximately by

$$d \simeq 1{\cdot}3\left(\frac{eV}{kT}\right)^{\frac{3}{4}} . h$$

where V is the modulus of the potential on the electrode relative to that of the plasma

T is the electron gas temperature

e and k are the electronic charge and Boltzmann's constant

h is the Debye shielding radius $= \left(\frac{kT}{4\pi Ne^2}\right)^{\frac{1}{2}}$

and N is the electron concentration.

In a laboratory plasma the Debye shielding radius is commonly a fraction of a millimetre. It is of the order of 1 cm in the ionosphere and rises to tens of centimetres in interplanetary space. Because the electric field of the probing electrode is largely localized within a distance of a few Debye lengths, the current-voltage characteristic of the electrode is primarily determined by the concentrations and energies of the charged particles constituting the plasma. Of these the electrons normally contribute strongly to the form of the characteristic as their high velocity insures that their random flux in the plasma is high. An actual characteristic for electron current to a probe on *Explorer VIII* is given in Fig. 1. The slope of the curve gives the electron temperature, the intersection of the dotted lines gives the space potential and the random electron flux.

It transpires indeed that full analysis of the current-voltage curve enables such quantities as the electron and ion concentrations, the plasma potential, the electron and ion temperature (and even the electron energy distribution in non-thermal plasma) to be determined.

2. Magnitude of Probe Currents

The Debye length is the characteristic length in a plasma. By taking it as the metric the mathematics may be given a parametric form which makes for a ready comparison between different circumstances. Thus it can be shown that the random flux of electrons in a thermal plasma crossing a sphere of radius equal to the Debye length is

$$i_D = \left[\frac{1}{2\pi} \cdot \frac{e}{m} \cdot \left(\frac{kT}{e}\right)^3\right]^{\frac{1}{2}}$$
$$= 1{\cdot}49 \times 10^{-11} T^{3/2} \text{ amps}$$

If a be the radius of a spherical probe and we define $S = a/h$, then the random electron current to the probe when it is at plasma potential is

$$i_{e0} = 1{\cdot}49 \times 10^{-11} T^{3/2} . S^2 \text{ amps}$$

It is common for S to be of order of unity and in the ionosphere $T \simeq 1000°$ K so the currents expected are of order of $\frac{1}{2}$ microamp. The random ion currents are smaller by a factor equal to the root of the ratio of the particle masses—commonly $\sim \frac{1}{200}$, though

† Department of Physics, University College, London.

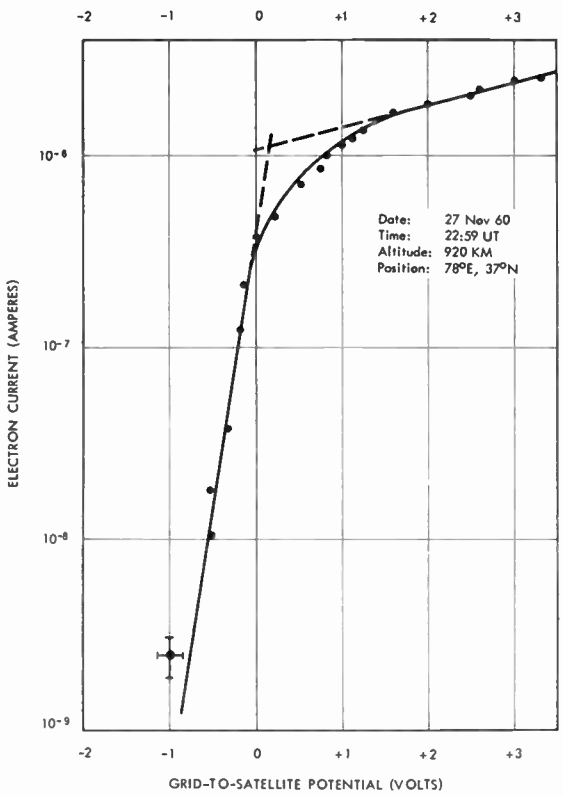


Fig. 1. Probe characteristic from plane gridded probe on *Explorer VIII.*

probes on hypersonic vehicles may collect ion currents greater than random by the order of the Mach number.

3. The Problem of the Vehicle Motion

In the laboratory both the theory and practice of Langmuir probes is often found to be a good deal more complex than it might seem at first sight. Prominent amongst the deviations from the simple theory are: (1) The fact that the electron current still continues to rise rather rapidly with increase of the probe potential in spite of the shielding effect of the sheath, and (2) that in a running discharge, where the temperature of the electrons is higher than that of the ions, the current of positive ions to the probe is controlled far more by the temperature of the electrons than by the temperature of the ions themselves. This fact leads to certain useful simplifications of the theory,[6] which unfortunately are not applicable in the isothermal conditions of the ionosphere.

On an unstabilized fast moving rocket or satellite there are a number of additional problems of major significance.

They may be listed as follows:

1. The problem of returning charges collected by the probe to the ionosphere again—the "current dumping" problem.
2. The effects of photo-emission from probe and vehicle structure.
3. The effects of vehicle motion on the relative velocities of the ions, especially the resulting anisotropy.
4. The effects of vehicle motion on vehicle potential.
5. The problem of making very small current measurements in the time and with the telemetry band width available.

In some circumstances other problems arise. For example, electrodes may become contaminated; vehicle outgassing may affect the particle concentrations or motions, or it may prove impracticable to position a probe at an adequate distance (several sheath thicknesses) from the rest of the structure. In what follows, however, attention will be concentrated on the more fundamental problems listed above. The limited history of Langmuir probes on space vehicles is the history of an endeavour to overcome these difficulties.

A rocket or satellite in the ionosphere adopts a potential with respect to its surroundings such that the current of positive ions arriving plus the current due to photo and secondary electrons leaving is equal to the current of electrons arriving. The flux of photo-electrons is between 10^{-9} and 10^{-8} amps cm^{-2} of normally illuminated surface. At the F layer maximum this is only a few per cent of the electron random flux but as altitude increases the electron random flux falls until at a distance of about half an Earth radius it is comparable to the photo-emission. Nearer the Earth a space vehicle may therefore be expected to adopt a potential somewhat negative to the space. At greater distances it will become positive so as to limit photo-emission. Of course, in the Earth's shadow it will always tend to be negative. Measurements to date suggest that the potential of space vehicles near the Earth usually varies from a fraction of a volt up to 2 or 3 volts negative.

As the aspect of the vehicle changes both the flux of ions swept up by the vehicle and the photo-emission change. As a result the potential of the vehicle changes. It is this changing reference potential which has so far made successful measurements of electron temperature rather rare. Moreover the potential of the vehicle may change, not only because of changing aspect but because of the finite impedance between the vehicle and the ionosphere. Unless this impedance is made very small compared with the impedance of the probe to its surroundings the varying current to the probe and thus from the vehicle to the ionosphere will result in significant variations in vehicle potential. In practice this means that the area of a probe which is intended to be driven positive should be about one

ten thousandth of the area of the conducting surface of the vehicle. Thus a vehicle 1 m in diameter might employ a probe 1 cm in diameter.

The group at Michigan University who pioneered the use of Langmuir probes in space vehicles with their use of V-2s, have attempted to overcome these problems by using a sphere carrying needle probes. A symmetrical double probe system in the shape of a dumb-bell has also been used. The symmetrical system has the advantage that the currents are fairly large but the disadvantage that only a very small part of the *i-V* curve is available for the measurement of electron temperature whilst the interpretation of the curve in terms of electron density is uncertain. Results obtained, however, seem to be too much in variance with the expected atmospheric model to be likely to be correct. The needle probes on the sphere ejected from the rocket is a better arrangement in that it makes it possible to obtain complete Langmuir curves up to space potential on the probe.

Bourdeau and his collaborators[4, 5] had a high degree of success with *Explorer VIII* by using a rotationally stabilized system and by taking complete characteristics in a time short compared with a rotational period. They also made use of gridded probes, as did the Russians in *Sputnik III*.[6] The use of grids makes it possible to separate ion and electron components of current and also to eliminate the photo-emission component.

In *Sputnik III* a gridded spherical probe was used to trap positive ions so as to give a measure of the current swept up and so of the concentration. The spherical shape made the trap insensitive to vehicle aspect. In *Explorer VIII* retarding potentials on a grided plane probe were used to determine the energy of arrival of the positive ions. The high data rate of the telemetry enabled curves to be obtained before the aspect of the probe had changed greatly. The system was not able to resolve individual ion species but it did show that the mean molecular weight at an altitude of 1000 km is just under 16 e.m.u.

Successful flights with Langmuir probes have also been made by Ichimiya, Takayama and Aono,[7] and by the group at University College, London. The former have found that a pair of crossed rings behaves very much like a spherical probe but, of course, has a much reduced photo-electric current. The workers at University College have used spherical probes with two small diametrically opposed inset electrodes, one of which is of necessity shadowed, in order to remove the effect of the photo-current. As already noted, a spherical probe is insensitive to vehicle aspect. Plane probe systems have also been used.

Although this paper is only dealing with Langmuir probes it should be noted that several probe methods depending on the use of radio frequencies near or at the critical frequency have been used with success.

4. The Anglo-U.S. Langmuir Probe Experiments

The instrumentation for the University College experiments on the first Anglo-U.S. satellite will be described later in this symposium. As the probe instrumentation presents a new approach in space experiments, based on some fairly extensive laboratory experience, the principle behind them will be described here.

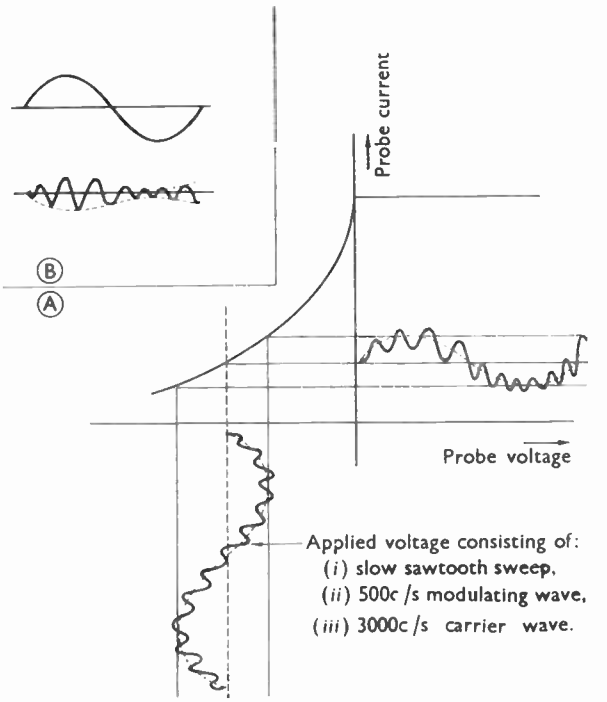


Fig. 2. Illustrating the measurement of first and second derivatives of an *i-V* characteristic by modulation methods. The output current contains the two a.c. components shown in inset B. The amplitude of the upper curve provides a measure of the first derivative while the fractional depth of modulation in the lower curve gives the electron temperature.

Telemetry capacity in the satellite is such that complete probe curves cannot be obtained in a time small compared with the vehicle rotational period. The satellite is roughly spherical in shape with a gridded spherical probe for measurement of ion energies (and hence masses) mounted on the spin axis. Two simple plane probes with guard rings for the determination of electron temperature and concentration are mounted normal to the spin axis, one on the axis at the opposite end to the ion probe and the other on a 4 ft arm extending at right angles to the vehicle axis.

The principle of operation of the electron temperature probes is as follows. The current to the probe when the probe is negative is given by

Fig. 3. Gridded sphere and circuitry for determination of ion mass spectrum and temperature.

$$i_e = i_{e0}\left[\exp\left(-\frac{eV}{kT}\right)\right] - i_+ - i_p$$

where V is the amount by which the probe is negative to the plasma

i_+ is the positive ion current to the probe

i_p is the photo-electric current from the probe.

Differentiating with respect to V we can write, since i_+ and i_p are nearly constant,

$$\frac{di_e}{dV} = i_{e0}\left(-\frac{e}{kT}\right).\exp\left(-\frac{eV}{kT}\right)$$

and again $$\frac{d^2i_e}{dV^2} = i_{e0}\left(\frac{e}{kT}\right)^2.\exp\left(-\frac{eV}{kT}\right)$$

whence $$\frac{di_e}{dV}\Big/\frac{d^2i_e}{dV^2} = -\frac{kT}{e}$$

Now if we apply two small audio-frequency oscillatory voltages to the probe the resulting current will contain a mixture of the two with a modulation of one upon the other. The amplitude of the wave is a measure of (di_e/dV) and the fractional depth of modulation is a measure of $\dfrac{d^2i_e}{dV^2}\Big/\dfrac{di_e}{dV} = \dfrac{e}{kT}$. See Fig. 2.

The fractional depth of modulation remains unchanged over most of the characteristic while the voltage on the probe is driven by a saw-tooth generator. Hence the electron temperature may be determined in the vehicle and be directly fed to the telemetry. From the value of (di_e/dV) when $(d^2i_e/dV^2 = 0)$ (i.e. at space potential), the electron concentration may be found unencumbered by the photo currents. This procedure has the advantage that the result is insensitive to small fluctuations in vehicle potential.

The determination of the energy spectrum of the ions proceeds in a similar way. Druyvesteyn has shown that the energy spectrum of electrons striking a probe is given by

$$f(E) \propto V^{\frac{1}{2}}\frac{d^2i_e}{dV^2}.$$

If a spherical probe is used this result holds for anisotropic distributions, and if the probe is surrounded by a grid slightly negative to the probe the result may be applied to the stream of positive ions striking it.

We have already seen that (d^2i/dV^2) may be obtained from the mixing of two audio frequencies so that in this way the energy distribution function may be found. The central energy of a peak in the distribution provides a measure of the masses of the related group of ions, while the width of the peak enables an estimate of ion temperature to be made. It is anticipated that the resolution will be sufficient to enable the important constituents H^+, O^+, and N_2^+ to be resolved.

5. References

1. M. J. Druyvesteyn, "Der niedervoltbogen", *Z. Phys.*, **64**, pp. 781–98, 1930.
2. R. L. F. Boyd, "The collection of positive ions by a probe in an electrical discharge", *Proc. Roy. Soc.*, A, **201**, pp. 329–47, 1950.
3. R. L. F. Boyd and N. D. Twiddy, "Electron energy distributions in plasmas. I", *Proc. Roy. Soc.*, A, **250**, pp. 53–69, 1959.
4. R. E. Bourdeau, 2nd COSPAR Symposium on Space Science, 1961.
5. J. E. Allen, R. L. F. Boyd and P. Reynolds, "The collection of positive ions by a probe immersed in plasma", *Proc. Phys. Soc.*, B, **70**, Pt. 3, pp. 297–304, 1957.
6. K. I. Gringauz, V. V. Bezrukikh and V. D. Ozerov, *Artificial Earth Satellites*, Issue 6, p. 63, 1961.
7. T. Ichimiya *et al.*, "Measurement of positive-ion density in the ionosphere by sounding rocket", *Nature*, **190**, No. 4771, pp. 156–8, 8th April 1961.

Manuscript received by the Institution on 3rd July 1961 (Paper No. 684).

U.D.C. 621.319.5: 621.9

Spark Machining Fundamentals and Techniques

By

G. V. SMITH, B.Sc.†

Presented at the South Western Section's Convention on "Aviation Electronics and its Industrial Applications" held in Bristol on 7th–8th October, 1960.

Summary: The principles on which spark erosion machines operate and some of the various types of spark generators and electrode servo controls are described. The relation between metal removal rates and surface finishes produced by the so-called "relaxation" circuit is established and improvements to the basic circuit are discussed. Brief descriptions of the developments in high power spark-arc circuits are given and methods of electrode manufacture with various types of materials are reviewed.

1. Introduction

The basic requirement of any industry is to be able to produce goods, whether they be aircraft, ships or nuts and bolts, as well and as cheaply as possible. It follows that the production engineer must be provided with plant and machinery having the highest possible efficiency.

Machines in the metal-working industries are today being made more efficient by the application of control techniques and it will soon be possible to say that these machines are working close to their maximum theoretical efficiency. It is important to realize, however, that there is a fundamental limitation in a machine which chips, hacks, cuts or grinds metal. This limitation is the strength of the tool such as the milling cutter, the drill, or the lathe tool, which is removing the metal.

For many years the aircraft industry, in particular, has demanded new metal alloys of increasing toughness and strength for use in engines and structures. The metallurgist has provided these and is working now to produce even stronger materials. In other industries, materials such as tungsten carbide, which have great hardness and can work at high temperatures, are now in common use.

Conventional machining methods, however, are proving inadequate in dealing with these materials and the cost of machining even the most simple part is alarmingly high.

For this reason new machining methods are being introduced which can overcome the limitations of the conventional machine-tool. One of the most important of these new methods is spark-machining which employs electrical energy directly for the machining of metals. Other names commonly used to describe the same machining principle are Spark-erosion and Electrical Discharge Machining (E.D.M.).

2. Fundamental Principles

2.1. *The Spark Discharge*

As its name implies, spark-machining employs the energy present in a spark discharge to erode metal. If two pieces of metal are separated by a small air-gap and an increasing d.c. voltage is applied to them, at some point the air becomes completely ionized and current flows across the gap. During the first hundred microseconds or so the path of the current has a very small cross-section area and therefore the current density is extremely high. The current path then gradually broadens and the current density therefore is reduced. The discharge during this first hundred micro-seconds is known as a spark, which degenerates after this period into an arc.

Owing to the very high current density in the spark, temperatures of several thousand degrees Celsius are attained in its path. If the discharge ceases before the arc forms and the two pieces of metal are carefully examined it is found that small volumes of metal at either end of the discharge have been vaporized, and craters produced on the metal surfaces. It is also observed that the piece of metal connected to the positive terminal of the d.c. supply has lost far more material than that connected to the negative, and the amount of erosion is not dependent upon the hardness of the metal. When the experiment is performed using a suitable liquid dielectric between the metals, it is found that the erosive action at the anode is considerably intensified.

† Sparcatron Limited, Tuffley Crescent, Gloucester.

2.2. *History*

Such experiments and observations have been made by many people during the last hundred years or so, but the first serious proposal to use a train of spark discharges for machining metals was made by B. R. and N. I. Lazarenko in 1946.[1] Since that time many workers in countries all over the world have developed this technique of spark-erosion and applied it to a very large field of metal-working. The first British patent was granted to Rudorff in 1950.[2]

During the early stages of development, spark-machining circuits and machines were limited to classes of work where they provided the only practical method of machining, for instance, the "tapping" of holes in tungsten carbide and the shaping of very hard alloy steels. However, recent developments, as described later in this paper, have made possible the machining of more common metals at the same rate as by conventional methods and in some cases faster.

An important advantage of spark-machining is that no mechanical strain is produced in the material being machined, as is so common with conventional tools. Even after a hardening process has been applied to them, it is possible to machine high-tensile steels without distortion.

3. Mechanical Arrangement of the Machine

Figure 1 shows the arrangement used in the majority of spark machines. The piece of material being machined, probably steel or tungsten carbide, known as the work-piece, is immersed in a tank filled with dielectric fluid. This fluid quenches and removes the vaporized metal more efficiently than air and also prevents oxidation of the heated metal in the craters

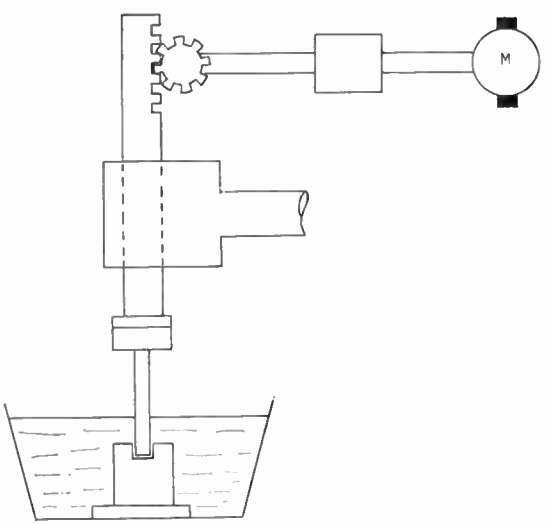


Fig. 1. Mechanical arrangement of electrode and work-piece in spark-machining.

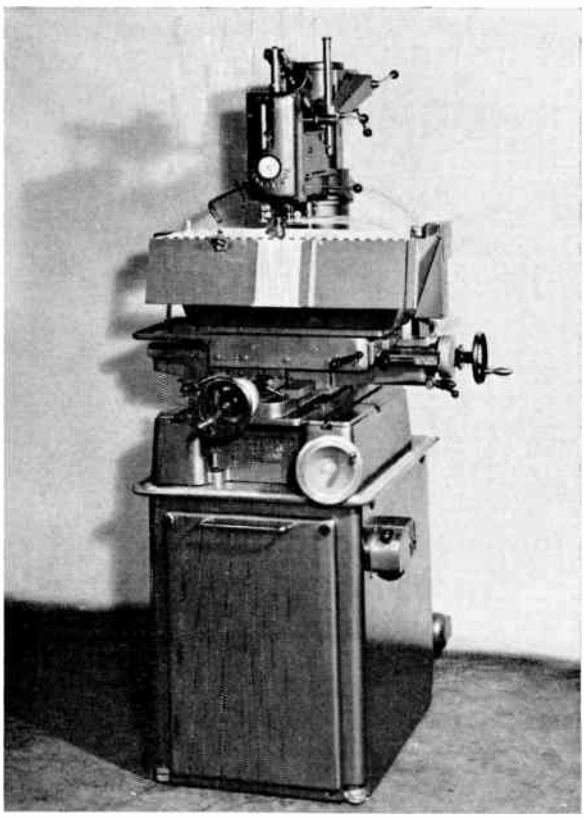


Fig. 2. A typical spark erosion machine.

produced. Liquids commonly used for this purpose are paraffin and light oils.

