VOL. 16 No. 1, pp. 1-32 JULY 1954

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

THE PRODUCTION OF RADIO-ISOTOPES

by A. H. W. ATEN *) and J. HALBERSTADT **).

539.167.3

The use of artificially prepared radioactive elements (radio-isotopes) for biological and medical purposes as well as in industry and in many branches of scientific research is rapidly gaining in importance. Radio-isotopes are prepared either in a nuclear reactor or by means of a particle accelerator; the following article deals with the production of isotopes with special reference to the products obtained from the Philips synchrocyclotron at Amsterdam.

The artificial production of hitherto unknown and usually radioactive isotopes of most of the elements has led to the development of a valuable means of investigation in many fields, viz. "indicator" or "tracer" techniques, details of which were described some years ago in this Review [1]). In medicine, too, radio-isotopes have become very important for diagnostic and therapeutic purposes. The enormous demand that has arisen for all kinds of radio-isotopes is met by production in a number of world centres with the aid of nuclear reactors or particle accelerators.

The purpose of this article is to describe the methods used in producing artificial radioactive substances in the Philips synchrocyclotron at Amsterdam [2]). Some general remarks will be followed by a discussion of the nuclear reactions commonly employed in the production of radio-isotopes and the methods of irradiation (design of the target). In conclusion, a description will be given of the chemical methods for the separation of the isotopes in the pure state.

General characteristics of nuclear reactions [3])

Radio-isotopes are prepared by nuclear reactions, stable nuclei being transformed by exposure to a beam of "projectiles" in the form of fast, light nuclei. The projectiles mostly used are alpha particles ($_2He^4$), deuterons (nuclei of the hydrogen isotope of atomic weight 2, i.e. $_1H^2$, usually written $_1D^2$), and neutrons ($_0n^1$). Deuterons and alpha particles are produced in the cyclotron. Alpha particles are also obtained from radioactive substances, but not in quantities which are economic. Neutrons are obtained mainly from nuclear reactors.

It is possible to visualize the reaction between the projectile and the nucleus in the first instance as a "melting", resulting in a "compound nucleus" with a very high energy content (kinetic energy of the projectile + binding energy); such a nucleus is very unstable and therefore immediately disintegrates with the emission of one or more particles.

Neutrons play a very important role as projectiles. In the first place, even slowly moving neutrons are easily absorbed by nuclei, since they are not electrostatically repelled. In this case the unstable compound nucleus does not as a rule disintegrate but emits a gamma quantum, and drops into a more stable state which is, however, usually radio-active.

When fast or slow neutrons strike a heavy nucleus such as that of uranium or thorium, something rather different may take place, viz. nuclear fission.

*) Institute for Nuclear Physics, Amsterdam.
**) Isotope Laboratory, N.V. Philips Roxane, Amsterdam.

1) A. H. W. Aten and F. A. Heyn, The use of isotopes as tracers, Philips tech. Rev. 8, 296-303, 1946; The technique of investigation with radioactive and stable isotopes, Philips tech. Rev. 8, 330-336, 1946.
2) For a description of this cyclotron see Philips tech. Rev. 12, 241-256 and 349-364, 1950-51 and 14, 263-279, 1952/53.
3) For a comprehensive and detailed review of the leading concepts in nuclear physics see, e.g., S. Glasstone, Source book on atomic energy, Macmillan, London, 1950.

The unstable compound nucleus does not assume a more stable form by emitting one or more lighter particles, but by division into two almost equally heavy nuclei, each of which is radioactive. Moreover, in this process of fission one or more neutrons are expelled, and this explains why the uranium pile is such a copious source of neutrons.

Most radio-isotopes are β-active, which means that their disintegration is accompanied by the emission of a positive or negative electron. Some are transformed into another element because the nucleus absorbs an outer electron from the K-level (K-capture).

When a specimen is bombarded by a quantity (p) of a certain kind of projectile, each projectile does not produce a radioactive atom; only a fraction k (<1) of the projectiles is effective and the number of radioactive atoms produced is thus $N_0 = kp$. The radioactivity (number of disintegrations per second) of these atoms is $N_0 \ln 2/T_{\frac{1}{2}}$, where $T_{\frac{1}{2}}$ denotes the half-life of the particular kind of atom in seconds. As p particles each having a charge of Ze represent an electrical charge of pZe, the yield in activity is accordingly:

$$\frac{N_0 \ln 2/T_{\frac{1}{2}}}{pZe}, \quad \ldots\ldots \quad (1)$$

or, expressed in microcuries per microampere-hour [4]): $4\times10^{11}k/ZT_{\frac{1}{2}}$. In practice, the value of k lies between 0·0001 and 0·001.