A metal electrode is moved towards the work-piece by means of an arrangement such as a rack and pinion or lead-screw driven by an electric motor (M). The function of the motor is to maintain a constant minimum gap between the electrode and the work-piece as a continuous train of sparks erodes away material.

Since a spark will occur always at the shortest gap between electrode and work-piece it can be seen that, apart from slight deformation due to the erosion of the electrode, the shape of the electrode will be "copied" into the work-piece.

Figure 2 shows a typical spark-erosion machine and Fig. 3 is a photograph of an impression made in die steel using this machine. Fuller descriptions of the erosion process have already been given in the literature.[3]

4. Spark Erosion

4.1. *The Relaxation Circuit*

The most simple circuit used for producing a series of sparks is shown in Fig. 4.

Capacitor C is connected across the electrode and work-piece and is charged from a d.c. supply via

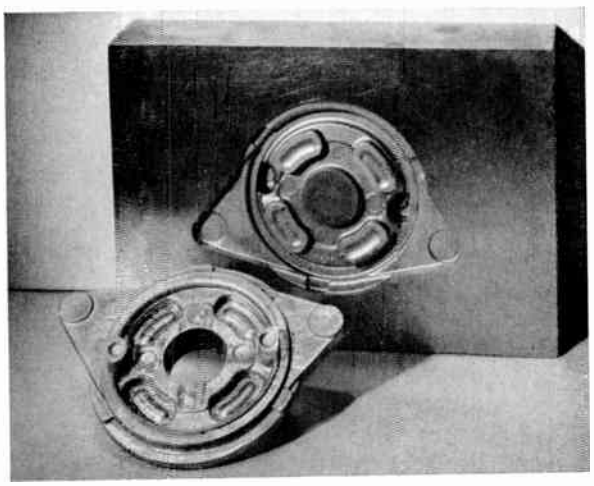


Fig. 3. Example of an electrode and impression made by spark machining.

resistor R. The voltage across C rises until the gap between the electrode and work-piece breaks down. The capacitor then rapidly discharges its stored energy as a spark and its voltage drops to zero. The gap then de-ionizes and C again charges through R. The voltage across C therefore, has the characteristic saw-tooth appearance, as shown in Fig. 5, of the well-known relaxation oscillator. Here the spark-gap acts as the non-linear element.

4.1.1. Metal removal rate and surface finish

The size of crater produced by a single discharge is related to the energy E of the spark. This energy may be expressed as that stored in the capacitor when the voltage equals the break-down value V_B. This is: $E = \frac{1}{2}CV_B^2$. As this energy is increased the roughness of the work-piece surface being machined is also increased. With constant value of V_B therefore, the surface roughness increases with the value of C.

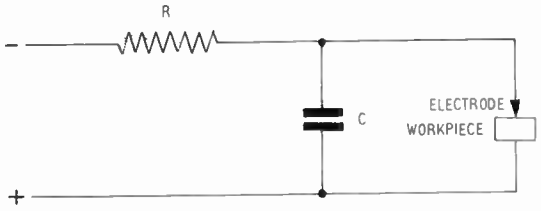


Fig. 4. The relaxation circuit.

The amount of metal removed per spark cannot be expressed simply as proportional to E since variations in such parameters as melting point, conductivity and geometry of the work-piece, are not negligible. However, we can express the rate of metal removal γ (gamma) as the product of the spark repetition frequency F and some function of the energy f(E); i.e. $\gamma = F.\mathrm{f}(E)$. If we neglect the time taken for the capacitor to discharge it can be simply shown that

$$F = \frac{1}{CR \log_e \left(\frac{V_S}{V_S - V_B} \right)}$$

where V_S is the supply voltage,
and thus

$$\gamma = \frac{1}{CR \log_e \left(\frac{V_S}{V_S - V_B} \right)} . \mathrm{f}(E).$$

In other words, the rate of metal removal is theoretically governed by the values of V_B, V_S, C and R.

The maximum values of V_B and V_S are determined by the safety of the equipment operator and insulation considerations. It is very rare, therefore, to find

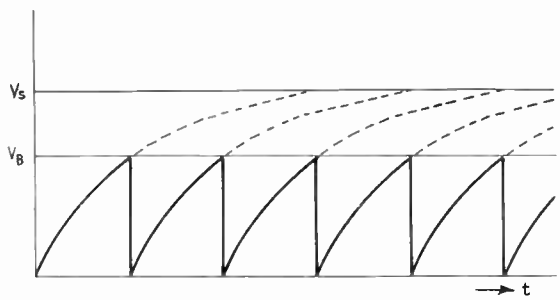


Fig. 5. Voltage across capacitor C in the relaxation circuit.

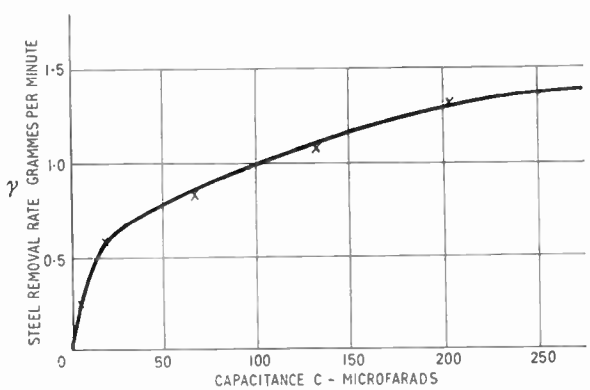


Fig. 6. Steel removal rate vs. capacitance in the relaxation circuit.

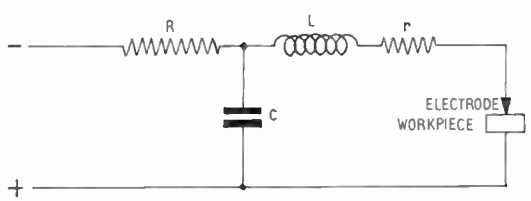


Fig. 7. Relaxation circuit showing high-frequency parameters.

voltages much above 200 volts being used for V_S. Both V_B and V_S may be regarded as fixed quantities.

The values of C and R are determined by the operation of the relaxation oscillator. As mentioned above, the value of C determines the surface finish, and if a small value of C is used to give a smooth surface finish, the value of R must be made relatively large. If R is too small the capacitor voltage rises after a discharge more rapidly than the dielectric can de-ionize and a sustained arc is formed between the electrode and work-piece, causing damage to the surface.

If a large value of C is used, giving a rough surface, R can be made relatively small but cannot be reduced indefinitely without, again, continuous arcing occurring across the gap.

In practice with a value of 150 and 90 volts for V_S and V_B respectively, the metal removal rate varies as shown in Fig. 6 where the minimum value of R has been used for each capacitor value.

As can be seen in practice, the stock-removal rate drops as the frequency increases, i.e. with finer finishes. This is largely due to the drop in power output of the relaxation circuit at small values of C. The normal procedure in spark-machining, therefore, is to remove the bulk of metal by "roughing" at high speed (using a large capacitor) followed by a "fine-finishing" operation with a low-valued capacitor to reduce the surface to an acceptable smoothness.

4.1.2. The discharge circuit

The relaxation circuit is re-drawn in Fig. 7 to include the high-frequency parameters of the discharge circuit. The leads connecting the electrode and work-piece to the capacitor possess both resistance r and inductance L.

To reduce power loss the connection resistance r is made low and therefore in the discharge circuit it is usually true that the value of r is much smaller than that of $2\sqrt{(L/C)}$. Thus, the sudden change in capacitor voltage at discharge gives rise to a current having the oscillatory form shown in Fig. 8. This oscillatory condition not only gives rise to a low peak spark energy, with a consequent loss of machining efficiency, but the current reversals produce increased electrode erosion.

Obviously every precaution must be taken in spark-machine circuits to reduce the value of L. Many designs arrange for the capacitor to be mounted adjacent to the electrode and connections are made with coaxial or basket-woven cables.

4.1.3. Inductive charging

The limitation in metal-removal rate using the relaxation circuit is largely due to the inter-relation of the de-ionization mechanism of the dielectric and the charging of the capacitor.

One method which is employed to improve this condition is shown in Fig. 9 where inductance is inserted in the charging portion of the relaxation circuit. This has the effect of delaying the rise in voltage on the capacitor after the discharge has occurred and the dielectric is allowed more time to de-ionize. Since a lower value of R can be employed with a given value of C, an improvement in maximum metal removal rate of 20% to 30% can be achieved by this modification.[4]

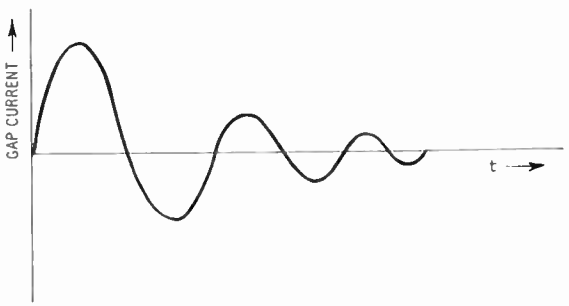


Fig. 8. Discharge current in the relaxation circuit.

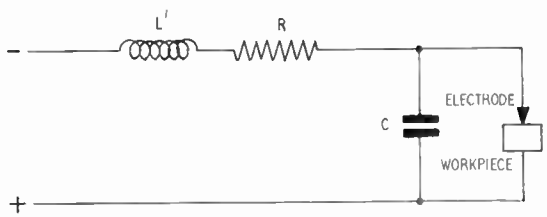


Fig. 9. Inductive charging in the relaxation circuit.

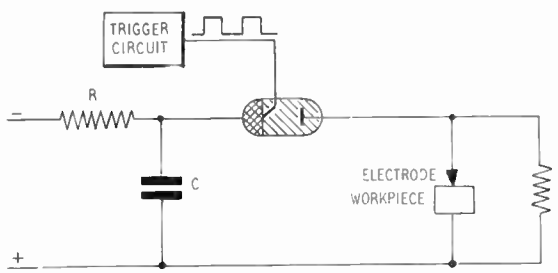


Fig. 10. Discharge control by ignitron.

4.2. *Independent Circuits*

It has been recognized by most spark-machine designers, however, that only by ensuring complete independence between the discharge and the charging cycles can any outstanding improvement be made in metal removal rates, which are, under ideal conditions, limited to 2 grammes of steel per minute with the relaxation circuit.

A simple improvement to circuit design is shown in Fig. 10. An ignitron is connected in the discharge circuit and is "fired" by externally applied trigger pulses. The value of R can be made very low without arcing occurring and the circuit is capable of some 6–7 g/min steel removal. A serious drawback to this arrangement is the time required for the mercury vapour in the ignitron to de-ionize. This sets a limit on the discharge repetition frequency to a few hundred cycles per second. Maximum metal removal rates are then obtained only with very rough surface finishes.

Various circuits, similar in principle to those employed in radar work for producing pulses up to a hundred microseconds duration, have been tried in spark-machining.[5, 6] Much space could be devoted to details of these pulse circuits, which are described in the literature, but to generalize, they are designed to produce high-energy pulses at low current and high voltage. For application to spark-machining it is necessary to introduce transformer coupling between the circuit and the electrode work-piece discharge path to give a high current transient. The design of this transformer is a serious practical problem since the spark-gap cannot be regarded as a stable load and any mismatch gives rise to oscillatory currents, resulting in poor machining efficiency.

A further type of circuit is shown in Fig. 11. Here, the capacitor is charged via a multitude of cathode-followers connected in parallel. Pulses from a multi-vibrator are fed to the grid circuits. This arrangement gives the advantage of full control over the charging cycle up to very high frequencies.

The one rather serious limitation lies in the number of valves required in parallel—some 70 or 80—which presents a major problem of unit failure when valves become unserviceable. Reliability of valves under this type of operation cannot be guaranteed.

5. Spark-Initiated Arc Erosion

5.1. *The Controlled Arc Discharge*

All circuits so far discussed employ a d.c. supply and have as their aim the production of true spark discharges, in other words, transients lasting for only 100 microseconds or so.

Due to the enormous difference between peak and mean power requirements it is usually necessary to employ an energy storage means, such as a capacitor. The charge and discharge of a capacitor is therefore the basic mechanism of true spark-erosion circuits. They have the advantage of simplicity but the disadvantage of limited performance at high frequency with high power.

It has been found that some of the problems encountered in spark-erosion circuits can be overcome by use of spark-initiated arcs—or prolonged spark discharges—of controlled duration. As mentioned earlier the current density in the path of an arc is much lower than that of a spark owing to the larger cross-section area of the arc. The temperature of the arc is therefore relatively low and has the effect, when used for machining, of melting rather than vaporizing the metal.

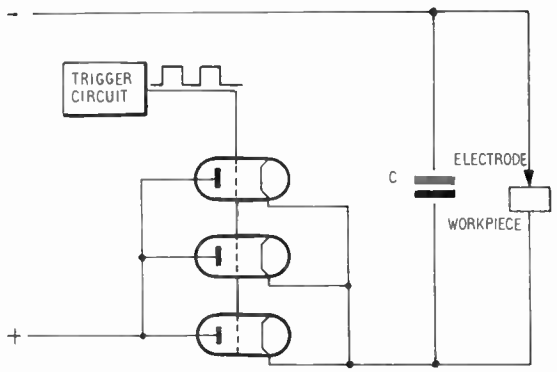


Fig. 11. Capacitance charging via cathode followers.

The circuit requirements in this case are much simplified. Instead of peak-to-average current ratios of 1000 or more, as is required for a spark circuit, an arc circuit needs only a ratio of 10 or less. It is possible for the circuit designer to dispense with the energy storage capacitor and consider circuits to provide high-power uni-directional pulses.

5.2. *Half-Wave A.C. Systems*

A simple and inexpensive circuit for producing uni-directional pulses to give controlled spark-initiated discharges is shown in Fig. 12. A 50 c/s single-phase mains supply is half-wave rectified and applied directly to the working gap. The discharge commences when the voltage reaches the break-down value of the gap, producing a spark which degenerates into and is maintained as an arc until the voltage falls below the ionization level of the gap (Fig. 13).

Metal removal rates of over 40 grammes of steel per minute have been obtained with this circuit in the laboratory but, generally speaking, the discharge duration is so long when using 50 c/s mains that excessive heating of the metal occurs. Modifications

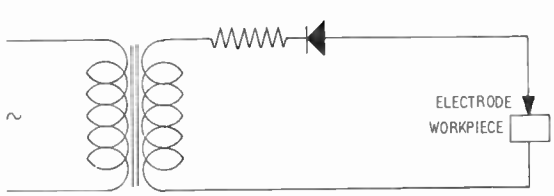


Fig. 12. Half-wave spark/arc circuit.

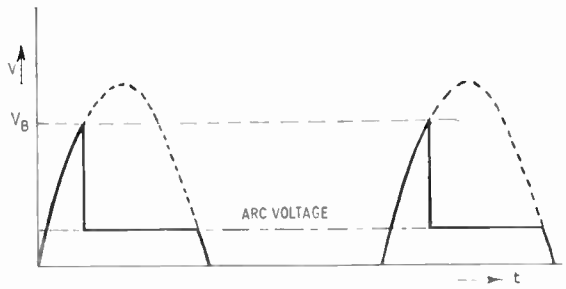


Fig. 13. Gap voltage characteristic with half-wave circuit.

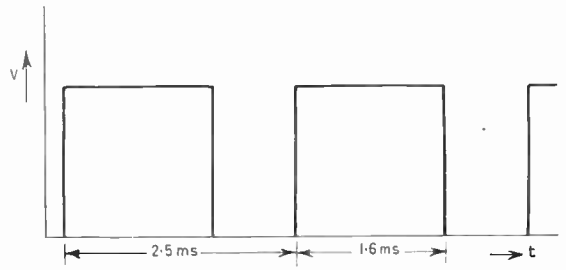


Fig. 14. Idealized voltage characteristic for spark arc pulse generator.

to this circuit include replacing the rectifier by a thyratron or ignitron and controlling its conduction angle by conventional methods. This reduces the excessive heating but the low frequency gives a poor relation between surface finish and metal removal rate.

A similar circuit connected to a high-frequency alternator of frequency up to 10 kc/s is capable of producing very good surfaces at relatively high metal removal rates. The disadvantage of this arrangement lies in the cost. Since the output is half-wave rectified, the machine is working at low efficiency and the cost of a high-frequency machine is high.

It has been found by experiment that, for medium metal removal rates of up to 20 g/min, a pulse repetition frequency of 400/s and pulse duration shown in Fig. 14 gives usable surface finish for "roughing" work without risk of damage by over-heating. A higher frequency will, of course, achieve a finer surface finish but the cost of the generator increases rapidly with frequency.

5.3. *Rotary Pulse Generators*

To obtain the voltage characteristic shown in Fig. 14 using equipment which is cheap and robust, specially-designed rotary pulse generators are used.[7, 8]

One such machine uses the same principle as a d.c. generator but the armature has only the same number of slots as there are poles. The pole-pieces

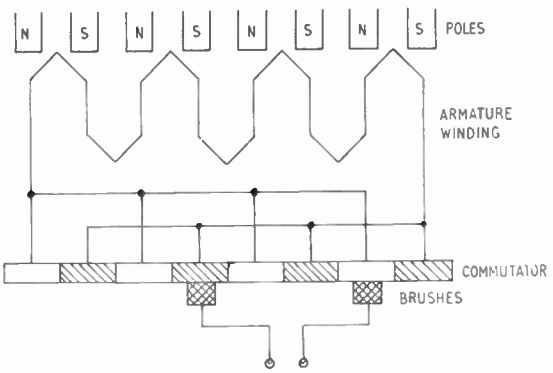


Fig. 15. Schematic arrangement of rotary pulse generator.

are much narrower than in a conventional machine, thus giving a pulsating voltage output, and the commutator has only the same number of segments as there are poles. Figure 15 shows the commutator arrangement and it can be seen that alternate segments of the commutator are linked together. The brushes are arranged to collect current from adjacent segments.

The typical machine shown in Fig. 16 is driven by a 15 h.p. motor and is capable of removing up to 20 grammes of steel per minute. This order of metal removal rate has made possible the economic production of large dies such as are used in the drop-forging industry.

6. Electrode Servo Control

The electric motor which drives the electrode is required to hold the minimum gap between electrode and work-piece at a constant value, which may be as small as one thousandth of an inch, as metal is eroded. The task of the motor is complicated by the

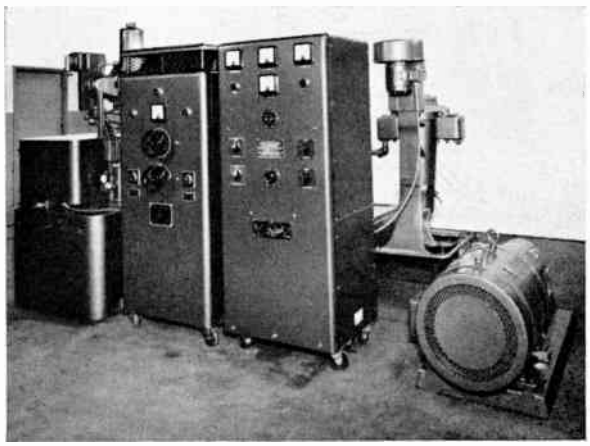


Fig. 16. A high-power spark machining installation showing rotary pulse generator and centrifugal filter.

occasional short-circuiting of the gap by metal particles and carbon produced by sparking. The electrode must be raised rapidly to break this short-circuit condition and to allow the particles to be cleared away by the fluid. The electrode must then be lowered again and machining continued with a minimum waste of time.

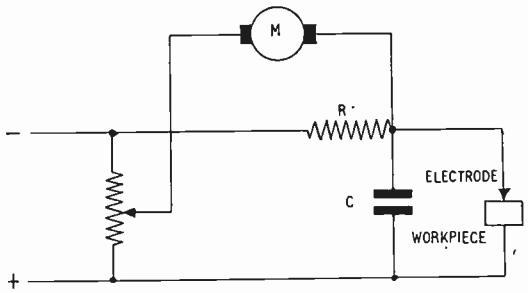


Fig. 17. A simple circuit for controlling the electrode servo-motor.

Since the mean voltage between the electrode and the work-piece and the voltage across the series resistor are a measure of the condition and size of the gap, this is used as a signal to control the movement of the motor.

A simple control circuit arrangement is shown in Fig. 17.

A permanent-magnet motor is driven by the unbalance of the bridge circuit formed by R, the gap and a potentiometer connected across the d.c. supply.

It can be seen that this arrangement is a closed loop servo system with sensitivity and stability determined by the loop gain, time delays and inertias of the components. Care is necessary to ensure that sufficient damping is present to prevent the electrode "overshooting" or oscillating. Electronic and magnetic amplifiers are also used to drive either d.c. or a.c. motors.

Another important type of electrode control employs hydraulic power, which is used particularly for heavy electrodes. The electrode is coupled to a hydraulic cylinder which is operated by oil-flow via an electrically-operated valve. This valve is actuated by a similar signal to the motor in other systems. Other configurations include an hydraulic motor driven via an electro-hydraulic valve.

7. Electrodes

7.1. *Electrode Wear*

The exact nature of the spark discharge is not agreed by the many investigators working on the phenomenon.[9–12] Certain metallurgical aspects of spark-machined surfaces have received considerable attention but the work done so far is not conclusive.[13, 14]

It can be stated, however, that more heat is dissipated at the anode than at the cathode during a discharge. By making the work-piece the anode, it is generally ensured that a majority of material is eroded from it, but nevertheless, a certain amount of material is also eroded from the electrode. Naturally, this is a disadvantage since the shape of the electrode is altered as machining progresses and the original form is not accurately reproduced in the work-piece by a single electrode.

The amount of erosion suffered by the electrode compared with that by the work-piece is referred to as "wear-ratio" and depends upon the physical and chemical properties of both the electrode and work-piece materials.

7.2. *Materials*

Four main factors determine the suitability of a material for use as an electrode. These are:

(i) The maximum possible machining rate.
(ii) The "wear-ratio".
(iii) The ease with which it can be fashioned to the shape required, and
(iv) Its cost.

Obviously it is not possible to put these factors into a preferred order as different work demands a different economic approach.

From purely technical considerations it is possible to specify a material such as silver-tungsten alloy as the most efficient electrode having a high metal removal rate and a very low "wear-ratio" but the cost of such an electrode under most conditions is of course prohibitive. Generally speaking, by using a sufficient number of electrodes of material having a high wear ratio, it is possible to produce the same accuracy of machining as with a single electrode of material with a low wear ratio. With the technical and economic conditions in mind, therefore, four classes of material are most commonly used for electrodes.

The first of these contains high-conductivity materials such as copper and brass which have a fairly good performance against the requirements. They are reasonably cheap, easily fashioned from solid, have good machining rates and fairly low wear-ratios.

The second class embraces zinc-based alloys which are most useful where a master die is available to enable a large quantity of electrodes to be cast cheaply. On high-power machining, such as with rotary generators, these alloys have a slightly lower machining rate than brass or copper but the wear-ratio is good. For fine-finishing, however, the wear-ratio is considerably worse. Recent developments, however, have led to the production of zinc alloys with much improved performance.

A third class includes refractory materials such as tungsten, tungsten carbide and graphite. These materials are difficult to handle but have extremely low wear-ratios, which makes them suitable for special applications.

The fourth and final group is that of mixtures of classes one and three containing materials such as copper-tungsten and copper-graphite. These combine the erosion resistance of tungsten and graphite with the good machining rates of copper. Their cost is rather high and copper-graphite is not very easily fashioned. Copper-tungsten, however, is often used where very accurate work is required with a single electrode.

7.3. *Manufacture*

Electrodes may be manufactured in a number of different ways. The most obvious method is by conventional machining techniques using drills, lathes, etc. This method is laborious and to reduce the difficulty of making complicated electrodes in copper, a technique has been developed for spraying molten copper into a plaster or zinc pattern prepared from a model. In this way the copper electrode is built up having the accurate shape of the original model. A second technique applicable to the zinc-alloy electrodes, is casting into a metal mould. After preparation of the mould, this is a very cheap production method and is commonly used by die manufacturers.

8. The Dielectric Fluid

As mentioned earlier, a fluid is used to fill the gap between electrode and work-piece which acts as a dielectric until it is ruptured by the spark. When this occurs an ionized path is created in the fluid which must "heal" rapidly to allow the next discharge to occur at the shortest gap between electrode and work-piece.

The fluid must also quench the molten and vaporized metal produced by the spark and remove it from the gap. When electrodes having large areas are being used, quenched material and colloidal carbon produced by sparking in the fluid cannot easily disperse and tends to build up, forming a short-circuit across the gap. To prevent this happening it is usually necessary to pump the fluid through the gap under pressure providing a "flushing" action. This is accomplished where possible by drilling small holes in the electrode through which the fluid is pumped.

This flushing action also has an effect on the metal removal rate and wear-ratio of the electrode. Generally, good flushing improves the metal removal rate, probably by reducing the number of short-circuits and thus raising the efficiency of machining. At the same time, the reduction in the number of short-circuits also lessens electrode erosion and a smaller wear-ratio is obtained.

It is essential that the fluid pumped through the gap should be free from metal particles and carbon. It is normal practice, therefore, to filter the fluid before it is pumped through the electrode. A very simple filter consisting of paper elements, sintered metal or ceramic cylinders through which the fluid is passed is usually employed in the smaller spark-machines. In the case of high-power machines, however, the volume of metal particles and carbon produced make a larger capacity filter necessary. Quite often a centrifuge type is employed as seen in Fig. 16.

9. Conclusion

The relatively simple electrical requirements of spark erosion machining have been met, up to now, by circuits such as the relaxation type which have

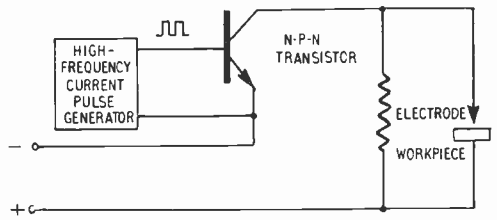


Fig. 18. Transistor switching circuit for high-frequency spark machining.

combined good reliability with a rather low efficiency. High-power low frequency spark-initiated arc circuits have increased the speed of metal removal but machines are urgently needed which produce a good surface finish at a high metal removal rate. This requirement can only be met by high-power, high-frequency circuits and attempts have been made to use thermionic valves. However, since ruggedness and reliability are important factors in machine-shop equipment, these attempts have not been wholly successful.

The recent introduction of rugged high-power semi-conductors capable of high frequencies has given the spark-machine designer fresh fields to explore. By using the "switching" characteristic of a transistor it is possible to control quite large powers in a spark-gap load. A suitable schematic arrangement is shown in Fig. 18. Such a circuit is capable of providing very high frequency voltage pulses to the gap, thus producing the high-power, high-frequency operation which is so desirable in spark-machining. Such circuits[15] are now being introduced in commercial equipment which will largely revolutionize the performance hitherto obtained with the relaxation circuit.

10. References

1. B. R. Lazarenko and N. I. Lazarenko, *Stanki I., Instrument (U.S.S.R.)*, **17**, No. 12, pp. 8–11, 1946 and **18**, No. 2, pp. 4–8, 1947.
2. D. W. Rudorff, Brit. Pat. No. 637,872, 1950.
3. D. W. Rudorff, "Machining of hard metals by electrical methods", *Proc. Instn Mech. Engrs*, **171**, No. 14, pp. 495–511, 1957.
4. A. V. Nosov and D. V. Bykov, "Working Metals by Electro-sparking". D.S.I.R. translation published by H.M.S.O., 1956.
5. E. M. Williams, "Theory of electric spark machining", *Electrical Engineering (N.Y.)*, **71**, 257–60, March 1952.
6. E. M. Williams and J. B. Woodford, "Electronic considerations in theory and design of electric spark machines" *Trans. Inst. Radio Engrs*, March 1955.
7. A. L. Livshits and I. S. Rogachev, "Generators for electric impulse-machining of metals", *Elektrichestvo*, 77, No. 3, pp. 19–23, March 1957. (In Russian.)
8. Yu. V. Mordvinov, "Noncommutator pulse generators for feeding electro-spark machines", *Vestnik Elektro-Promyshlennost*, **28**, No. 1, January 1957.
9. K. A. Macfadyen, "Some researches into the electrical conduction and breakdown of liquid dielectrics", *Brit. J. Appl. Phys.*, **6**, pp. 1–7, No. 1, January 1955.
10. A. S. Zingerman, "Role of Joules-Lenz heat in electrical erosion of metals", *Zh. Technicheskoi Fiziki (U.S.S.R.)*, **25**, No. 11, October 1955.
11. B. N. Zolotykh, "The physical characteristics of electro spark machining of metals", *Schriftenreihe Verlag Technik* **175**, 1955.
12. B. Salvage, "The dielectric breakdown of some simple organic liquids", *Proc. Instn Elect. Engrs*, **98**, Pt IV, pp. 15–22, 1951. (I.E.E. Monograph No. 2, May 1951.)
13. O. Ruediger, "Structural influence and quality of surface finish from e.d.m.", *Jahrbuch der Oberflaechentech*, **16**, 1960.
14. H. Opitz, "Finish, accuracy and structure of work pieces machined electrically", *Microtecnic*, **9**, No. 1, 1955.
15. Sparcatron Ltd., British Pat. No. 812,012 etc.

Manuscript first received by the Institution on 17th November 1960 and in final form on 13th June 1961 (Paper No. 680).

Amplifier Gain and Stability

U.D.C. 621.375.1

By L. G. CRIPPS, B.A.† AND J. A. G. SLATTER, B.Sc.†

We have been examining the gain and effects of internal feedback ("stability") of amplifiers in terms of four-pole parameters. Our analysis utilizes a quantity $1/\sqrt{(G_F G_R)}$ where G_F is the forward and G_R the reverse transducer gain of the system. Boothroyd[1] and Venkateswaran[2] have used a stability factor S, for the conjugate matched case, defined in terms of other parameters, which is equal to this quantity $1/\sqrt{(G_F G_R)}$. Venkateswaran[2] proves a number of theorems regarding S, and implicit in his analysis is the result that, for conjugate matching, a value of $S>1$ indicates that the source and load conductances may be removed without oscillation occurring. In addition, it may be shown that a small change of load admittance from y_L to $y_L + \Delta y_L$ implies a change of input admittance given by $|\Delta y_{IN}|/g_{11} = (1/S)|\Delta y_L|/g_{22}$. Thus S is a meaningful quantity, and further, it is directly measurable.

Two important new results can easily be deduced. Utilizing the definition $S = 1/\sqrt{(G_F G_R)}$ and expressions for G_F and G_R for *arbitrary* terminations, it may be shown that

$$\sqrt{(G_F/G_R)} = SG_F = |y_{21}/y_{12}| = |z_{21}/z_{12}| = \text{etc.}$$

Thus both the ratio G_F/G_R and the quantity SG_F are device parameters *completely independent of source and load terminations*. (Venkateswaran[2] has already deduced the above equations for the matched case only.) Hence, firstly, gain × "stability factor" product is an invariant figure of merit. Secondly, since the terminations do not affect the SG_F product, we can choose convenient ones, e.g. 50Ω or 75Ω resistive source and load, to measure G_F/G_R, yielding SG_F. None of the problems associated with conjugate matching arise, and the measurement is extremely simple.

Our analysis will form the basis of a future publication, which will include a discussion of the use of S as a circuit design parameter and an examination of its significance, particularly for terminations other than those giving conjugate matching.

We are grateful to a number of our colleagues for several helpful discussions.

† Mullard Research Laboratories, Redhill, Surrey.

References

1. A. R. Boothroyd, "The transistor as an active two-port network", *Scientia Electrica*, 7, p. 3, 1961.
2. S. Venkateswaran, "An invariant stability factor and its physical significance", Institution of Electrical Engineers Monograph No. 468E, September 1961.

Manuscript received by the Institution on 14th October 1961 (Contribution No. 40).

# of current interest . . .

Contract for "Atlas" Computer

An *Atlas* electronic digital computer has been ordered for the National Institute for Research in Nuclear Science. The computer, together with the necessary buildings, will cost about £3½ million. It will be installed at the Institute's Rutherford High Energy Laboratory, Harwell, for common use by the Universities, the United Kingdom Atomic Energy Authority, Government Departments and the N.I.R.N.S. itself. It should be ready for use early in 1964. As in the case of the Institute's other facilities, university requirements for use of the computer will be judged on their scientific merits, and when accepted will be met without charge to the Universities.

The *Atlas* computer, made by Messrs. Ferranti, has been developed in co-operation with scientists at the University of Manchester, where the prototype is now being assembled. The United Kingdom Atomic Energy Authority and certain universities have substantial requirements for time on such a machine, but the *Atlas* can cope with so much work that it was decided to provide one machine for common use in the first instance, and the Institute were invited to manage it.

The Atomic Energy Authority is the largest user of computers in this country; large amounts of computation have gone, for example, into the design of nuclear power reactors and the study of their behaviour. The National Institute itself will require a very large amount of computation in the work of interpreting the experiments made with the 7000 MeV proton synchrotron *Nimrod* which is under construction at the Rutherford Laboratory. It will be used also in theoretical studies of future high energy accelerators, involving calculations of the motion of charged particles in electric and magnetic fields which vary both in space and in time.

Of the Government scientific establishments, the Meteorological Office is likely to be one of the main users. The idea of predicting the weather by computation is about 50 years old but the scale of the task made this quite impractical until the electronic computer was developed; the mathematical problem is the integration of the partial differential equations for the pressure and temperature distribution and the motion of the atmosphere.

Generation of High Magnetic Fields

A continuous magnetic field of 126 000 gauss, believed to be the highest ever produced, has been generated at the National Magnet Laboratory of the Massachusetts Institute of Technology. The field was achieved using a solenoid magnet about the size of a grapefruit, invented and patented by Dr. Henry H. Kolm of the Laboratory.

Such a high continuous, magnetic field is difficult to achieve because of the problem of getting enough electric power into a small volume and removing the heat that it produces. The magnet designed by Dr. Kolm consists chiefly of a ribbon of thin copper, six inches wide at one end, tapering to an inch and a half at the other end and 135 ft long. This ribbon has over three thousand square channels cut into it, and when it is wound into a roll between insulating material the channels align like spokes in a wheel and water can be forced through them for cooling purposes.

In the centre of the coiled copper is a tube measuring only one inch in diameter by two inches long. The magnetic field is produced in this tube by an electric current of 10 000 amperes, and 320 gallons of water per minute are pumped through the slots to remove the heat produced in the consumption of 1·88 MW of electricity. This is the present limit of the power supply at M.I.T. but the magnet may be capable of producing still larger magnetic fields when a source of 8 MW is installed.

Future Screw Thread Practice and the Electrical Industry

The British Electrical and Allied Manufacturers' Association has had the problem of future screw thread practice under consideration for some years and following the decision of the International Standards Organization to issue recommendations for both unified inch and metric systems a detailed inquiry was made amongst individual member firms.

The BEAMA Council has now issued a statement summarizing the results of this enquiry. In this statement the BEAMA Council points out that the inquiry was directed towards long term consideration of all the many relevant factors which should influence the industry's future screw thread policy. These considerations included the importance to the electrical industry and its customers of having a common screw thread policy, the significance to the industry's export trade, not only in the Commonwealth and Europe but throughout the world, and generally an attempt to foresee long term trends.

The inquiry has revealed that a substantial majority opinion amongst the BEAMA members is that it is desirable to maintain a common normal production practice within the electrical and allied industries and that when the time comes to make a change it should be to the I.S.O. metric thread. In the light of this concensus of opinions of its members the BEAMA intends to initiate discussions with the major customer, trade, and Government interests with which they are concerned so as to seek to work out an acceptable long-term plan for giving effect to this future screw thread policy.

U.D.C. 523.16: 629.19

Determining Local Concentrations of Charged Particles in the Ionosphere and Interplanetary Space

Methods used on Soviet Rockets and Earth Satellites and some results obtained

By

Dr. K. I. GRINGAUZ†

Presented at the Convention on "Radio Techniques and Space Research" in Oxford on 5th–8th July 1961.

Summary: Electron concentration determinations in the ionosphere by radio wave dispersion measurements using transmissions from rockets are discussed with reference to the results obtained by various workers. Charged particle trap experiments carried on satellite and space probes are also described, and some conclusions are reached regarding the distribution of charged particles in the sunlit part of the Earth's gas envelope.

1. Introduction

Several descriptions of Soviet rocket and satellite experiments aimed at studying local ionospheric concentration have been published during the last three years.[1–9] Therefore this paper will mention only some of the essential features of the experiments and some of results obtained; full and systematic descriptions of methods and experiments which would be impossible in such a report are omitted.

When we speak about measurements of local concentrations of charged particles in the ionosphere, we mean those measurements each of which makes it possible to determine the concentration at a definite altitude, in definite time, and without the use of observations from ionospheric stations or suppositions on the height distribution of charged particle concentration. Since experiments aimed at ionospheric studies by means of radio waves emitted from Earth satellites do not satisfy these conditions they will not be considered in this paper.

2. Electron Concentration Measurement Experiments by Means of Radio Waves Emitted from Rockets

Coherent unmodulated radio waves with frequencies $f_1 = 24$ Mc/s, $f_2 = 48$ Mc/s and $f_3 = 144$ Mc/s are radiated from vertically launched geophysical rockets of the U.S.S.R. Academy of Sciences. To determine the electron concentrations at various altitudes, radio wave dispersion measurements by the phase method are carried out at several points on the Earth (with the use of frequency combinations 48–144 Mc/s and 24–144 Mc/s). Faraday rotations of polarization planes of each radio wave received are also recorded.

Figure 1 shows a geophysical rocket in flight. This rocket reaches a height of about 470 km. The important feature of the rocket is the extreme closeness of

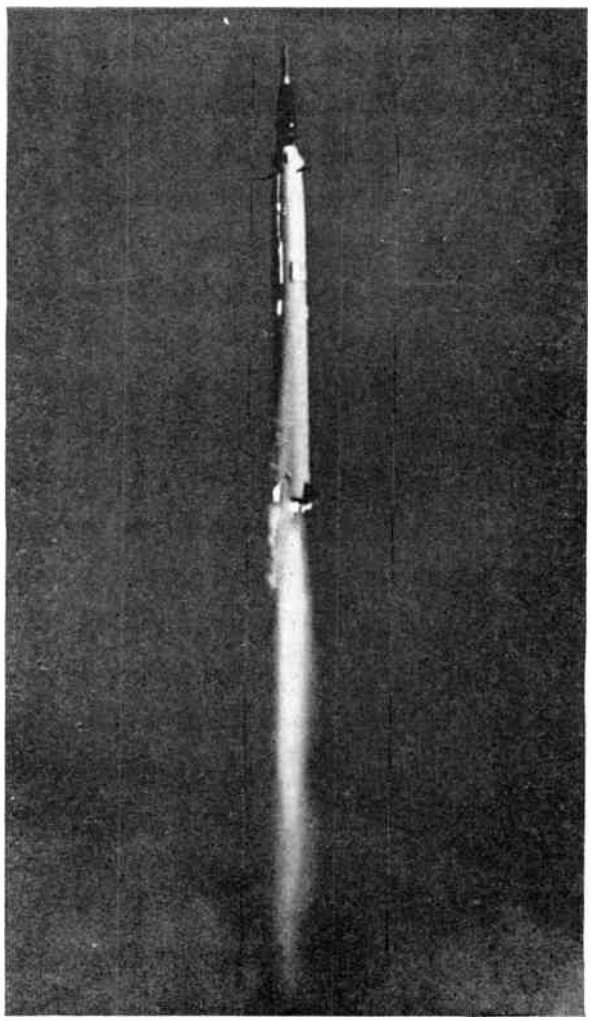


Fig. 1. Geophysical rocket in flight.

† Radio Technical Institute of the U.S.S.R. Academy of Sciences, Moscow.

its trajectory to the vertical, which makes it possible to ignore its horizontal velocity while processing experimental results. The second important feature of the rocket is its complete stabilization after power cut-off with respect to three mutually perpendicular earth-bound axes. Thus the only cause of rotation of the plane of polarization of radio waves received on the Earth is the Faraday effect. Therefore, the height distribution of electron concentration can be obtained from observing the rotation of the polarization plane at one frequency.

Results of determining electron concentration height-distribution obtained by the dispersion method and by the Faraday rotation method are in good agreement.

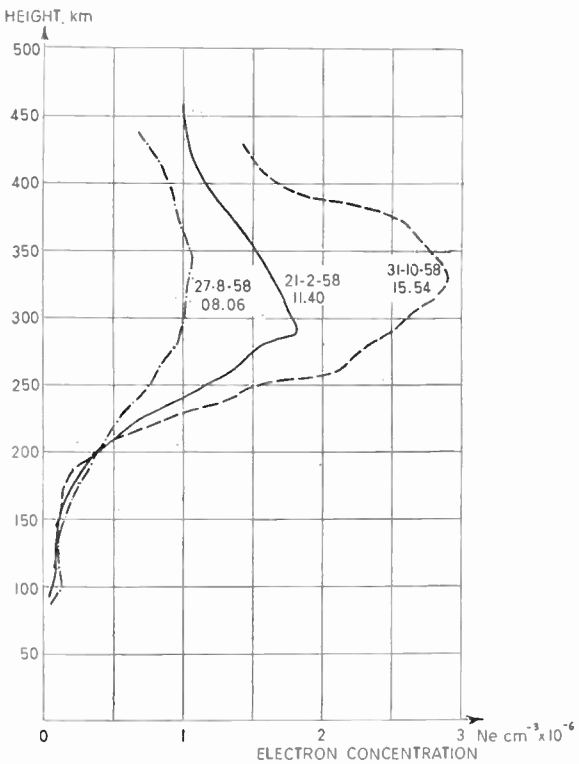


Fig. 2. Variation of electron concentration with height.

The variability of the ionosphere (including its outer part lying above the maximum of ionization of the F-layer) is to some degree characterized by the three electron concentration height distributions shown in Fig. 2. All three distributions were obtained in 1958 by means of rockets of the type shown earlier, above the same point (in middle latitudes of the European part of the U.S.S.R.) with the same instrumentation. The first curve refers to the launching on 21st February (the beginning of the launching at 11.40), the second curve refers to the launching on 27th August (beginning at 08·06), the third one shows the launching on 31st October (beginning at 15.54). From the curves it is seen that values of electron concentration vertical gradients above the F-region maximum over the same point can vary considerably in time.

It is well known that during electron concentration measurements made from vertically launched rockets by the phase dispersion method, time variations of the total electron content in a column under the rocket are starting to exert a growing influence as the rocket approaches the apex of its trajectory.

It is easy to show that the change of the phase difference of two coherent radio waves received on the ground during the time Δt from a rocket flying in the ionosphere and reduced to one frequency can be presented as

$$\Delta\phi = \Delta\phi_{\text{loc.}} + \Delta\phi_{\text{int.}} \qquad \ldots\ldots(1)$$

As the frequencies considered are sufficiently high, then we have

$$\Delta\phi_{\text{loc.}} = K\overline{n_e v_h}\Delta t$$

$$\Delta\phi_{\text{int.}} = K\frac{\partial}{\partial t}N_e\Delta t$$

K is a constant which depends on the instrumentation parameters and frequencies received.

$\overline{n_e}$ is the mean electron concentration in the height range Δh passed through by the rocket in the time Δt.

$\overline{v_h}$ is the rocket's mean vertical speed in this range.

$N_e = \int_0^h n_e\,dh$ is the total electron content in a vertical column from the observer to the beginning of the interval Δh.

The term $\Delta\phi_{\text{loc.}}$ is related to the height increment Δh for the time Δt and depends on electron concentration $\overline{n_e}$ in the interval Δh.