Another important quantity often required is the activity of a given isotope per gram of the material. If the specimen contained only the active atoms (quantity N_0, atomic weight A), the weight of the specimen would be $(N_0/N_A)A$, where N_A is Avogadro's number (6×10^{23} atoms per gram atom). The activity is $N_0 \log 2/T_{\frac{1}{2}}$ and hence the mass per unit activity in milligrams per millicurie is:

$$(N_0/N_A)A/N_0 \ln 2/T_{\frac{1}{2}} \approx 10^{-13} A\, T_{\frac{1}{2}} \text{ mg/mc} \quad (2)$$

This is a measure of the specific activity of the *pure* radioactive isotope. For a *diluted* isotope, i.e. when the radioactive atoms are mixed with stable atoms, this quantity must be multiplied by the appropriate dilution factor. In the case of dilution by N_1 stable atoms of the same element, this factor is $(N_1+N_0)/N_0$.

4) The curie was originally a measure of a quantity of radon gas, viz. the quantity that is in equilibrium with 1 gram of radium. It is now taken to be the quantity of a radioactive substance exhibiting 3.7×10^{10} disintegrations per second.

Review of some useful nuclear reactions

Reactions with neutrons

When bombardment takes place with neutrons, e.g. in the production of radioactive sodium,

$$_0n^1 + {}_{11}Na^{23} \rightarrow {}_{11}Na^{24} + \gamma,$$

it is a disadvantage that the radioactive product is obtained among a large quantity of inactive material from which it cannot be chemically separated. Against this drawback, neutron bombardment has the advantage that the "absorption cross-section" (the chance of the neutron striking the material being actually captured by a nucleus) is often very high, especially in the case of the slow thermal neutrons. In addition to this, if a powerful source of thermal neutrons is available (e.g., a nuclear reactor) some very active specimens can be obtained. Thus gold for example can be activated to a strength of several hundred millicuries per gram. In the production of Sr^{89}, on the other hand, only a weakly active material is obtained, the strength being not much more than a few microcuries per gram.

In a few instances, owing to a fortunate circumstance, it is nevertheless possible to achieve complete separation of an isotope during neutron bombardment. The gamma quantum (of frequency ν) ejected when an atomic nucleus captures a neutron possesses not only the energy $h\nu$, but also a momentum $h\nu/c$ (h is Planck's constant and c the velocity of light). In accordance with the law of the conservation of momentum the nucleus (of mass m) emitting this quantum will have a velocity v in the opposite direction, such that $mv = h\nu/c$. The corresponding kinetic energy $\frac{1}{2}mv^2$ is usually 10 to 100 times greater than that of the chemical bond, so that the bombarded atom is wrenched from any other atom with which it may be in chemical combination (Szilard-Chalmers effect). For example when ethyl iodide is activated with neutrons, the radioactive free iodine can be separated from the bombarded liquid almost entirely free from non-radioactive iodine content (carrier-free).

In certain cases the required radio-isotope is not obtained direct, but occurs as a radioactive product of an intermediate unstable atom. This applies to the production of radio-active $_{53}I^{131}$ obtained by the neutron bombardment of $_{52}Te^{130}$ in a nuclear reactor. Here the primary process produces radio-active $_{52}Te^{131}$ (half-life 25 min). As a result of β-emission, the $_{52}Te^{131}$ is converted into the radioactive $_{53}I^{131}$:

$$_0n^1 + {}_{52}Te^{130} \rightarrow {}_{52}Te^{131}, \qquad {}_{52}Te^{131} \xrightarrow[25\text{ min}]{\beta^-} {}_{53}I^{131}.$$

the half-life of which is 8 days:

$$_{53}I^{131} \xrightarrow[8\text{ days}]{\beta^-} {}_{54}Xe^{131}.$$

Some hours after bombardment the tellurium thus contains a quantity of radioactive I. It will be clear that such a method can only be useful in cases where the half-life of the daughter isotope is long compared with that of the parent isotope.

Many other reactions are possible with neutrons in addition to that just mentioned. The following may be given as examples:

$$_0n^1 + {}_{17}Cl^{35} \rightarrow {}_{15}P^{32} + {}_2He^4,$$
$$_0n^1 + {}_{17}Cl^{35} \rightarrow {}_{16}S^{35} + {}_1H^1.$$

Neutrons with higher energies are required for these processes (roughly 1 MeV), which can also be obtained in a nuclear reactor. Another example is:

$$_0n^1 + {}_{16}S^{32} \rightarrow {}_{15}P^{32} + {}_1H^1.$$

For reactions involving even faster neutrons ($>$ 1 MeV) it is necessary to use a particle accelerator. Neutrons are then produced as by-products from the bombardment of targets by deuterons or alpha particles. The energy of such neutrons is greater than the maximum energy of the neutrons produced in the nuclear reactor. Some examples of reactions obtained with high-velocity neutrons are:

$$_0n^1 + {}_{27}Co^{59} \rightarrow {}_{26}Fe^{59} + {}_1H^1,$$
$$_0n^1 + {}_{13}Al^{27} \rightarrow {}_{11}Na^{24} + {}_2He^4,$$
$$_0n^1 + {}_{11}Na^{23} \rightarrow {}_{11}Na^{22} + 2\,{}_0n^1.$$