The term $\Delta\phi_{\text{int.}}$ is related to the changes in the N_e value for the time Δt. Such N_e changes take place as a result of the regular diurnal variations of the state of the ionosphere and as a result of motions of non-homogeneous formations in the ionosphere.

With sufficiently large values of the $\overline{n_e v_h}$ product $\Delta\phi_{\text{loc.}} \gg \Delta\phi_{\text{int.}}$ and the second term in (1) can be ignored. It is evident, however, that with very small $\overline{n_e v_h}$ values (which in particular take place near the rocket trajectory apex where v_h passes through zero) the measured value of $\Delta\phi$ depends only on $\Delta\phi_{\text{int.}}$, i.e. the value of $\partial N_e/\partial t$, the rate of change of the total electron content in a vertical column below the rocket.

Values of $\partial N_e/\partial t$ were measured near upper points of the trajectories of the rockets launched on 21st February and 27th August 1958. They were determined as $\frac{1}{K}\left(\frac{\Delta\phi}{\Delta t}\right)$ with $v_h \simeq 0$ (which corresponded to

$h \simeq 470$ km in one case, and to 450 km in the other) and $\Delta t = 0{\cdot}3$ seconds. In both cases values $\partial N_e/\partial t$ were close to 5×10^9 cm^{-2} s^{-1}.

One should bear in mind that these measurements refer to the period of maximum solar activity and that lower $\partial N_e/\partial t$ values are possible for vertical columns at the same height.

From the above results of measurements it is evident that in general it is impossible to determine local electron concentration in interplanetary space near the space vehicle by means of terrestrial observations of radio waves emitted from the rocket, since the influence on radio waves of processes occurring in the Earth's ionosphere will be greater than the effect made by the change of the optical length of path of radio waves which is due to the interplanetary ionized gas near the rocket. In fact, if one takes the electron concentration in interplanetary gas as equal to $n_e \simeq 10^2$ cm^{-3} (actually n_e we think may be much lower) and if the rocket velocity $v_r \sim 10$ km/s, then, nevertheless, n_e. $v_r \simeq 10^8$cm^{-2}s^{-1} and $\partial N_e/\partial t$ exceeds this value more than by one order of magnitude.

It is of interest, however, to consider the possibility of determining the total electron concentration N_{eR} by radio methods on the path from the Earth to the space vehicle.

Knowing N_{eR} and the distance to the rocket it is possible to determine $\overline{n_e}$—the mean electron concentration in interplanetary space, if one is confident that N_{eR} considerably exceeds N_{eE}, which is the total electron content of the part of the path of the radio waves in the Earth's ionosphere. Evaluation of this possibility was made in 1959 by Kelso[10] who considered the case of $\overline{n_e}$ determination by measuring the difference of group paths of two pulses of radio waves with different frequencies, simultaneously emitted from a space vehicle. Taking $\overline{n_e} \simeq 10^2$ per cubic centimetre the author concluded that $N_{eR} \gg N_{eE}$ with $f_1 = 100$ Mc/s and $f_2 = 400$ Mc/s at a distance of 10^8 km and that the difference between the group paths of radio waves with the above frequencies is about 40 km. This can be recorded with confidence.

The drawback of the above computation is that it assumes an excessively high value of electron concentration in interplanetary space n_e (~ 100 cm^{-3}). At the same time, the method of measurement under discussion (i.e. reception of the signals on the Earth as two pulses and determination of delay of one in respect to the other) requires a wide-band receiver. It is difficult with such a system to ensure reception of the signals transmitted with powers acceptable for rocket-borne transmitters over the necessary distances of the order of 100 million km.

The experimental variation considered by Soviet radiophysicists E. E. Mityakova, N. A. Mityakov and V. O. Rappoport from the Gorky University[11] is more to be recommended. They proposed to conduct the experiment in the following way. Radio waves are transmitted from a space vehicle at three frequencies f_1, f_2 and f_3, in-phase and sinusoidally amplitude modulated with a frequency Ω. Frequencies f_1 and f_2 are close to each other in magnitude, but frequency f_3 essentially differs from them. By observing the rotation of planes of polarization of the radio waves with frequencies f_1 and f_2 the total electron content in the Earth's ionosphere N_{eE} should be determined as it was made in Evans' well known experiments at the Jodrell Bank Observatory when signals reflected from the Moon were used.[12] Measurements of the phase shift of the modulation envelopes of radio waves with frequencies f_1 and f_3 (or f_2 and f_3) will make it possible to determine the difference between the group paths of these waves and consequently the N_{eR} value.

The mean electron concentration in interplanetary space can be obtained as $(N_{eR}-N_{eE})/R$, where R is the distance from a space vehicle to the Earth.

The possibility of taking into account and excluding the total electron content in the Earth's ionosphere as well as the possibility of using receivers with much narrower bands (at the expense of sinusoidal modulation) are considerable advantages of this variant of the experiment. Its sensitivity to electron concentration is much higher than in Kelso's version.

However, it is likely that the determination of the mean electron concentration of interplanetary ionized gas by means of radio waves emitted from space vehicles can become very difficult, even in the case of the most sensitive variants of such an experiment. These apprehensions are based on the fact that experiments with traps of charged particles on board space probes give reason to believe that the concentration of stationary ionized gas in interplanetary space is very low.

Let us return to the lower part of the ionosphere.

Figure 3 shows height distribution of electron concentration obtained by the phase dispersion method during the vertical launching of the U.S.S.R. Academy of Sciences' geophysical rocket which began at 06.43 on 15th July 1960 in the middle zone of the European part of the U.S.S.R. (local time).[13] The maximum electron concentration at a height $h = 105$ km is of particular interest. Earlier we repeatedly observed small maxima at these altitudes. They were also observed in the American experiments conducted by Seddon and Jackson.[14]

The same geophysical rocket launched on 15th June, 1961, carried the radio-frequency ion mass spectrometer of the Bennett type installed in the capsule which was separated from the rocket to make measurements in an ionospheric region not contaminated by the rocket.

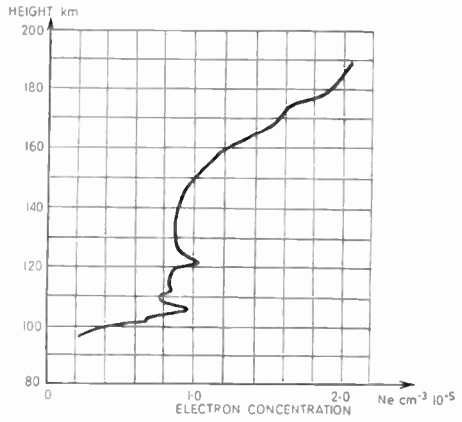


Fig. 3. Variation of electron concentration with height.

The analysis of measurements by the ion mass spectrometer performed by V. G. Istomin makes it possible to give an interesting explanation of the above-mentioned electron concentration maximum.[15] In the region of this maximum there were revealed peaks in positive ion mass spectra corresponding to mass numbers 24 and 26. Magnesium isotopes correspond to these mass numbers.

Besides there were revealed peaks at the same altitudes with mass number 40 which corresponds to calcium ions (Ca^+). The magnesium ion concentration was estimated as n^+ Mg $= 1 \times 36 \times 10^4 cm^{-3}$. This estimation was made by the use of maximum values of the relative intensity of magnesium ion peaks with respect to the total intensity of all ion components recorded by the mass spectrometer. The values of electron concentrations at these altitudes determined by the dispersion radio-interferometer were also used. Calcium ion concentration was estimated as $n^+Ca \simeq 540\ cm^{-3}$.

Measurements were made during the activity of the daytime meteoric streams of Arietides and Perseides. Comparison of values of the height of the layer, in which Ca^+ and Mg^+ ions are revealed, with data on the height of the layer of CaII line of twilight airglow as well as comparison of concentrations of these ions with relative distribution of magnesium and calcium in stone meteorites, have led Istomin to the conclusion that the Mg^+ and Ca^+ ions which were recorded are the product of the interaction of the atmospheric molecules with swiftly flying Mg and Ca atoms evaporated from meteors. If this supposition is true, then the rather frequently observed small electron concentration maximum in the region of the height $h \sim 105$ km has a meteoric origin.

It should be pointed out that though these mass-spectrometric observations were made in morning hours, they indirectly confirm considerations in favour of the meteor origin of the nocturnal E-layer expressed by Nicolet in 1955.[16]

3. Experiments made with Charged Particle Traps on Soviet Satellites and Space Probes

Langmuir probe shielding by means of a grid to which an electrical potential is supplied was successfully used for the first time by R. L. F. Boyd in a plasma of a gas discharge and was described by him in 1950.[17] Such a "shielded" probe moving in a plasma with a speed exceeding thermal ion velocities acquires a number of new properties.[2]

In 1958 we used such spherical shielded probes called ion traps on board *Sputnik III* to study the concentration of positive ions in the ionosphere. Later all Soviet space probes, including the Venus probe launched on 12th February, 1961, carried charged particle traps in which a third electrode—an inner grid—was introduced to suppress photo-emission from collector. This photo-emission is caused by solar ultraviolet emission and limits the instrument sensitivity to the charged particle streams getting into traps from the surrounding medium. Figure 4 shows the placing of charged particle traps on a capsule with

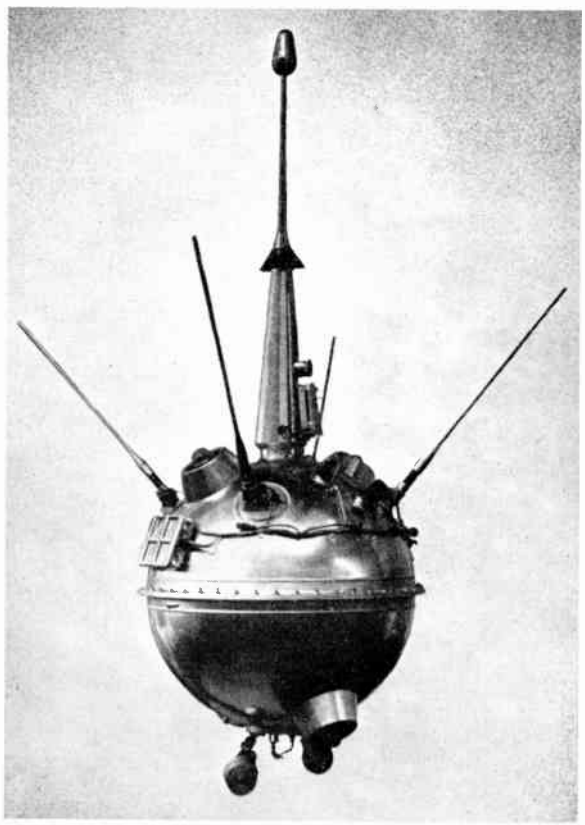


Fig. 4. *Lunik II.*

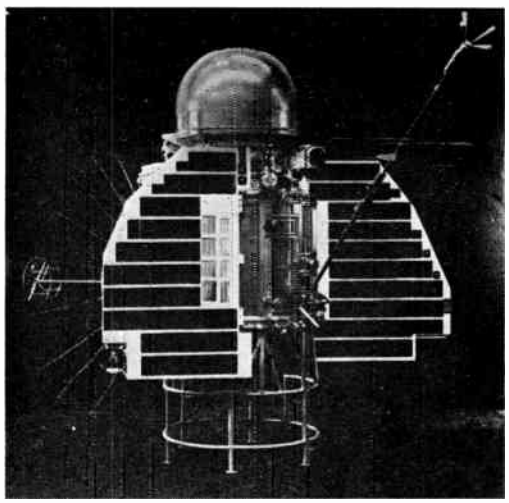


Fig. 5. Venus probe automatic interplanetary station.

Lunik II scientific instrumentation. Figure 5 shows the placing of traps on the Venus probe automatic interplanetary station.

Lack of time prevents the description of the peculiarities of the technique of measurements by means of charged particle traps on *Sputnik III* and each of space probes. There is also no opportunity of discussing theoretical considerations on the basis of which recorded currents in the trap collector circuits were recalculated into charged particle concentrations. These data were published in Soviet scientific periodicals[6, 7, 8] and were reported in April this year at the Second International Space Science Symposium in Florence.[18]

Only some conclusions made on the basis of the results of charged particle trap experiments and referring to the Earth's ionized gas envelope and the ionized gas in interplanetary space will be discussed.

More than 10 000 ion volt-ampere characteristics ("retardation curves") corresponding to the altitude ranges up to 1000 km were obtained by *Sputnik III* in the period from 15th May to 2nd June, 1958. It was established that at the altitudes of *Sputnik III* flight, the electron concentration is equal to the ion concentration. A number of positive ion concentration measurements were made along the trajectories of *Luniks I, II* and *III* launched in 1959.

A combination of the results of these measurements together with the results of electron concentration measurements made in 1958 by means of vertically launched rockets (which were cited at the beginning of the paper) makes it possible to compose a tentative charged particle distribution in the sunlit part of the Earth's gas envelope in a period close to maximum solar activity (1958–1959). Such distribution is shown in Fig. 6. Dotted lines indicate the region in which there are no data. Measurements of ion mass spectra have revealed that at altitudes up to 1000 km the ionosphere mainly consists of O^+ ions. The height variation of the charged particle concentration makes one believe that at altitudes from 1000 km to 2000 km a transition takes place from the oxygen ionosphere to the hydrogen one which mainly consists of hydrogen ions—protons with concentration of the order of 10^3 cm^{-3}. At altitudes of more than $\sim$ 15 000 km the negative vertical gradients of the charged particle concentration increase and at altitudes about 20 000 km the concentration reaches the value of $\sim$ 100 particles per cm^3. It continues to decrease with the increase of height.

In the graph of Fig. 6 the dates on which the experimental dots were obtained are indicated.

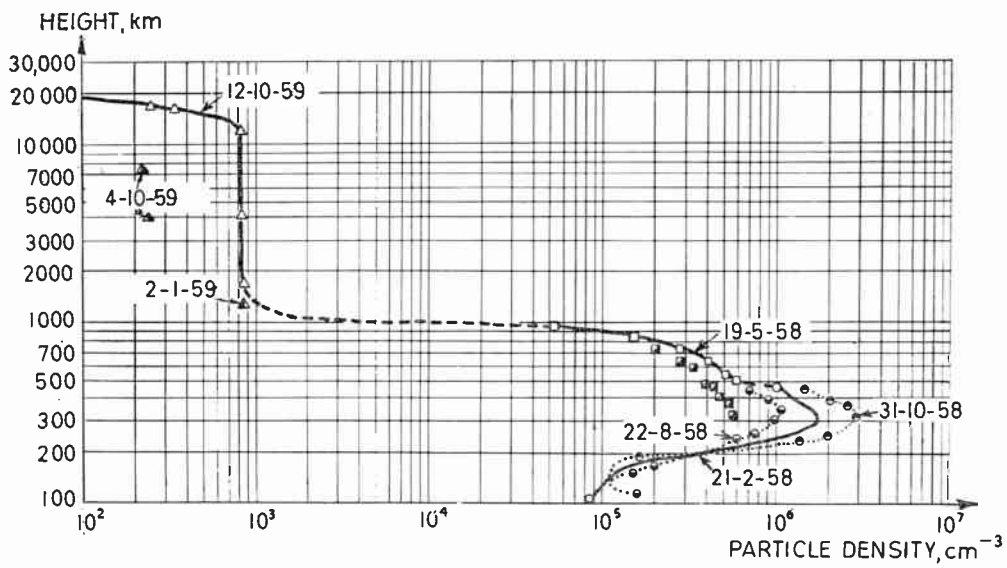


Fig. 6. Variation of the density of charged particles with height.

Data obtained with *Lunik III* (launched on 3rd October, 1959) at distances greater than 200 000 km from the Earth show that the concentration of charged particles with low (thermal) energies of the presumably existent stationary interplanetary gas is very low (apparently not more than 1·5 particles per cm^3).

On 13th September, 1959, during *Lunik II* flight solar corpuscular streams (streams of ionized gas ejected by the Sun) were for the first time recorded in interplanetary space beyond the Earth's magnetic field by means of charged particle traps.

Currents created by the positive ions of the corpuscular streams were repeatedly observed in traps during *Lunik III* flight. There was a good correlation between the presence of such currents and *K*-indexes which characterize geomagnetic disturbances.

The maximum solar corpuscular stream observed up to present time with charged particle traps was recorded on 17th February during the radio contact with the automatic interplanetary station of the Soviet Venus probe at a distance of $\sim$ 1 900 000 km from the Earth.

The density of this stream was $\sim 10^9$ $cm^{-2}s^{-1}$. A magnetic storm was observed at this time on the Earth. The stream velocity determined from the time of delay of the commencement of the magnetic storm in respect with a flocculus passing through the central meridian of the solar disc amounted to $\sim$ 400 km/s. This means that electron concentration in the stream was approximately 25 cm^{-3}.

One can believe that experiments aimed at measuring the mean electron concentration in interplanetary space by means of radio signals transmitted from space vehicles will make it possible to measure the mean electron concentration in sufficiently intense corpuscular streams.

The minimum stream of charged particles from the surrounding medium which can be recorded in three-electrode traps is limited by the value of the current of the photo-electrons incident from the inner grid to the collector while the trap is lit by the Sun. This limiting value of the recorded stream can be considerably lowered if the charged particles getting into the trap from the outside are modulated with a certain frequency (so that the photo-electron current remains unmodulated) and records are made at the modulation frequency.

At present instrumentation has been prepared to measure by charged particle traps, not only the density of the stream of solar corpuscles, but also their energy spectrum.

There are no doubts that charged particle traps will be one of the effective means of exploration of the Earth's ionosphere, the ionized gas in interplanetary space and ionospheres of other planets.

4. References

1. K. I. Gringauz, *Doklady Akad. Nauk S.S.S.R.*, **120**, p. 1234, 1958.
2. K. I. Gringauz and M. H. Zelikman, *Uspekhi Fiz. Nauk*, **63**, issue 1b, p. 239, 1957.
3. V. I. Krassovsky, *Proc. Inst. Radio Engrs*, **47**, p. 289, 1959.
4. K. I. Gringauz and V. A. Rudakov, *Artificial Earth Satellites*, Issue 6, p. 48, 1961.
5. K. I. Gringauz, V. A. Rudakov and V. A. Kaporsky, *Artificial Earth Satellites*, Issue 6, p. 33, 1961.
6. K. I. Gringauz, V. V. Bezrukikh and V. D. Ozerov, *Artificial Earth Satellites*, Issue 6, p. 64, 1961.
7. K. I. Gringauz, V. V. Bezrukikh, V. D. Ozerov and R. E. Rybchinsky, *Dokl. Akad. Nauk S.S.S.R.*, **131**, p. 1301, 1960.
8. K. I. Gringauz, V. I. Moroz, V. D. Kurt and I. S. Shklovsky, *Astronom. J. (U.S.S.R.)*, **37**, No. 4, p. 716, 1960.
9. I. S. Shklovsky, *The New Scientist*, **8**, No. 198, p. 591, 1st September, 1960.
10. J. M. Kelso, *J. Atmos. Terrestr. Phys.*, **16**, No. 3–4, p. 360, 1959.
11. E. E. Mityakova, N. A. Mityakov and V. O. Rappoport, *Izvest. Vysshykh Uchebnykh Zaved, Radiofizika*, **3**, No. 6, p. 949, 1960.
12. J. V. Evans, *Proc. Phys. Soc.*, **69B**, p. 953, 1956.
13. V. A. Rudakov, *Artificial Earth Satellites*, Issue 11, 1961 (now in press).
14. J. C. Seddon and J. E. Jackson, I.G.Y. Data Center, A. Rockets and Satellites, No. 1, p. 146, 1958.
15. V. G. Istomin, *Dokl. Akad. Nauk S.S.S.R.*, **130**, No. 5, p. 1066, 1960.
16. M. Nicolet, "Meteors" (Ed. P. R. Keuser), p. 99 (Pergamon Press, London, 1955).
17. R. L. F. Boyd, *Proc. Royal Soc.*, **A201**, p. 329, 1950.
18. "Space Research", Volume II (North Holland, Amsterdam, 1961—*in press*).

Manuscript received by the Institution on 5th July 1961 (Paper No. 681).

U.D.C. 52/5: 629.19

The Scientific Uses of Earth Satellites

By

J. H. BLYTHE, M.A., Ph.D.†

Presented at the Convention on "Radio Techniques and Space Research" in Oxford on 5th–8th July 1961.

Summary: A survey is given on the present position of certain branches of research which have been considerably influenced by satellite experiments and the value of the new methods is demonstrated. Topics discussed include the earth's gravitational field, its atmosphere, the upper regions of the ionosphere, magnetic fields near to the earth and in interplanetary space, and the zones of trapped radiation.

1. Introduction

Many satellites have been placed in orbit since the first *Sputnik* of four years ago, and a number of experiments of geophysical interest has been successfully carried out. To give some idea of the emerging picture of the value of these experiments the most immediate method is to survey the state of our knowledge of certain phenomena, selecting those areas in which satellite research has most changed our ideas. To preserve a correct perspective, limitations imposed by the nature of the orbits will be pointed out, and alternative experimental techniques will be indicated.

The following topics have been selected for discussion:

Gravitational field of the earth.

Structure of the earth's atmosphere.

Structure of the ionosphere above the F2 layer maximum.

Magnetic fields near to the earth and in interplanetary space.

The zones of trapped radiation.

2. The Gravitational Field of the Earth

Pre-satellite measurements of the gravitational field of the earth were made by observing the period of a pendulum, or by timing bodies in free fall. These were supplemented by the use of gravimeters, carefully designed spring balances, as relative instruments to interpolate between the absolute measurements. The study of the gravitational field has been inseparably linked with that of the figure of the earth ever since 1743 when Clairaut calculated the latitude variation of gravity on the assumption that the figure of the earth was very nearly that of an ellipsoid in fluid equilibrium. Clairaut's formula enabled the ellipticity of the earth's figure to be calculated, independently of trigonometrical determinations, from the observed variations of gravity over a wide range of latitude. The accuracy of geodetic triangulation has not been adequate to detect any significant departure of the true ellipticity from the "hydrostatic" value, and it has been assumed as the "basic hypothesis of geodesy" that the earth's gravitational field is very nearly that of a fluid in equilibrium. On this hypothesis a number of astronomical methods are available for finding the earth's ellipticity, in addition to the gravity method above. The most accurate of these uses the precessional constant (i.e. the mean rate of precession of the earth's axis about the pole of the earth's orbit), and the most recent value[1] of the hydrostatic flattening, derived in this way, is

$$e^{-1} = 299{\cdot}8 \pm 0{\cdot}3$$

By this means the gravitational potential of the earth, U, may be calculated. In terms of zonal harmonics it is written

$$U = \frac{GM}{r}\left[1 - \sum_{n=2}^{\infty} J_n \left(\frac{a}{r}\right)^n P_n(\cos\theta)\right] \quad \ldots\ldots(1)$$

where G is the gravitational constant, M and a are the earth's mass and mean radius, r is the distance from the centre of the earth, $P_n(\cos\theta)$ is a Legendre polynomial, θ is the co-latitude, and the J_n are dimensionless constants. The expression above is not a complete representation of the earth's field since it does not permit any variation with longitude. At the present stage the tesseral and sectorial harmonics are omitted since they have not been measured with sufficient accuracy to be usefully discussed. Values of J_n may be calculated from the flattening on the hydrostatic hypothesis which make all odd-numbered harmonics vanish. In view of the uncertainties of this method values were given only for J_2 and J_4, which could be justified by gravity measurements. It is now possible to check these values by satellite observations.

King-Hele has shown that departures of the earth's gravitational field from spherical symmetry give rise to three main perturbations of the orbit of a near earth satellite.[2] There is, first, an oscillation in radial distance of about 1 km; second, a rotation of the major axis of the ellipse; and, third, a rotation of the orbital

† Marconi's Wireless Telegraph Co. Ltd., Research Laboratories, Great Baddow, Essex.

plane of the ellipse about the earth's axis—the regression of the nodes. Fortunately these effects may be readily differentiated from perturbations due to other causes, such as the reduction of the semi-major axis due to drag (see Section 3), the slow variation of the angle of the inclination of the orbit arising from the rotation of the atmosphere, or the perturbations in perigee height produced by solar radiation pressure.[3] The rate of regression of the nodes, $\dot{\Omega}$, is a specially useful quantity since it may be measured very accurately. It is given by[4]

$$\dot{\Omega} = \sum_{n=2}^{\infty} J_n F_n + \sum_{m=2}^{\infty} \sum_{n=2}^{\infty} J_m J_n F_{mm} + \ldots \quad \ldots\ldots(2)$$

where the F_n and F_{mm} are orbital constants. It turns out that the odd-numbered harmonics have little effect, and that only the J_2^2 term in the second expression on the right contributes significantly to the errors. Hence in effect each value of $\dot{\Omega}$ gives one linear relation between the even-numbered J_n. Thus the even-numbered J_n may be determined from observations on a number of satellites, with different orbital inclinations to avoid ill-conditioned equations. King-Hele and Merson first used observations on *Sputnik II* and *Vanguard I* to find improved values of J_2 and J_4. These values were incompatible with the hydrostatic flattening. A final blow for the "basic hypothesis of geodesy" was the detection by the N.A.S.A. group in America of a non-vanishing J_3 component from periodic variations in the eccentricity of *Vanguard I*. From the analysis of a long period of observations in *Vanguard I* and other satellites the N.A.S.A. group has obtained[5] values of the coefficients J_2, J_3, J_4 and J_5. More recently King-Hele[4] has given a value for J_6 obtained from observations on *Sputnik II*, *Vanguard I*, and *Explorer VI*. His latest value for the flattening is $e^{-1} = 298{\cdot}24 \pm {\cdot}02$. Thus there are now available enough data to test theoretical explanations of the departures from fluid equilibrium, and there is considerable activity in this field.

It has been noted that the discrepancy in the observed bulge is in the sense to be expected for a decelerating earth. In fact, with the present rate of tidal deceleration, 5×10^{-22} rad.s^{-2}, the observed ellipticity is that for an earth in fluid equilibrium and the angular velocity it had 10 million years ago. O'Keefe has calculated the lifetime expected for the harmonics on the hypothesis that they are maintained by viscous forces as only of the order of a thousand years. He concludes that the anomalies in mass distribution must be supported either by mechanical strength or by hydrodynamic forces such as those arising from convection. It has been shown that convection currents might be able to account for the results, but the required thermal efficiency is too high on present simple theories.[6]

The depth of the anomalies is not given by the satellite data. Munk and McDonald have shown that there is no relation between the observed harmonics and the distribution of continents, and suggest that density variations in the mantle determine the low-order zonal harmonics.[7] They point out that the tesseral harmonics would provide a test of this hypothesis. According to Heiskanen the sectorial and tesseral harmonics are probably best determined by gravity surveys.[8]

3. Structure of the Earth's Atmosphere

The main effect of the atmosphere on a satellite is to cause it to lose energy, which manifests itself as a shortening of the period and a reduction of the semi-major axis. Most of this reduction is taken up in the height of the apogee, the perigee height remaining fairly constant throughout the life of a satellite. The period T of a satellite can be observed very accurately, and most of the work on atmospheric density to date has used the observed rate of change of period, $\mathrm{d}T/\mathrm{d}t$, as starting point. The atmospheric density ρ is assumed to vary exponentially with height y, and to be independent of longitude and latitude, so that

$$\rho = \rho_0 \exp(-y/H) \quad \ldots\ldots(3)$$

where the value of H, the scale height, is not known exactly. It can be shown[9] that if H^* is the best estimate of H, the air density, ρ^* say, at a height $1/2H^*$ above perigee height is given in terms of the rate of change of orbital period, $\mathrm{d}T/\mathrm{d}t$, by the equation

$$\rho^* = -\frac{\mathrm{d}T}{\mathrm{d}t} \frac{m}{3FSC_D} \sqrt{\frac{2e}{\pi a H^*}} \times \left\{1 - 2e - \frac{H^*}{8ae} + 0\left(e^2, \frac{H^{*2}}{ae^2}\right)\right\} \quad \ldots\ldots(4)$$

where a is the semi-major axis of the orbit, e is its eccentricity, F is a factor allowing for the rotation of the atmosphere which normally lies between 0·9 and 1·0, C_D is a drag coefficient which normally lies between 2·0 and 2·3, and S is the effective cross-sectional area of the satellite perpendicular to the direction of motion. A mean cross-sectional area can be computed from the dimensions on reasonable assumptions about the way that the satellite is tumbling in flight. In the case of *Sputnik II* and the rockets of *Sputniks I* and *III*, for which m and S are unknown, useful results may be obtained by comparison with *Sputnik I*. All terms on the right-hand side of (4) are therefore known, and hence the atmospheric density at perigee may be determined. This method has been exploited from the earliest launchings, since it provides information on a region of the atmosphere not previously accessible. Rocket measurements extended to about 220 km, and densities above that

height had to be obtained by extrapolation on the basis of an approximate theory, and were subject to large errors.

A limitation of the satellite drag method arises because certain other physical effects, including radiation pressure and electrical drag, produce a similar perturbation of the orbit. Calculations of the effect of radiation pressure, which show the effect to be small, have been verified by observation on *Vanguard I*.[3] The situation with electrical drag is not so satisfactory. A charged body moving through the ionosphere will experience an electrical drag, but the magnitude of the drag is not readily calculable. If the potential is of the order of -10 volts, as measurements on *Sputnik III* indicate, theory seems to show that the electrical drag is small.[10] However, satellite potentials are in general unknown, and electrical drag remains a major uncertainty in the densities derived at altitudes between 400 km and 700 km. A detailed survey of the present position is given by Chopra.[11] Another limitation is that the density is obtained at a single point, perigee (or, more precisely, an average over the orbit which is heavily weighted near perigee) the position of which varies in a way completely out of the control of the observer once the satellite is in orbit. Despite these drawbacks, observation of many satellites over a long period of time has revealed some interesting new features of the atmosphere.

It soon became apparent that the density was subject to variations which made the business of long-term prediction a rather hazardous one. It has now been established that solar activity is a determining factor for day-to-day variations in the density at heights between 200 km and 300 km, with the density tending to be high when solar activity, measured by sunspot number and radiation at 20 cm, is high. King-Hele has also found evidence of a long-term variation, in the sense that densities now appear to be some 20% lower than in 1957–8, the time of sunspot maximum.[12]

Another remarkable effect of the sun on the earth's atmosphere was observed by Jacchia, who noted that on two occasions there were abrupt increases in the drag on *Sputnik III* about one day after a large solar flare.[13] This suggests that corpuscular streams from the flare influenced the atmosphere. The geomagnetic field presumably directed the solar particles into auroral latitudes, where they caused the atmosphere heating and expansion giving the increased drag on *Sputnik III*.

These solar effects have been found at heights of 200 km to 300 km. At greater heights, 400 km to 700 km, the most striking effect is a large diurnal variation, first noted by Jacchia from a reduction in the slope of the period-versus-time curve of *Vanguard I* when the perigee point entered the earth's shadow. The density changes by about 10 to 1 between day-time and night-time at 700 km. It appears to depend on the zenith distance of the sun, which will produce a variation with latitude and season. Thus the density should be high by day in the equatorial regions and low near the poles in winter. The data is not yet good enough to show whether there is a slowly varying component matching that found at low altitudes.

A point requiring some discussion is that the satellite data does not show the strong variation with latitude apparently found using rockets. For example, the summer day-time atmospheric density found in 1957 200 km above Fort Churchill, Canada, latitude 59° N was about five times the corresponding value found in 1951 at White Sands, New Mexico, latitude 32° N. Some variation with latitude is indicated by the satellite drag data, in that densities from the Russian satellites flying at high latitudes tend to be higher than those from American satellites. However, the effect is not marked, and King-Hele[12] remarks that of 24 densities obtained from satellites at heights between 180 km and 300 km none departs from the mean curve by a factor of more than 1·6. It has been suggested that the latitude effect is associated with atmospheric heating by energetic charged particles (Fort Churchill lies in the auroral zone), possibly those in the Van Allen belts. The effect may therefore peak at auroral latitudes, whereas the satellite density is the average over a wide range of latitudes. It is also likely that much of the difference between the two rocket measurements is to be associated with the different epochs of the sunspot cycle.

Attempts have been made to interpret the density measurements in terms of the composition and temperature which together control the scale height in equation (3) above. There is very little direct information available on the composition of the atmosphere at heights of several hundred kilometres, but, on the basis of certain assumptions as to the nature of the equilibrium and adopting the composition at 120 km given by rocket experiments, Bates has derived an atmosphere which gives a reasonable fit with density data.[14] His results have some interesting features. He finds that atomic oxygen is the main constituent of the atmosphere in this region, and that the amount of atomic hydrogen is much less than previously supposed. Mass spectrometric observations would be of great interest to confirm these deductions. Another unexpected feature is the high temperature gradient found just above the E layer, as much as 30° K per km at 120 km. While it is clear that there is a marked falling off in this gradient, the data are not yet good enough to decide whether the temperature continues to increase above the peak of the F layer, or whether it becomes effectively constant. Analysis of more

extensive data by Priester and Martin has revealed a "wiggle" in the density-height curve, which corresponds to a temperature inversion in the F1 layer.[15] Theoretical investigations of the high regions of the atmosphere have been made by Johnson[16] and Singer.[17]

Atmospheric research in the immediate future will be directed to obtaining further data of atmospheric density, and it is of great interest that the manometers in *Sputnik III* gave good results, as this is a much more flexible technique than the orbital observations which only give data near perigee. Data on composition are also required, and there are plans to fly mass spectrometers.

4. Structure of the Ionosphere above the F2 Layer Maximum

A great number of research centres were investigating the ionosphere before the advent of satellites, and many of them have now added the observation of satellite signals to their repertoire, since this promises a means of investigating the electron density above the maximum of the F2 layer, a region of the ionosphere which is screened from conventional swept-frequency sounders by the lower regions of the layer. As a result of the magnitude of effort a great diversity of techniques has been applied and we have space only to outline the main types of experimental methods. Methods based on the presence of a transmitter in the satellite include the observation of the Faraday rotation, observation of the Doppler effect, and observation of the satellite "rise" and "set". Most popular, in view of the experimental simplicity, has been the observation of the Faraday rotation.

It is well-known that when a linearly polarized wave propagates in an ionized medium containing a magnetic field, Faraday rotation of the place of polarization occurs at a rate given by

$$\mathrm{d}\Omega = \frac{K}{f^2} NH\cos\theta\,\mathrm{d}s \qquad \text{.....(5)}$$

where $\mathrm{d}\Omega$ is the amount of rotation in a distance $\mathrm{d}s$, f is the wave frequency, N is the electron density, H is the magnetic-field and θ is the angle between the field and the direction of propagation. The constant K is equal to $\frac{e^3\mu_0}{8\pi^2 m^2 c\varepsilon_0} = 2{\cdot}97\times10^{-2}$ m.k.s. units. For heights small compared with the earth's radius the magnetic field does not vary greatly, and hence we may write

$$\Omega = \frac{K}{f^2}\,\overline{H\cos\theta}\int N\,\mathrm{d}s \qquad \text{.....(6)}$$

Thus the total rotation angle is proportional to, and may therefore be used to measure, the integrated electron content of the ionosphere, that is the number of electrons in a vertical column of unit area from the ground to the height of the satellite. This principle was first applied to measure the total electron content of the ionosphere using moon echoes.

The Faraday rotation manifests itself by a regular fading of the amplitudes of the signal received on a dipole, at a rate of the order of 1 c/s for the 20 Mc/s signals from the *Sputniks*. According to equation (6) the fade corresponding to zero rotation angle occurs for $\theta = 90°$, that is for propagation transverse to the magnetic field. In principle this may be identified by a minimum in the fading rate. Then the rotation angle corresponding to the point nearest overhead, the time of which is found independently, may be obtained by counting fades between the two points on the record. This gives one measure of the integrated ionospheric electron content. Another estimate may be obtained directly from the fading rate, on the assumption that the ionosphere is spherically stratified.

It turns out that an important correction is necessary to the simple ideas outlined above. The Faraday rotation may be regarded as due to interference between the ordinary ray and the extra-ordinary ray of magnetic-ionic theory, and path differences of these rays, neglected above, produce significant departures from equation (6). It is possible to make allowance for this effect, and also for the effect of horizontal gradients, on the basis of a model of the ionosphere, such as that given by ionospheric sounding. However, the procedure involves ray tracing through the ionosphere, which has a certain nuisance value as it means that considerable computation is required to reduce the data.

When the necessary corrections are made, reasonable values for the integrated electron content are obtained. Garriott gives the results of an analysis of an 8-months period, and his results show the annual variation of the total electron content.[18] He gives an estimate of the average electron-density profile using the slow variation in the height of passes over his station due to rotation of the perigee position mentioned above (Sect. 2). He also shows that the total electron content is near normal on magnetically disturbed days in which the maximum electron density falls by 25 to 40%. This implies a severe distortion of the layer.

The effect of the ionosphere on the Doppler frequency shift provides another technique for ionospheric investigation. It turns out that the Doppler frequency is affected by two terms, the larger a function of the integrated electron content, and the smaller related to the electron density at the height of the satellite. Thus the measurements may give both the integrated electron content, and, when combined with Faraday data, density at the satellite.

If the effect of the ionosphere is to be estimated by comparison of an observed Doppler frequency-time curve with one computed from orbital data an unreasonable accuracy is required of the orbital data. A more satisfactory method is to arrange two or more harmonically related transmitters on the satellite, say on frequencies f_1 and f_2 where $f_2 = 2f_1$. The orbital Doppler frequency shift is proportional to frequency, whereas the ionospheric effects decrease with frequency. Thus the ionospheric effects may be isolated by frequency doubling the received signal on f_1 and mixing it with the received signal on f_2. Attempts were made at Great Baddow to apply this technique on the 20 Mc/s signal radiated by *Sputnik III* and its second harmonic. Records were obtained, but telemetry modulation at an unfortunate frequency hampered the analysis. A more convenient object for this technique is *Transit IIA*, which radiates phase-related signals of good strength on 54 Mc/s and 324 Mc/s. It is hoped that this technique will yield better values of the integrated electron content than the Faraday rotation, as it uses the mean of the Doppler frequencies of the ordinary and extra-ordinary rays, which is large compared with the difference, which is measured by the Faraday rotation, and should be less affected by path splitting.

The observation of the "rise" and "set" of satellites appears to have been the most successful early method of obtaining data on the upper F layer. *Sputniks I* and *II* were equipped with 20 Mc/s and 40 Mc/s transmitters of such high power, about 1 watt, that they could easily be received on amateur equipment. Hence the time of "rise" and "set" was found at a large number of points. The results were analysed by assuming various shapes for the F layer above maximum, taking the shape below given by ionospheric sounding, and tracing rays by conventional methods to find predicted times of "rise" and "set". When these agree with the observed times the assumed shape is a reasonable approximation to the true one. The first quantitative data on the profile of the upper F layer was obtained by this means.[20]

Consideration of the relative paucity of results that have been obtained which directly increase our knowledge of the upper F layer, in relation to the labour involved, prompts the conclusion that the satellite methods tried so far have been inadequate. This is emphasized by the recent development of so-called "incoherent scatter" radars sensitive enough to detect a return from electrons in the ionosphere.[21] These give directly the complete profile of the ionosphere above the radar. It is also possible to derive an estimate of the electron temperature from the broadening of the return.

There is a place for satellite research, if a more satisfactory experimental technique is evolved. Ways of measuring the electron and ion density at the satellite are to hand in such instruments as the ion trap flown in *Sputnik III*, the Langmuir probe and the aerial impedance experiment suggested by Storey.[22] Experiments of this type are being flown on *Explorer VIII*, and the results will be of great interest. A satellite designed to give good values of the integrated electron content has been described by Garriott and Little.[23] However, these experiments do not give the profile of the upper F layer. This might be obtained by flying a number of satellites in circular orbits at heights of, say, 350 km, 400 km, 450 km, etc., but a more elegant solution is to fly a swept-frequency sounder of suitable design, and work is in hand on this project in Canada.[24]

In contrast to the large-scale structure, the small-scale irregularities have been quite effectively investigated using a simple beacon transmitter. These irregularities manifest themselves as bumps in Doppler frequency-terms curves, as fluctuations in apparent satellite direction, and as fluctuations in record amplitude. Yeh and Swenson have shown that the night-time effects are correlated with "spread-F", as are the scintillations of radio stars.[25] They originate at heights near 200 km and below 300 km, mostly north of latitude 40° N. The day-time scintillation originates in small inhomogeneous patches distributed over wide ranges of latitude. Seasonal variations are slight.

5. Magnetic Fields near to the Earth and in Interplanetary Space

The dominant role played by magnetic fields in many cosmical phenomena has been recognized of recent years, and theoretical investigations have been started into the complex relationships that exist between magnetic fields and ionized matter in motion. These studies have as their goal the elucidation of, for example, the magnetic field of the earth, magnetic storms and aurorae, the nature of sunspots, the origin of the solar cycle, and cosmic ray intensity variations. It is clear that experimental investigations by means of satellites and space probes of the morphology of the magnetic fields are of key importance.

The main dipole component of the earth's magnetic field has been recognized since Gilbert's work in 1600, but even today the theory of its origin is incomplete. It appears likely, however, that a satisfactory explanation is offered by the "dynamo" theory which ascribes the field to convection currents in the earth's molten core. An alternative explanation is given by Bailey's hypothesis that stars like the sun carry a net charge $-Q$ given by[26]

$$Q = \beta G^{\frac{1}{2}} M \text{ e.s.u.} \qquad \ldots\ldots(7)$$

where β is a pure number of the order of 0·03, M is the stellar mass, and G is the constant of gravitation.

The orbital motion of the earth in the sun's electric field gives rise to a magnetic field of $6{\cdot}9 \times 10^{-3}$ gauss, and according to Bailey's theory this is sufficient to saturate a highly permeable shell some 20 km or so thick near the surface of the earth and hence to produce the observed geomagnetic field. Bailey also accounts for a number of other phenomena on the same hypothesis. This explanation differs so much from the dynamo theory that it should be possible to discriminate unequivocally between them by experiment, and Bailey has suggested several experiments. It appears to the author that the measurement of the magnetic field of the Moon can be included among such experiments, for if the permeability of lunar material is similar to that of terrestrial material the Moon should possess an appreciable field. However, no magnetic field was detected by the Russian probes *Lunik I* and *Lunik II* or by the American probe *Pioneer IV*. This is in accordance with the dynamo theory, since the Moon is not expected to have a liquid core.

Superimposed on the main field observed at the earth's surface are certain fields which vary in time, including the solar daily variations S_q, the lunar daily variation L, and the complex phenomena classed as magnetic storms. These effects are ascribed to currents in the atmosphere, predominantly in the D and E regions in the case of S_q and L. These current systems have been detected by rocket-borne magnetometers. A large part of the field observed in magnetic storms is also due to currents flowing above the earth's surface. The theory of Chapman and Ferraro assumes that an electrically neutral stream of particles emanates from the Sun, and on reaching the earth interacts with the geomagnetic field to set up a ring current round the earth at a height of several earth radii. This ring current, together with currents induced in the earth's surface, produces the observed field. To resolve the many uncertainties involved in the detailed development of this theory, direct observations of the growth and decay of the currents are required. Currents have been detected by magnetometers on *Explorer IV* and *Pioneer V* of about 10^6 amps at 50 000 to 60 000 km height, and currents have also been detected by *Lunik I* at 20 000 to 25 000 km height.[27]

Another interesting feature of the geomagnetic field is the way in which it joins on to the interplanetary field. Parker has developed the theory that the high temperature of the solar corona causes it to expand continuously, producing a "solar wind".[28] Decay times in interplanetary space are of the order of thousands of years, so that magnetic fields are effectively "frozen" into the solar wind. Thus the Sun will have a virtually radial field, rendered somewhat spiral by rotation, which terminates in a region of disordered field roughly between the orbits of Mars and Venus. It is suggested that variations in this region resulting from changes in solar conditions produce the observed modulation of the intensity of primary cosmic rays incident on the earth. It is also to be expected that the solar wind will screen off the earth's magnetic field, and cause it to terminate at a certain height. Calculations show that with currently reasonable wind densities the cut-off should occur at about six earth radii, or 40 000 km. Magnetometer observations on *Explorer IV* and *Pioneer V* show a complicated situation, with violent fluctuations in the field above the ring currents observed at 50 000 to 60 000 km, and these fluctuations persist to 100 000 km where the earth's field appears to terminate.[29] It has been suggested that the fluctuations are hydromagnetic waves launched by the solar wind.