Bombardment with neutrons in the cyclotron is invariably a secondary process and the yield is relatively small. The preparation of thermal neutrons (energy $<$1 eV) from the high-velocity neutrons produced in the cyclotron entails the disadvantage that the neutrons must be passed through a fairly thick layer of paraffin wax which has to be located outside the cyclotron. Consequently the current density of the thermal neutrons is low in the case of the cyclotron. This difficulty is less important for bombardment in the cyclotron with fast neutrons; the specimen to be irradiated can then be introduced into the acceleration chamber, in the region of the target.

Summarising, to obtain highly radio-active material by bombardment with thermal or low-energy neutrons ($<$1 MeV) it is preferable to make use of the nuclear reactor. Less active specimens can also be obtained from the cyclotron. Bombardment in the cyclotron has the advantage that the slow neutrons outside the cyclotron chamber are entirely free from other particles; it is then known that these are the only radiations involved.

Most products obtained by bombardment with neutrons can also be produced by direct bombardment with deuterons.

Reactions with deuterons

Bombardment with deuterons of relatively low energy gives reactions of the following kind:

$$_1D^2 + {}_{11}Na^{23} \rightarrow {}_{11}Na^{24} + {}_1H^1$$
or: $$_1D^2 + {}_{30}Zn^{66} \rightarrow {}_{31}Ga^{67} + {}_0n^1.$$

The first of these is equivalent to the simple absorption of a neutron. With higher deuteron energy ($>$10 MeV) another kind of reaction becomes possible, viz.

$$_1D^2 + {}_{29}Cu^{65} \rightarrow {}_{30}Zn^{65} + 2\,{}_0n^1.$$

At still higher energies the expulsion of more particles, e.g. 3 neutrons might be anticipated. Such reactions cannot be easily observed however, for they are always accompanied by other reactions, and also because the final product is often a stable isotope. Nevertheless, there are certain instances in which the process may be observed, as for example in the reaction:

$$_1D^2 + {}_{92}U^{235} \rightarrow {}_{93}Np^{234} + 3\,{}_0n^1.$$

Particularly with nuclei of not too high atomic number, various additional reactions occur with high deuteron energies, which are relevant to the preparation of radio-isotopes, e.g.

$$_1D^2 + {}_{11}Na^{23} \rightarrow {}_{11}Na^{22} + {}_1D^2 + {}_0n^1,$$
$$_1D^2 + {}_{27}Co^{59} \rightarrow {}_{26}Fe^{59} + 2\,{}_1H^1,$$
$$_1D^2 + {}_{13}Al^{27} \rightarrow {}_{11}Na^{24} + {}_2He^4 + {}_1H^1,$$
$$_1D^2 + {}_{12}Mg^{24} \rightarrow {}_{11}Na^{22} + {}_2He^4,$$
$$_1D^2 + {}_{30}Zn^{64} \rightarrow {}_{27}Co^{58} + 2\,{}_2He^4.$$

It is worth noting that the first three of these reactions yield the same result as absorption of a neutron followed by emission of two neutrons, a proton or an alpha particle. The same products can therefore be obtained by bombardment with high-energy neutrons. However, as it is easier to project a large number of deuterons on to a specimen than a large number of high-energy neutrons, preference is usually given to bombardment with deuterons.

Reactions with alpha particles

Bombardment with alpha particles is used for preparative processes only in exceptional cases. In certain instances this technique has definite possibilities, however, as in the preparation of the most widely used isotope of astatine (At):

$$_2He^4 + {}_{83}Bi^{209} \rightarrow {}_{85}At^{211} + 2\,{}_0n^1.$$

Nuclear fission

We have already seen that the absorption of a neutron by heavy nuclei can result in nuclear fission. In the case of U^{235} the absorption of a thermal neutron is sufficient to effect this. With other nuclei such as U^{238} and Th^{232} neutrons of high kinetic energy are necessary. Nuclear fission is employed for isotope preparation in cyclotrons only in isolated cases such as in the bombardment of metallic thorium with deuterons. This is because the fission process can take place in several different ways resulting in a large number of different isotopes being produced simultaneously. Consequently the radioactivity of each type of atom is, in itself, relatively low. At the same time, there are several isotopes which can be produced only by nuclear fission.