Whilst it may be argued that the fields near to the earth are consistent with the solar wind hypothesis, the same is not true of the interplanetary field detected by *Pioneer V*.[29] The measured component of the field, normal to the spin axis, was roughly constant at about $2{\cdot}7 \times 10^{-5}$ gauss, whereas on the solar wind hypothesis it should have become zero, and then reversed. The value is too large to be part of the Sun's dipole field, and if it is assumed that the galactic field penetrates the solar system the behaviour of cosmic ray particles is very difficult to understand. However, Bailey has pointed out that the behaviour of the field is in accordance with his hypothesis of an electrically charged Sun, although the magnitude is rather small.[30] Sudden increases in the interplanetary field are also observed, up to perhaps 4×10^{-4} gauss, and it has been assumed that these increases are associated with clouds of charged particles emitted by the Sun. Without enquiring into the origin of the field Dungey has pointed out that the incidence on the earth of a plasma containing a magnetic field may be expected to produce currents distributed like the well-known S_q system.[31]

The satellite and space probe observations described above have shown that conditions in the solar system are far more complex than had been supposed. More observations are required to describe this situation, and experiments are in preparation, including a space probe which will detect plasma as well as measure magnetic field, so that the relation between them may be observed in detail.

6. The Zones of Trapped Radiation

One of the most interesting results of satellite research has been the discovery of belts of charged particles trapped in the earth's magnetic field. The particles were detected in 1958 by Geiger tubes carried on *Explorer I* and *Explorer III* and also by a scintillation detector on *Sputnik III*.

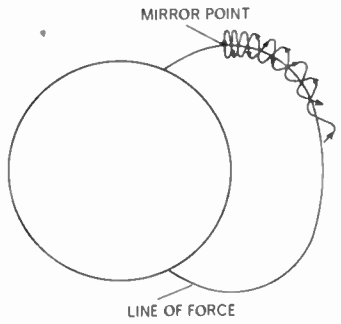


Fig. 1. Trajectory of a trapped particle.

The motion of such particles had previously been investigated theoretically by Stormer and Alfven, who showed that particles spiral along the lines of force, the pitch of the spiral increasing as the particles approach the earth. At a "mirror point" the pitch is reversed, and the particle reflected, as shown in Fig. 1. It then undergoes the same process at the conjugate point.

As a result of the curvature of the lines of force, and the radial gradient of field strength, the trajectories of the particle undergo a slow drift in azimuth, positive and negative particles drifting in opposite directions.

Collisions with atmospheric particles scatter the particles out of their trajectories, resulting in a decay of the trapped radiation. In the *Argus* experiment of August–September 1958 this process was studied by observing the behaviour of a layer of β-decay electrons artificially injected by a small nuclear explosion at high altitude. The lifetime of these electrons was found to be about a week.[32] A notable feature of the experiment was the observation of aurorae and geomagnetic effects at the injection point and its conjugate.

The intensity structure of the two principal radiation zones as derived by Van Allen and his associates from *Explorer IV* and *Pioneer III* observations is shown in Fig. 2.[33]

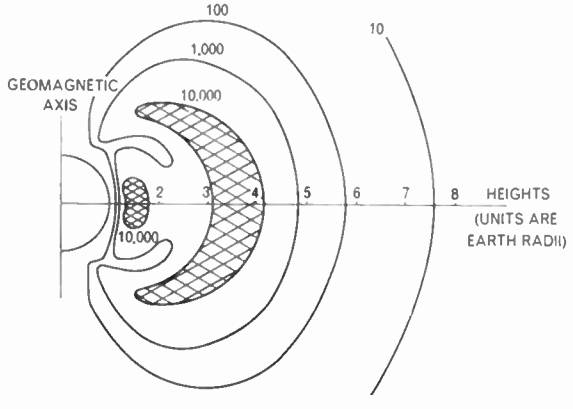


Fig. 2. The zones of trapped radiation.

The inner zone is relatively stable, whereas the outer zone is very variable. Structure has been detected in the outer zone, and on occasions a stable and distinct peak has been detected between the inner and the outer belts. These variations are apparently associated with solar activity.

The composition of the layers has been investigated by using a variety of detectors, and it has been shown that there are considerable differences in the composition of the two layers. For example a counter telescope carried by *Explorer VI* showed that energetic particles, i.e. protons with energy greater than 75 MeV, or electrons with energy greater than 13 MeV, occur only in the inner belt. On the basis of these experiments, it has been possible to estimate the density and energy spectra of the electrons and protons in the two layers.[33]

There has been some speculation on the source of the trapped radiation and a number of authors have drawn attention to the so-called cosmic ray "albedo" source. The neutrons produced by cosmic ray disintegrations in the atmosphere will move outwards without deflection, and occasionally decay in the upper atmosphere. The decay products of a neutron are a proton, an electron, and a neutrino. It appears that this mechanism can account for the protons and electrons in the inner belt but does not provide sufficient energy for the outer belt.[34]

Suggestive clues about possible mechanisms for the outer belt are given by certain correlations that have been observed. For example, the Geiger tube on *Explorer VI* showed a depletion of the outer zone within a day or so of a magnetic storm on 31st March 1960. Aurorae were also observed. The zone recovered within a week or so, showing at times considerable detailed structure. No great flux of energetic particles was detected in excursions of *Explorer VI* beyond the belt, which seems to rule out the possibility that the electrons arrive from the Sun with their full energy. Parker has suggested a mechanism whereby the electrons gain energy from collisions with magneto-hydrodynamic waves.

The correlation found between the outer belt and aurorae led Van Allen to suggest that aurorae follow a perturbation of the geomagnetic field which allows the "dumping" of outer belt particles. This suggestion enables correct predictions to be made of the altitude variation, and also of the isochasms in the arctic and antarctic zones. However, McIlwain has shown by a rocket experiment that the energy distribution of auroral electrons is not the same as in the outer belt.[35] It is evident that further work is required before the relation between the outer belt and aurorae is understood.

7. Acknowledgment

This paper is published by permission of the Director of Research, Marconi's Wireless Telegraph Co. Ltd.

8. References

1. J. A. O'Keefe, "Zonal harmonics of the earth's gravitational field and the basic hypothesis of geodesy", *J. Geophys. Res.*, **64**, p. 2389, 1959.
2. D. G. King-Hele, "The effect of the earth's oblateness on the orbit of a near satellite", *Proc. Roy. Soc.*, (*London*) **A.247**, p. 49, 1958.
3. P. Musen, R. Bryant and A. Bailie, "Perturbations in perigee height of *Vanguard 1*", *Science*, **131**, p. 935, 1960.
4. D. G. King-Hele, "Evaluation of the second, fourth and sixth harmonics in the earth's gravitational potential", *Nature* (*London*), **187**, p. 490, 1960.
5. J. A. O'Keefe, A. Eckels and R. K. Squires, "The gravitational field of the earth", *Astronomical J.*, **64**, p. 245, 1959.
6. A. L. Licht, "Convection currents in the earth's mantle", *J. Geophys. Res.*, **65**, p. 349, 1960.
7. W. H. Munk and G. J. F. McDonald, "Continentality and the gravitational field of the earth", *J. Geophys. Res.*, **65**, p. 2169, 1960.
8. W. A. Heiskanen, "The latest achievements of physical geodesy", *J. Geophys. Res.*, **65**, p. 2827, 1960.
9. D. G. King-Hele, "Density of the atmosphere at heights between 200 km and 400 km, from analysis of artificial satellite orbits", *Nature* (*London*), **183**, p. 1224, 1959.
10. P. J. Wyatt, "Induction drag on a large negatively charged satellite moving in a magnetic-field-free ionosphere", *J. Geophys. Res.*, **65**, p. 1673, 1960.
11. K. P. Chopra, "Interaction of rapidly moving bodies in terrestrial atmosphere", *Rev. Mod. Phys.*, **33**, p. 153, 1961.
12. D. G. King-Hele and D. M. C. Walker, "Density of the upper atmosphere and its dependence on the sun, as revealed by satellite orbits", *Nature* (*London*), **186**, p. 928, 1960.
13. L. G. Jacchia, "Corpuscular radiation and the acceleration of artificial satellites", *Nature* (*London*), **183**, p. 1662, 1959.
14. D. R. Bates, "Some problems concerning the terrestrial atmosphere above about the 100 km level", *Proc. Roy. Soc.* (*London*), **A.253**, p. 451, 1959.
15. W. Priester and H. A. Martin, "Earth satellite observations and the upper atmosphere", *Nature* (*London*), **188**, p. 200, 1960.
16. F. S. Johnson, "The exosphere and upper F region," *J. Geophys. Res.*, **65**, p. 2571, 1960.
17. S. F. Singer, "Structure of the earth's exosphere", *J. Geophys. Res.*, **65**, p. 2577, 1960.
18. O. K. Garriott, "The determination of ionospheric electron content and distribution from satellite observations: Part 1. Theory of the analysis; Part 2. Results of the analysis," *J. Geophys. Res.*, **65**, pp. 1139 and 1151, 1960.
19. W. J. Ross, "The determination of ionospheric electron content from satellite doppler measurements: Part 1. Method of analysis; Part 2. Experimental results," *J. Geophys. Res.*, **65**, pp. 2601 and 2607, 1960.
20. H. E. Newell and J. W. Townsend, "Report on Moscow I.G.Y. Conference", *Science*, **129**, p. 79, 1959.
21. V. C. Pineo, L. G. Kraft and M. W. Briscoe, "Some characteristics of ionospheric backscatter observed at 440 Mc/s", *J. Geophys. Res.*, **65**, p. 2629, 1960.
22. L. R. O. Storey, "A method for measuring local electron density from an artificial satellite", *J. Res. Nat. Bur. Stand.*, **63D**, p. 325, 1959.
23. O. K. Garriott and C. G. Little, "The use of geostationary satellites for the study of ionospheric electron content and ionospheric radio-wave propagation", *J. Geophys. Res.*, **65**, p. 2025, 1960.
24. R. C. Langille and J. C. W. Scott, "The Canadian Defence Research Board topside sounder satellite", Brit.I.R.E. Convention paper, Oxford, 1961. To be published in *J. Brit.I.R.E.*
25. K. C. Yeh and G. W. Swenson, "The scintillation of radio signals from satellites", *J. Geophys. Res.*, **64**, p. 2281, 1959.
26. V. A. Bailey, "Existence of net electric charges on stars", *Nature* (*London*), **186**, p. 508, 1960.
27. E. J. Smith, P. J. Coleman, D. L. Judge and C. P. Sonett, "Characteristics of the extraterrestrial current system: *Explorer VI* and *Pioneer V*", *J. Geophys. Res.*, **65**, p. 1858, 1960.
28. E. N. Parker, "Interactions of the solar wind with the geomagnetic field", *Phys. of Fluids*, **1**, p. 171, 1958.
29. P. J. Coleman, L. Davis and C. P. Sonett, "Steady component of the interplanetary magnetic field: *Pioneer V*", *Phys. Rev. Letters*, **5**, p. 43, 1960.
30. V. A. Bailey, "Apparent steady component of the interplanetary magnetic field", *Nature* (*London*), **189**, p. 44, 1961.
31. J. W. Dungey, "Interplanetary magnetic field and the auroral zones", *Phys. Rev. Letters*, **6**, p. 47, 1961.
32. J. A. van Allen, C. E. McIlwain and G. H. Ludwig, "Satellite observations of electrons artificially injected into the geomagnetic field", *J. Geophys. Res.*, **64**, p. 877, 1959.
33. J. A. van Allen, "The geomagnetically trapped corpuscular radiation", *J. Geophys. Res.*, **64**, p. 1683, 1959.
34. S. F. Singer, "Latitude and altitude distribution of geomagnetically trapped protons", *Phys. Rev. Letters*, **5**, p. 300, 1960.
35. C. E. McIlwain, "Direct measurements of particles producing visible auroras", *J. Geophys. Res.*, **66**, p. 2727, 1960.

Manuscript first received by the Institution on 3rd May 1961 and in final form on 1st July 1961 (Paper No. 682).

U.D.C. 591.11: 621.37

Continuous Electronic Recording of the Activity of the Perfused Frog Heart

By

I. A. BOYD, M.B., Ph.D., B.Sc.†‡

AND

W. R. EADIE (*Associate Member*)‡

Presented at a meeting of the Medical and Biological Electronics Group in London on 24th March, 1960.

Summary: A biological research project is described in which specialized electronic equipment was developed in order to obtain permanent records of the action on live tissue of extremely low concentrations of certain active substances.

1. Introduction

The purpose of this paper is to describe some of the problems which have arisen in obtaining permanent records of the behaviour of living tissue. In the present research project, in which the heart of a frog is made to operate under critically controlled conditions, it has proved necessary to adapt or develop electronic and other equipment to achieve the required degree of sensitivity and stability.

The paper is in three parts:

(1) An outline of the background and object of the investigation.

(2) A description of the experimental and recording technique. This includes a discussion of the problems, particularly hydrodynamic ones, involved in obtaining a high degree of stability in the behaviour of the heart preparation itself, in addition to a description of the electronic instrumentation. A simple explanation of the action of the heart is included for those not familiar with physiology.

(3) A brief review of the results so far obtained.

2. Background of the Investigation

Not so long ago 1 gramme of a substance was regarded in medical science as quite a small quantity. Continued improvement in methods of detection and measurement has shown, however, that much smaller quantities, such as the nanogramme (10^{-9} g), of certain substances may affect the behaviour of living tissue. In a few specialized fields of research, e.g. on vitamins or "trace elements", the picogramme (10^{-12} g) is a significant quantity.

In recent years the authors have become increasingly aware that many errors may arise in biological research due to an inadequate appreciation of the great sensitivity of live animal tissue to very small quantities of substances, whose presence may not even be suspected. Later in this paper it will be shown that a marked alteration in the behaviour of a heart may be produced by 10^{-13} g of a substance.

The research at present in progress has three main aims:

(*a*) To study the physiological behaviour of a frog heart under various experimental conditions, using more accurate methods of recording than have been used in the past. These experiments include variation of physical conditions, such as the pressure of the fluid entering the heart and the resistance against which it must pump out the fluid.

(*b*) To determine the "biological threshold" of action of certain substances on the heart preparation whose physiological behaviour has been studied in project (*a*).

(*c*) To trace some of the sources of "contamination" frequently present in research involving living tissue, and to eliminate these effects if possible. By "contamination" is meant effects produced, not by the substance under study, but by something else inadvertently reaching the tissue. Bacterial or chemical impurities may be present in the water used in the preparation of solutions, or they may be introduced into the solutions from vessels containing them or tubing through which they flow.

The usual measure of the purity of water is its conductivity, but this only gives an indication of the concentration of ions present; it gives no indication of the presence or absence of non-ionized substances, many of which are biologically active. Resin de-ionizers are being used increasingly as sources of water for biological and other work. It is claimed that they produce water of great purity because its conductivity is greater than 4 MΩ/cm. It may be almost ion free but its effects on a biological preparation may differ from that of water of similar conductivity from a glass still.

† Institute of Physiology, University of Glasgow.
‡ Boyd Medical Research Institute, 17 Sandyford Place, Glasgow, C.3.

In an investigation, such as the present one, which involves quantities of the order of 10^{-12} g to 10^{-15} g, contamination is a real problem. For example, if a solution of 10^{-15} g/ml of a biologically active test substance is put into a container which is not absolutely clean and sterile it soon contains contaminating substances in concentrations greater than that of the test substance. Again, if a pipette is used to transfer a solution of, say, 1 μg of a substance it cannot be used again in the same experiment; even though it is rinsed several times it may still contain enough of the original substance to interfere with a later test. All glassware, therefore, must be thoroughly cleaned and sterilized, and no piece may be used twice in one experiment.

Further, since the biological "threshold" is often beyond the physical or chemical threshold of detection by the means at present available, rigid control procedures are essential. In these the experiment is carried out exactly as planned but with the test substance absent. Only by doing this can one be sure that any effect obtained when the test substance is used is in fact due to its presence and not to some other unknown factor.

The present work, therefore, requires experimental and recording techniques of a very high standard.

3. Experimental and Recording Techniques

To express the requirements in electrical terms, a biological test preparation and a recording system, both of high gain, high stability, and low noise level, are needed.

3.1. *The Preparation*

A frog heart is used for the following reasons:

(*a*) It is spontaneously active. It is easier to detect small changes in the behaviour of an already active preparation when a test substance is applied to it than it is to detect small responses produced in a tissue which is inactive at the start.

(*b*) The behaviour of the heart can be measured in terms of several different variables, rather than a single one. During a test only one of these may show a change, but if more than one changes, a check that the effect is real, and not an artefact in the recording, is provided.

(*c*) The heart responds to very small changes in its environment, particularly to changes in the nature of the fluid passing through it. A mammalian heart may be even more sensitive than a frog heart, but the latter has the advantage that it will work at room temperature and does not require special oxygenation procedures.

(*d*) If the frog heart is made to operate under dynamic conditions similar to those in which it operates in the living animal, and these are kept constant, then the spontaneous activity is very stable indeed. Thus, very small changes in activity produced by the addition of a test substance become significant.

3.2. *Physiological Considerations*

Since a knowledge of the normal behaviour of the heart is necessary for the understanding of the experimental method, a brief description is given here.

The function of the heart in any animal is to maintain a flow of blood through the vascular system, carrying the required nutrient materials to the various tissues and removing their waste products. It is, in fact, a muscular pump which creates a pressure in the large arteries, this pressure forcing the fluid on through the blood vessels until it eventually returns to the heart via the veins.

The heart muscle contracts many times a minute. In man the heart rate is about 70 beats/min, in frogs it is 30–60 beats/min. Fluid is ejected into the large arteries at each beat. The flow through the vessels is maintained even when the heart relaxes after a beat, by a system quite similar to electrical "smoothing". The large arteries have elastic walls, especially the aorta, which is the part of the arterial system nearest to the heart. When the heart ejects its contents, the aorta and arteries distend and the pressure within them rises. When the heart relaxes, valves prevent backflow of fluid into the heart. The fluid flows forward through the blood vessels, driven by the pressure, which acts against the resistance to flow provided by the vessels, which are narrow tubes. The process is similar to the intermittent application of charge to a capacitor with a leakage resistance. Each application of charge raises the electrical pressure, or voltage, which then decays exponentially depending on the time constant of the circuit. The heart ejects fluid into a system with volume elasticity (analogous to the capacitance) raising the arterial pressure, which then decays due to leakage of fluid through the resistance of the blood vessels.

3.3. *Principle of the Method*

The heart is perfused with a fluid (Ringer's solution) which has a certain ionic composition, osmotic pressure, and pH. The fluid is similar to frog blood, but without the blood corpuscles or the plasma proteins. It flows into the heart all the time, at a measured pressure, the "venous pressure", through a small glass tube (cannula) tied into one of the veins leading to the heart (see Fig. 1). The frog itself is killed before the experiment begins, but the heart continues to beat for many hours. The heart pumps

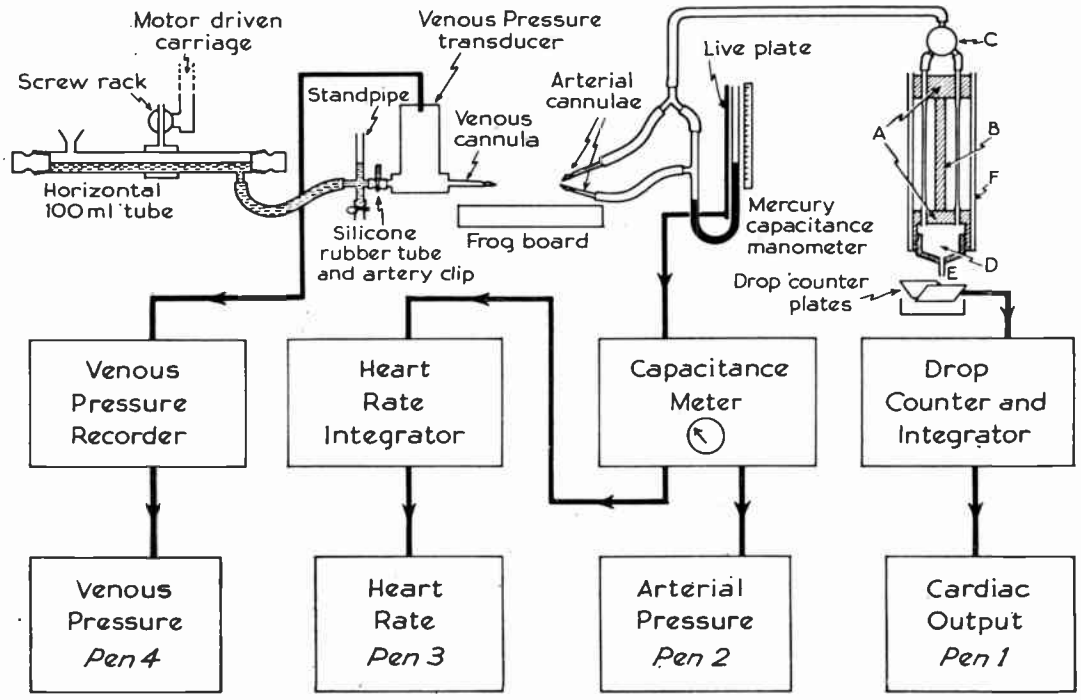


Fig. 1. Block diagram showing the experimental arrangement for continuously displaying, on a four-channel pen recorder, the arterial and venous pressures, output and rate of a perfused frog heart preparation.

the fluid out through small glass cannulae tied into the aortae, and thence through an artificial resistance made of drawn glass tubing (analogous to the resistance offered by the peripheral blood vessels in the living animal). The pressure developed by the heart (the "arterial pressure") is indicated by a small mercury manometer connected to the rubber tubing between the heart and the artificial resistance. The arterial pressure, the amount of fluid the heart pumps out per minute (cardiac output), and the number of beats of the heart per minute, are measured electronically and displayed on a chart, so that a continuous record of the behaviour of the heart is obtained. The pressure at which the fluid is made to flow into the heart (venous pressure) is also indicated on the chart continuously.

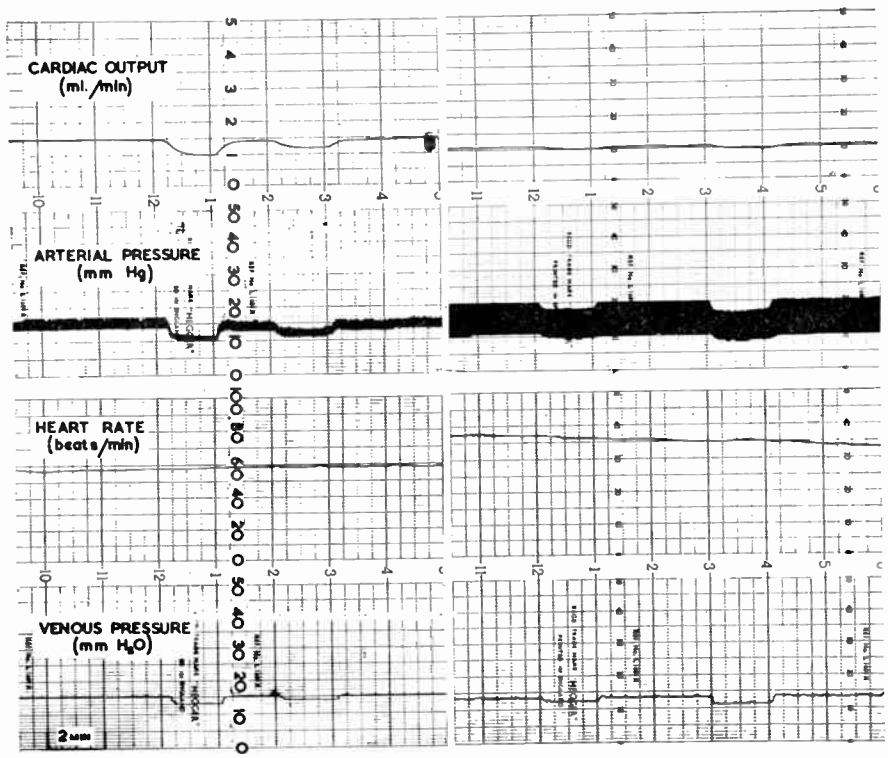
The dynamic conditions under which the heart operates may be varied by changing the value of the artificial arterial resistance, or of the venous pressure. The heart is, in fact, a pump of variable stroke and frequency.

To test the action of any substance on the heart, solutions of various concentrations of the substance are prepared in Ringer's solution. The arterial resistance and the venous pressure are fixed at values such that the heart produces an arterial pressure and an output of fluid similar to those found in living frogs. The heart is perfused with Ringer's solution, then with a chosen concentration of test substance in Ringer's solution for a given time, and then with pure Ringer's solution again. Any change in behaviour of the heart is recorded automatically on the chart.

3.4. *The Perfusion System*

Very small changes in the pressure of the fluid flowing into the heart produce considerable changes in its behaviour. The effect on two different frog hearts of lowering the venous pressure in small steps is shown in the recording of Fig. 2, which was obtained with the perfusion system now in use. The change in venous pressure, produced by reducing the height of the reservoir of fluid, is recorded on the lowest trace. In frog (*a*) a change from 15 to 13 mm of water alters the activity of the heart so that the systolic arterial pressure (i.e. the maximum pressure reached during each beat) falls from 17 to 11 mm Hg (about 30%), and the output falls from 1·5 to 1·0 ml/min (30%). A change in venous pressure of 1 mm H_2O produces a change of about 20% in both arterial pressure and output. In frog (*b*) the output is affected less, but the arterial pressure alters about 10% per millimetre of water change in the venous pressure.

Thus, changes of 10% or less in the value of the arterial pressure or output in any test procedure will



(a) (b)

Fig. 2. The effect on two hearts of reducing the venous pressure in small steps of 1 mm and 2 mm of H_2O. In frog (*a*), both arterial pressure and cardiac output fall considerably; in frog (*b*), the arterial pressure shows the largest change. The heart rate is not affected, though in (*b*) it is falling slowly throughout the record.

be significant only if the venous pressure is controlled to within a fraction of a millimetre of water. Further, the venous pressure must be recorded continuously throughout all tests to prove that any changes which take place in the activity of the heart are not attributable merely to changes in venous pressure.

There are a number of different devices in common use for providing fluid at constant pressure. Many of these depend on the Marriot principle, i.e. air bubbles escape from a fixed point below the surface of the fluid, the pressure being constant, supposedly, at the height of the air inlet port above the heart. In fact, the formation of the air bubbles at the inlet port results in marked fluctuations in the venous pressure. This is shown in Fig. 3, a record of an experiment in which the effects on a heart of perfusing from several types of "constant pressure" system were compared. With most of the methods illustrated, the venous pressure fluctuates, producing corresponding changes in the activity of the heart shown in the traces of output and arterial pressure. Further, all the systems which were tried, other than the one now used, which is described below, were difficult to fill and to sterilize.

The perfusion system at present in use consists of six horizontal glass tubes of about 100 ml capacity (one of which is shown in Fig. 1), with glass stoppers at both ends, funnels through which they are filled, and outlet tubes leading to six separate standpipes. Short silicone rubber tubes connect each standpipe to the pressure transducer, all the tubes except one being closed by artery clips. The transducer has a single outlet which is connected to a glass cannula tied into the posterior vena cava of the frog.

The height of the fluid column in the standpipe in use gives an approximate visual index of the venous pressure. The menisci in the different standpipes are brought to the same level by adjusting the screw racks holding individual horizontal "burettes". Perfusion from a particular burette is obtained by removing the appropriate artery clip and placing it on the connecting tube of the burette previously in use. Thus, changeover from one perfusion fluid to another may be effected with only a very small fluctuation in venous pressure. Any small difference in pressure is recorded by the transducer and indicated on a meter above the burettes; the difference is corrected at once by momentarily operating the motor-driven carriage on which all the burettes are mounted. This procedure is important because a fluctuation in venous pressure during a changeover from one solution to another may cause an alteration in heart rate which lasts many minutes. Also, a difference in venous pressure of a fraction of 1 mm of water may affect the arterial pressure produced by the heart.

The area of the fluid surface in each burette is large and the level falls slowly during perfusion, since the output of the heart is only 1 to 2 ml/min. The

venous pressure is easily maintained within the required range of stability of $\pm\frac{1}{4}$ mm of water; the slow fall in fluid level is compensated by raising the burette carriage fractionally once every few minutes.

Under these conditions the preparation is very stable, as is shown in period 4 of Fig. 3, and in Fig. 7.

3.5. *The Venous Pressure Recorder*

The requirements for the venous pressure recorder are six in number:

(1) Range. 0–50 mm of water.

(2) Sensitivity. It must be sufficiently sensitive for changes in pressure of $\frac{1}{4}$ mm of water to be seen clearly on the tracing of venous pressure.

(3) Linearity. Since the pen recorders have a wide linear scale, a pressure recorder with a linear response is desirable.

(4) Volume. The delay between the removal of the clip on the connecting tube and the action of the corresponding fluid on the heart, should be constant and as brief as possible. The delay depends on the volume of fluid between clip and heart. This volume is reduced by having the junction of the six connecting tubes actually in the manometer. The volume of fluid in the manometer itself must be small and must not vary with change of pressure.

(5) Materials. No metal or other materials which might release contaminating substances should come in contact with the solutions.

(6) Sterilization. All parts in contact with the perfusion liquids must be sterilized at 120° C at least, preferably 150° C.

The transducer is shown in Fig. 1, and in detail in Fig. 4.

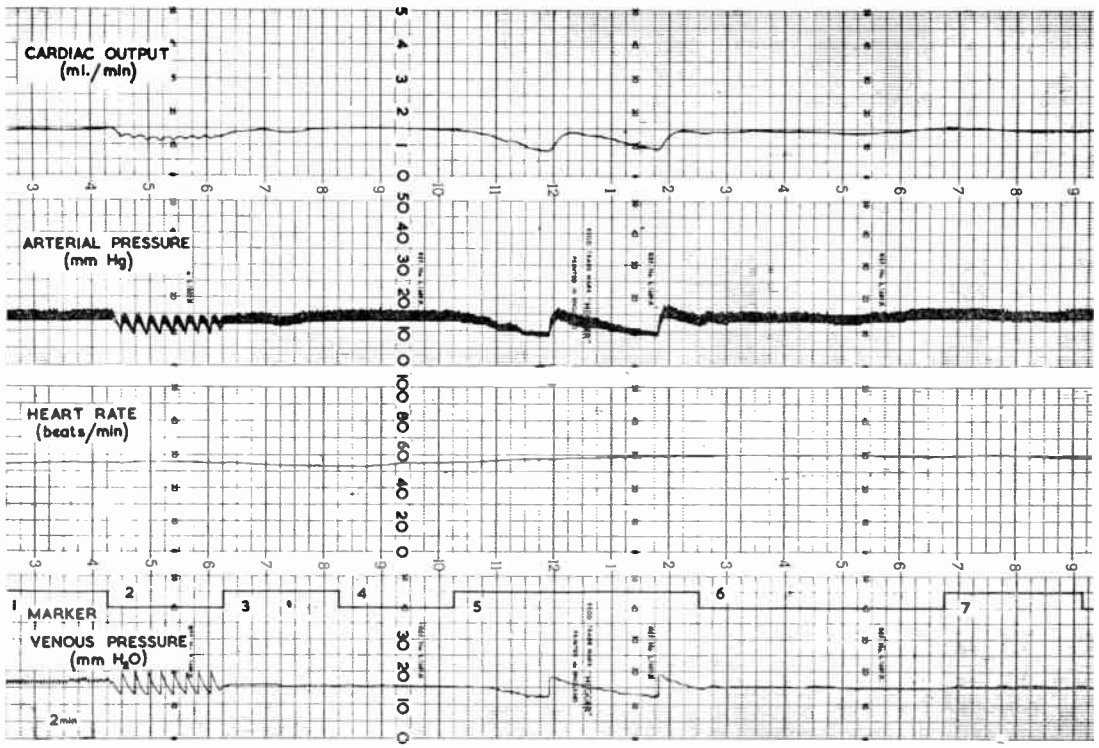


Fig. 3. The activity of a frog heart while perfused successively from different systems providing fluid at "constant pressure".

(a) Constant pressure burettes. Period 2 in figure; burette selected at random from a large number; venous pressure fluctuates 7 mm H_2O with formation of each air bubble at the inlet port, cardiac output and arterial pressure reflect these changes. Period 1; best burette available, fluctuations less serious.

(b) Marriot flasks. Period 5; air inlet of small diameter, marked change in venous pressure and hence in heart activity. Period 7; air inlet of optimum diameter 4–5 mm.

(c) Coils of tubing in horizontal plane: almost vertical fluid meniscus moves round coil as fluid is used up. Period 3; glass tubing. Period 6; polythene tubing. Fluctuation in venous pressure and heart activity present in both cases.

(d) Horizontal tube system at present in use (see text). Period 4; venous pressure and heart activity stable.

and by calibrating it against known pressures a recording of the arterial pressure is obtained on channel 2 (Fig. 1).

This system has two great advantages. Initial calibration is easy; the resistance is clipped off and the pressure is raised in steps of 10 mm Hg by injecting fluid from a syringe attached in place of the heart. Also, the excursion of the mercury column provides a continuous visual check on the correct functioning of the recording system. Any drift is indicated by a disparity in the readings of the mercury column and the proximity meter, and can be corrected at once by adjustment of one panel knob.

A high degree of mains stabilization is necessary. A change of 1 volt in the mains supply produces a change of 0·5% in the output of the particular Fielden proximity meter in use (Type PM2); this produces a movement of the pen of $\frac{1}{4}$ of a division on the chart, and is the limit of variation which can be tolerated.

3.8. *The Recording of Heart Rate*

The electrical output of the arterial pressure recorder rises and falls during each heart beat and provides a convenient signal for operating a heart rate-meter. The circuit, which follows standard practice, is shown in Fig. 5(*a*). The first part of the circuit converts the recurring wave of irregular shape (first inset), derived from the arterial pressure recorder, into a square wave of constant amplitude (second inset). This is differentiated and the negative-going component is suppressed (third inset). The remaining component is used to trigger a univibrator, so that a square wave of constant amplitude and duration is produced for each heart beat (fourth inset). These

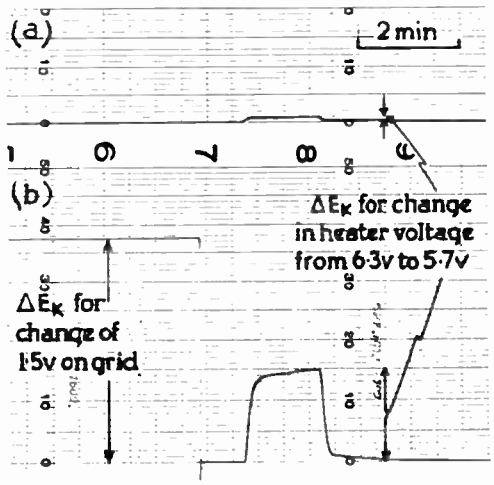


Fig. 6. The change in cathode voltage (ΔE_k) in a cathode follower when the heater voltage is reduced by 0·6 V. (*a*) With the two-valve cathode follower of Fig. 5(*b*), ΔE_k is small. (*b*) With a single valve arrangement, ΔE_k is large. For comparison, ΔE_k for a change of 1·5 V on the grid is shown in (*b*) also.

square waves are applied to the two 2 μF reservoir capacitors through a diode pump, developing a voltage across them. There is a constant leakage of charge through the 1 MΩ and 2·2 MΩ resistors, so that the voltage at the output terminals at any moment is proportional to the rate at which the square waves arrive, i.e. proportional to the heart-rate. The square waves are also applied to a loud-speaker amplifier circuit producing an audible signal for each beat.

The ratemeter output voltage operates the pen-recorder and a trace is produced on channel 3 (Fig. 1) which is calibrated in beats per minute. Since the calibration of the ratemeter depends on the fixed leakage across the reservoir capacitors, this part of the circuit must be isolated from external factors such as leakage in connecting leads and the relatively low input impedance of the pen-recorder amplifier. A cathode follower unit is used for this purpose. This normally consists of a single valve circuit in which the signal is applied to the grid and the output is taken from the cathode. The stability of such an arrangement for d.c. signals is poor, however, especially to changes in heater voltage. Figure 6 shows clearly the seriousness of such changes.

A two-valve cathode follower[2] has therefore been adopted. The circuit is shown in Fig. 5(*b*) and its stability to heater voltage changes is excellent, as shown in Fig. 6(*a*). The same change in heater voltage applied to the two-valve arrangement produces a change in cathode voltage of 15 times less than when applied to the single valve circuit (Fig. 6(*b*)).

The sensitivity of the ratemeter is such that it will operate from an arterial pressure change of $\frac{1}{2}$ to 1 mm Hg.

3.9. *The Recording of Cardiac Output*

The cardiac output in the frog is between 1 ml and 2 ml per minute. It is convenient to measure an output as small as this by counting the number of drops expelled by the heart each minute. The drops fall between two silver plates, with a potential difference of 300 V between them. Each drop short circuits the plates as it passes, producing a sharp electrical impulse which is used to trigger an integrating ratemeter similar to that used for the heart rate (Fig. 5(*a*)). A universal type of input circuit has been adopted for the ratemeters, incorporating the necessary −300 V supply.

The output from the meter is displayed on the first channel of the pen recorder (Fig. 1), and the trace is calibrated in ml/min. To avoid irregularities in the trace, the drops must fall in a regular stream, and not in a series of several in quick succession with a pause between each series, as occurs with many simple experimental arrangements. Also, the drops

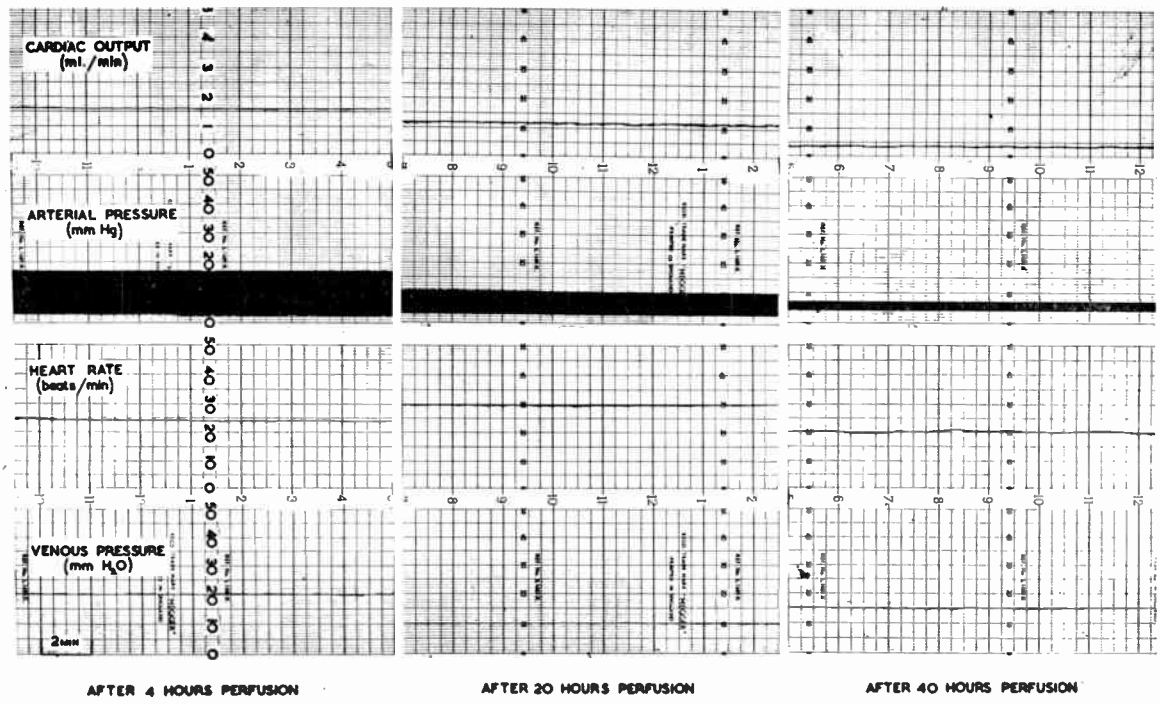


Fig. 7. Extracts from the recording from a frog 4 hr, 20 hr and 40 hr after setting up the preparation, showing its great stability if venous pressure and artificial resistance are accurately controlled.

must be of equal size throughout the experiment or the initial calibration of the trace will not be valid. The method of ensuring that these two conditions are met is described below.

3.10. *The Artificial Resistances*

Glass capillary tubes of about 1 mm bore were drawn out, and their approximate resistance was determined by measuring the flow of water through each with a constant pressure difference of 30 mm Hg between the ends. A tube of resistance 1 unit is defined as one through which a flow of 1 ml/min is produced by a pressure difference across it of 1 mm Hg. Eight tubes were selected covering a range of resistance from 5 to 50 units in approximately equal increments. The eight tubes are mounted in a circular arrangement so that their ends project through holes in two cylindrical blocks of perspex (A) rigidly connected by a perspex rod (B) (see Fig. 1). At the top end the resistances are connected by fine rubber tubing to the outlets of a specially constructed eight-way tap (C). A circular disc attached to the tap has eight notches, one of which engages with a spring-loaded ball-bearing in each position of the tap. This facilitates rapid changing from one resistance to another and ensures correct alignment of the tap bore in each position. At the lower end the tubes project into an airtight perspex chamber (D) with a single outlet (E). The lower part of this chamber is threaded into the perspex block, and can be removed for cleaning. The resistances are protected by an outer cover of large bore glass tubing (F), and the whole system is mounted vertically above the silver plates of the drop counter.

This resistance system has been developed after years of trial with simple arrangements none of which proved satisfactory. It has the following advantages.

(*a*) The resistance, when the tap is in any one position, is absolutely constant, and very stable behaviour of the heart is obtained, provided that the venous pressure is constant also. A constant resistance cannot be achieved, for example, when a screw-clip on rubber tubing attached to the arterial cannulae is used to provide the resistance.

(*b*) Change of resistance can be effected between one heart beat and the next, with minimum disturbance to the preparation, and with no interruption in the recording of cardiac output or arterial pressure. The outlet tube remains aligned over the silver plates; the size of the drops is unchanged and the initial calibration of the cardiac output recording equipment holds irrespective of which resistance is in use.

(*c*) The drops fall in a regular stream, whatever the heart rate, provided that the resistance used is

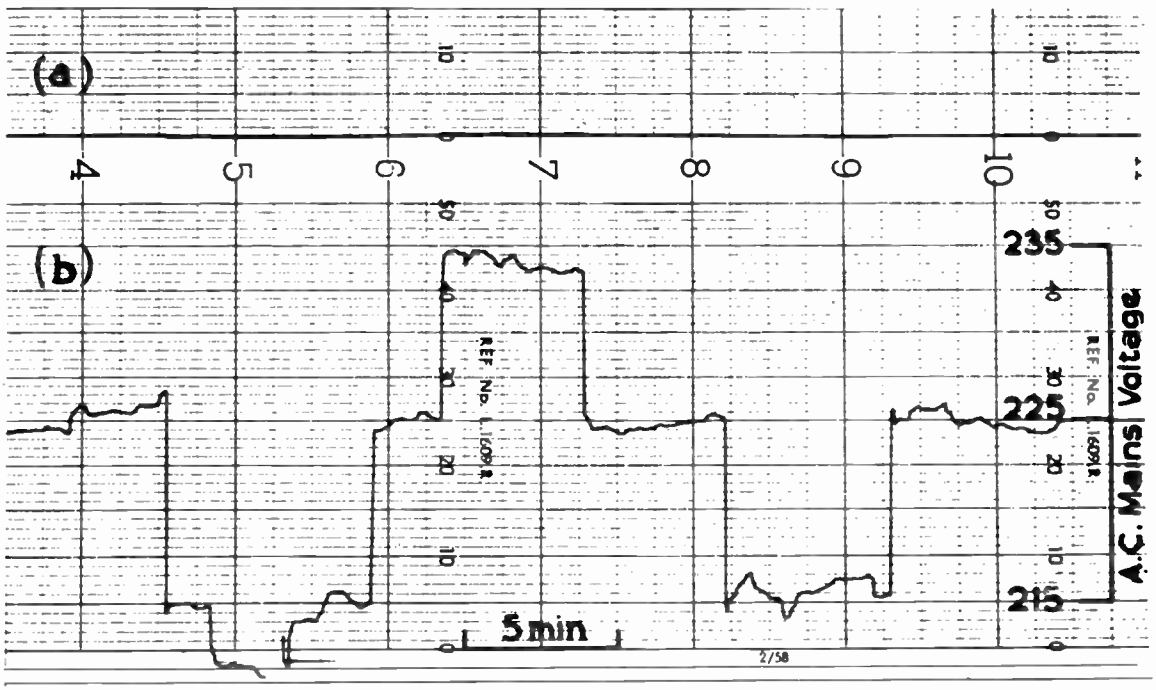


Fig. 11. (*a*) Tracing from servo-assisted pen-recorder with zero input, showing stability of pen-amplifier against supply voltage changes shown in (*b*); pen movement is negligible.
(*b*) Simultaneous recording, on a second channel, of the step changes of 10 V, superimposed on fluctuating mains voltage, applied to the pen-recorder amplifier whose output is shown in (*a*).

The pen recorder amplifiers are balanced internally against changes in supply voltages and can be operated directly from the mains.[3] The result of superimposing step changes of 10 V on the fluctuating mains voltage supply to one of these amplifiers is shown in Fig. 11. The voltage changes applied to the amplifier are recorded in (*b*). The effect on the zero of the pen itself is recorded in (*a*) and is scarcely visible.

4. Review of Results

The experimental and recording techniques described in Section 3 have been used in each of the three lines of investigation outlined in Section 2. A brief description of the kind of result obtained in one experiment of each type will now be given.

4.1. *The Dynamics of the Frog Circulation*

The recording of Fig. 12 is part of an experiment in which the dynamic conditions imposed on the frog heart were changed. The resistance against which the heart operated was altered in steps, as shown in the lowest tracing, by selecting different calibrated values of glass resistance tubes. As the resistance is increased the arterial pressure developed by the heart muscle rises, and the amount of fluid it ejects per minute falls. This result is of interest since it is assumed frequently that the output of the frog heart, and also of the mammalian heart, is independent of the resistance against which the heart works.[4]

4.2. *The Biological Activity of Small Quantities of Active Agents*

The action on the heart of very small amounts of acetylcholine has been studied in detail. Acetylcholine is a chemical substance which is liberated at the ends of nerves in the human body. Part of a typical experiment is illustrated in Fig. 13. At the start of this recording the heart was being perfused with pure Ringer's solution from burette 0, indicated by the marker pen. About 2 minutes from the start the clip on the connecting tube of burette 1, which also contained pure Ringer's solution, was removed and put on the tube of burette 0. The marker pen indicates this change. The heart was perfused from burette 1 for about 2 minutes, and then the clip was moved from connecting tube 1 back to tube 0. This constitutes the "control" perfusion. A change in the activity of the heart resulting from perfusion with an active agent can be considered significant only if this control procedure produces no appreciable change, as in this case.

At point 8 on the chart, perfusion from burette 2 was commenced. This burette contained acetylcholine in a concentration of 10^{-13} g/ml in the same

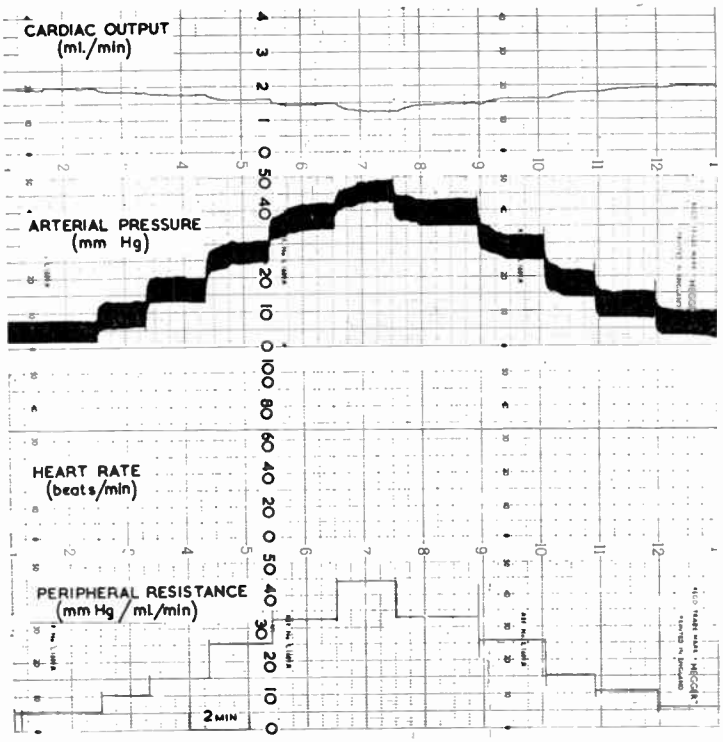


Fig. 12.

Response of a frog heart to increase in the artificial resistance against which it pumps out fluid; venous pressure maintained constant. Different resistances were selected in turn, as shown in lowest trace; with increased resistance arterial pressure rises, cardiac output falls and rate is unaffected.

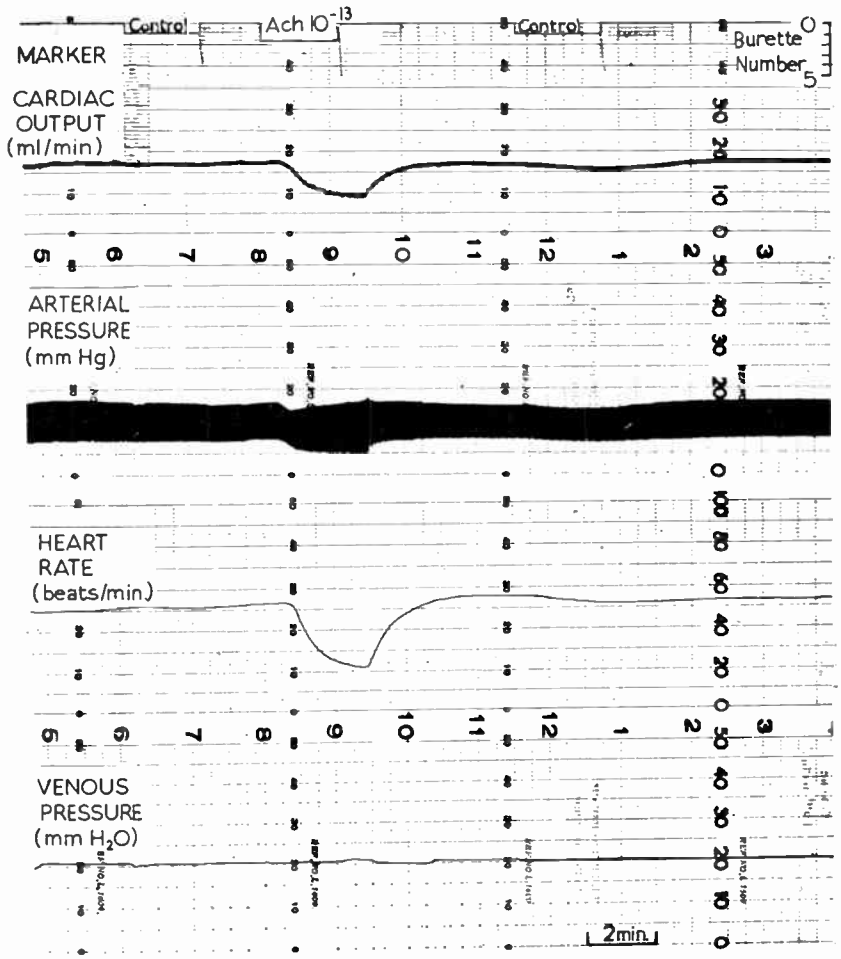


Fig. 13.

Response of a very sensitive frog heart to perfusion with 10^{-13} g/ml acetylcholine (Ach) in Ringer's solution, as indicated by marker at top of trace; depth of marker step shows which burette in use. Heart perfused with Ringer's solution from reservoir when marker in position 0; test preceded and followed by control perfusion with Ringer from burette 1, with little effect on heart. Test perfusion with Ach from burette 2 produces marked fall in heart rate, output and arterial pressure. Venous pressure tracing proves that these changes are not due to alteration in perfusion pressure.

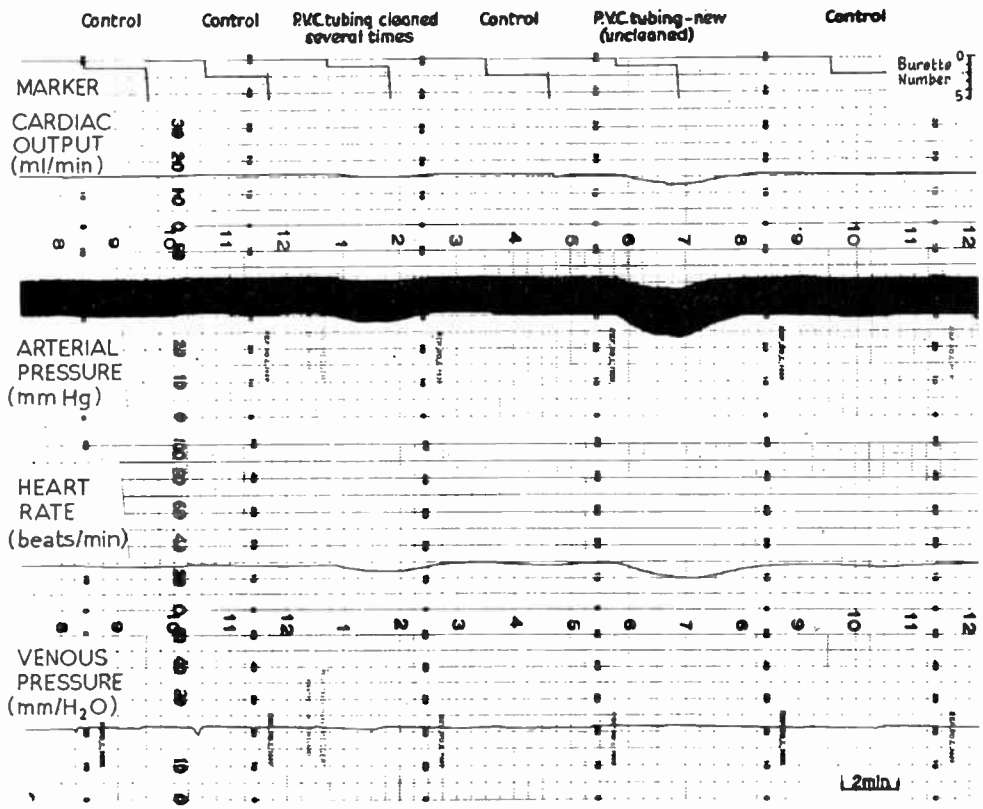


Fig. 14.

The effect on a frog heart of contaminating substances released from p.v.c. tubing. Perfusion with Ringer's solution, in which pieces of p.v.c. tubing were soaked previously for some hours, produces a fall in heart rate, output, and arterial pressure, more marked with new tubing but still occurring with tubing cleaned and sterilized several times. Heart is perfused with pure Ringer from reservoir when not perfused from a burette; little effect is produced by control perfusion with pure Ringer from burette 1 at start, and by control perfusions from burette 2 before and after test perfusions with contaminated Ringer from burette 1. Venous pressure tracing proves that changes in heart activity are not due to changes in perfusion pressure.

stock of Ringer's solution as was used to fill burettes 0 and 1. After a delay of about half a minute, during which the fluid containing acetylcholine passed through the venous pressure transducer and the veins leading to the heart, the heart rate, arterial pressure and output fell sharply due to the inhibitory action of acetylcholine on the heart. About 2 minutes later perfusion with pure Ringer's solution was recommenced and the heart recovered. The lowest trace on the chart shows that the venous pressure altered little during the test, so the effect was not due to a change in this pressure. A second control perfusion was carried out after the Ach test. A slight inhibitory effect occurred due probably to the fact that traces of acetylcholine were still present in the venous pressure manometer capsule, and were washed into the heart during the control perfusion.

This heart was very sensitive to the action of acetylcholine, more so than are most frog hearts. The acetylcholine solution is prepared by dissolving the solid in Ringer's solution and then diluting this with more Ringer's solution in steps of 1 in 100 by volume, until the required concentration is obtained. The extent of the dilution can be appreciated better, perhaps, if it is realized that a 10^{-13} g/ml solution would be obtained, approximately, by dissolving 1 g of acetylcholine in a square pond with sides half a mile long and a depth of six and a half feet!

4.3. *The Release of Contaminating Substances from Synthetic Tubing*

Many types of plastic or rubber tubing have been found to release substances into solutions passing through them, which affect the activity of a heart preparation. Part of the tracing from an experiment to test the toxic effect of p.v.c. tubing is shown in Fig. 14. Short lengths of p.v.c. tubing were allowed to lie for some hours in Ringer's solution. The same

quantity of Ringer's solution was left for the same time in an identical flask, but without the tubing in it. The Ringer's solution from the flask which had the tubing in it was the "test" solution. The other sample was the "control" solution.

Initially, the heart was perfused with pure Ringer's solution from burette 0, and then a short period of perfusion with control solution was carried out from each of burettes 1 and 2 in turn. The activity of the heart was little affected. Burette 1 was then filled with a test solution; in this case the p.v.c. tubing had been cleaned and sterilized several times before being soaked in the Ringer's solution. Perfusion with this test solution caused the heart rate, arterial pressure and output to fall. About 3 minutes later a control perfusion from burette 2 was satisfactory; the behaviour of the heart did not alter significantly. Then a second test solution was put into burette 1; in this case the p.v.c. tubing, which was soaked in the Ringer's solution, was new and had not been cleaned. This solution produced marked inhibition of the activity of the heart. A subsequent control perfusion was satisfactory, though the heart rate did fall very slightly.

Obviously the toxic property of the p.v.c. tubing has been exaggerated in this experiment by soaking it for some time in Ringer's solution, a situation which would not occur in practice. Nevertheless, the toxic effect is sufficient to interfere with any experiment in which dilute solutions of active agents reach a biological preparation by flowing through p.v.c. tubing, especially if the fluid is stationary in the tubing for more than a few minutes at any point in the experiment. Silicone rubber tubing, which has been cleaned and sterilized several times, has been found to be the most suitable tubing for use in this type of experiment.

5. Acknowledgments

The authors express their thanks to the Trustees of the Boyd Medical Research Institute for providing the space and facilities for this work, and to Dr. C. L. Pathak for his co-operation in some of the experiments from which recordings are reproduced. Grateful acknowledgment is made for financial support from the Glasgow Homoeopathic Hospital (Endowment Fund), the Homoeopathic Research and Educational Trust, and the B.H.A. Beit Research Fund.

6. References

1. I. A. Boyd and A. M. Mackay, "The blood pressure of the frog", *J. Physiol.*, **139**, No. 2, pp. 11–12*P*, December 1957.
2. J. G. Thomason, "Multi-valve cathode follower circuits—2", *Wireless World*, **63**, No. 8, pp. 373–377, August 1957.
3. I. A. Boyd and W. R. Eadie, "Improving the response of a recording galvanometer", *Electronic Engng*. To be published.
4. I. A. Boyd and W. R. Eadie, "The dynamics of the frog heart", *J. Physiol.*, **144**, No. 1, pp. 8–9*P*, November 1958.

Manuscript first received by the Institution on 9th December 1960 and in final form on 26th May 1961. (Paper No. 683.)

Radio Engineering Overseas . . .

The following abstracts are taken from Commonwealth, European and Asian journals received by the Institution's Library. Abstracts of papers published in American journals are not included because they are available in many other publications. Members who wish to consult any of the papers quoted should apply to the Librarian, giving full bibliographical details, i.e. title, author, journal and date, of the paper required. All papers are in the language of the country of origin of the journal unless otherwise stated. Translations cannot be supplied. Information on translating services will be found in the Institution publication "Library Services and Technical Information".

SEMI-CONDUCTOR PHOTOCELLS

An analysis is given in a recent Australian paper of the lateral photovoltage in the plane of a non-uniformly illuminated semi-conductor junction and of its use in photocells which indicate the position of a small area of illumination by visible or infrared radiation. It is shown that an important consequence of the presence of a heavily doped p^+-region contiguous with an n-region is the enhancement by orders of magnitude of the effective lateral diffusion length for holes in the n-region. Reference is also made to a twin-junction photocell made by alloying to one side of a thin slice of semi-conductor. This device may also be used to indicate the position of a light spot, and a comparison of its output and sensitivity to small movements of the light spot with that of the lateral photovoltage diode shows it to be superior under some conditions. Experimental observations are in good agreement with theory.

"Semi-conductor junctions as positional indicators of radiation", L. W. Davies. *Proceedings of the Institution of Radio Engineers Australia*, **22**, No. 8, pp. 509–12, August 1961.

MILLIMETRIC STANDING WAVE INDICATION

The theory and design of a new v.s.w.r. measuring line has been described by a German engineer. This measuring line can be used for measurements of the magnitude and the phase of reflection coefficients on certain types of surface wave lines (dielectric line, dielectric image line). The equipment operates like a quasi-optical device and for this reason it is particularly suitable for very short waves (millimetric waves) and in principle can be used for wavelengths down to the optical region. The probe, which travels parallel to the line, consists of two crossed wire grids penetrating the whole of the electro-magnetic field of the surface wave. Experiments have been made at a wavelength of 5 mm. In two designs for the dielectric line and the dielectric image line the residual v.s.w.r. was approx. 1·02 and the maximum measurable v.s.w.r. was approx. 1000.

"A new standing wave indicator for dielectric surface wave lines", G. Schulten. *Nachrichtentechnische Zeitschrift*, **14**, pp. 445–48, September 1961.

REMOTE CONTROL EQUIPMENT

The remote control of mobile manipulators for radioactive work is a difficult problem but a new approach has been developed for the French Atomic Energy Commission. This employs a single wire remote control technique. The article also describes automatic control circuits which operate the various functions.

"Remote control and operation of manipulating equipment in a radioactive area", J. Fuzellier and B. Boulenger. *Onde Electrique*, **41**, pp. 714–727, September 1961.

NOISE GENERATION AT MILLIMETRIC WAVELENGTHS

The positive column of a discharge in an inert gas, provided it is properly dimensioned, is eminently suitable as a noise source for measuring purposes. A method commonly used for centimetre waves is to pass the discharge tube obliquely through a waveguide to couple the noise source to the circuit. At wavelengths shorter than 7 mm, however, the waveguides are too small to make this system practical. The design described by two Dutch engineers avoids this difficulty. The discharge (in neon) takes place in a quartz-glass tube mounted *axially* in a circular waveguide, the latter acting as the anode. The noise energy is delivered to the circuit through a mica window. In the experimental generator for the 4 mm waveband the noise temperature is 21 000°K (maximum measuring error $\pm$ 1000°K) and the standing-wave ratio between 65 and 76 Gc/s has a maximum value of 1·8. It is suggested that with slight modifications it will be possible to make this type of noise source suitable for even shorter wavelengths.

"An experimental noise generator for millimetre waves", P. A. H. Hart and G. H. Plantinga. *Philips Technical Review*, **22**, No. 12, pp. 391–2, 1960–61. (In English.)

LONG SLOT ANTENNAS

An investigation by a German engineer of rectangular waveguides, in the narrow side of which has been cut a long slot, produces the following picture: the wave guided through the waveguide is partly radiated into the surrounding space, depending on the width of the slot, and as is to be expected, the radiated portion increases as the slot widened. However, in addition to the width of the slot the wall thickness also has an effect. When no additional precautions are taken this process has superimposed on it the excitation of a wave which is propagated with the velocity of light along the slot, termed a slot wave. This slot wave can easily be detected particularly in the near zone of the slot. A special shape of the slot is mentioned on which only a weak slot wave is generated and this permits a relationship between the width of the slot and the magnitude of the radiation between the width of the slot and the magnitude of the radiation to be derived. This again permits the calculation of the slot shape required for a given amplitude distribution. Since the distant radiation field is determined by the amplitude distribution (together with the phasing conditions) this constitutes a method for the realization of predetermined radiation patterns by means of long slot antennas. Two examples are given.

"Long slot antennas", Th. Heller. *Nachrichtentechnische Zeitschrift*, **14**, No. 9, pp. 441–44, September 1961.