Nuclear reactions employed in the Philips cyclotron

In the production of a given isotope in the cyclotron, that nuclear reaction should be selected which will give the maximum yield. It is also necessary to know what unwanted isotopes are likely to occur as a result of associated reactions.

The Philips synchrocyclotron is capable of accelerating deuterons up to an energy of 30 MeV and alpha particles up to 60 MeV; this is high enough to produce quite a large number of nuclear reactions. At the same time, these energies are not so high that subsidiary reactions occur which are difficult to control, viz. those in which many nuclear particles leave the bombarded nucleus. This simplifies the problem of separating the required product in the pure state. It has already been mentioned that with one or two exceptions, bombardment with alpha particles is not a very satisfactory method of producing radio-isotopes. Most preparative reactions are therefore effected by means of accelerated deuterons in accordance with the examples given above.

The deuteron reaction:

$$_1D^2 + {}_ZQ^A \rightarrow {}_ZQ^{A+1} + {}_1H^1,$$

in which $_ZQ^A$ is an isotope of mass number A and atomic number Z — this is also abbreviated to $(_1D^2, {}_1H^1)$ reaction [5]) — is rarely if ever employed in the Amsterdam cyclotron for the production of radio-isotopes for medical or biological applications, although this reaction occurs with every bombardment. There are two reasons why the $(_1D^2, {}_1H^1)$ reaction is not used. Firstly, it gives the same results as the $(_0n^1, \gamma)$ reaction in a nuclear reactor; and, although at the present time the yield from the $(_0n^1, \gamma)$ reaction in most nuclear reactors per unit time and per gram of basic material is roughly equal to that in the $(_1D^2, {}_1H^1)$ reaction in the Amsterdam cyclotron, preference is usually given to activation in the reactor since the latter can irradiate a greater quantity of material per charge for longer times and at a lower cost.

The more important reason for this preference, however, is that with bombardment in the cyclotron, using deuterons at 30 MeV, the $(_1D^2, {}_1H^1)$ reaction is accompanied by the $(_1D^2, {}_1D^2\ {}_0n^1)$ reaction, with a probability of not less than 1 : 10 with respect to the former. It is thus impossible, for example, to produce pure Na^{24} from Na^{23}, because relatively large quantities of Na^{22} occur [6]), in accordance with:

$$_1D^2 + {}_{11}Na^{23} \rightarrow {}_{11}Na^{22} + {}_1D^2 + {}_0n^1.$$

Fortunately, however, the probability of the following reaction:

$$_1D^2 + {}_{13}Al^{27} \rightarrow {}_{11}Na^{24} + {}_1H^1 + {}_2He^4$$

is fairly high, viz. 1 : 3 with respect to the reaction:

$$_1D^2 + {}^{11}Na^{23} \rightarrow {}_{11}Na^{24} + {}_1H^1,$$

so that it is possible, by irradiating aluminium, to produce large quantities of Na^{24} which are radiochemically pure and, moreover, carrier-free. The aluminium, however, must be entirely free from traces of magnesium, as this again produces unwanted Na^{22}.

Table I shows the half-life, target material and type of reaction as well as the yield in μc per μAh for a number of radio-isotopes. As the average beam intensity generally employed in the Amsterdam cyclotron is 30 μA (a value of 40 μA is now attainable), the figures given for the yield indicate the radio-activity obtained from 2 minutes bombardment. The activity of those radio-isotopes which disintegrate by K-electron capture and subsequent emission of X-rays of relatively long wavelength, was measured with a Philips Geiger-Muller X-ray counter tube [7]). This is a sensitive and accurate instrument by means of which absolute activity measurements of the radiation can be made.

[5]) Or, in general, the reaction designated *a*, *b* means that *a* is the projectile and *b* the particle(s) ejected by the compound nucleus.

[6]) The formation of Na^{22} together with Na^{24} is undesirable because the co-existence of two active forms of the element having different half-lives makes many experiments difficult to interprete. Moreover, the longer life of Na^{22} is a disadvantage in medical applications.

[7]) Philips tech. Rev. 13, 282, 1951/52.

Table I [8]). Yield of radio-isotopes from deuteron bombardment at 30 MeV.

Isotope	Half-life	Radiation	Target material	Reaction	$\frac{\mu c}{\mu Ah}$
Be^{7}	53 d	K, γ	$LiBO_2$*	d, 2n	150
Na^{22}	2.6 y	$\beta+$, γ	Mg	d, α	2
Na^{24}	15 h	$\beta-$, γ	Al	d, pα	2200
Mg^{27}	9.4 m	$\beta-$, γ	Al	d, 2p	48000
P^{32}	14 d	$\beta-$, γ	FeP*	d, p	400
V^{48}	16 d	K, γ	Ti	d, 2n	350
Cr^{51}	26 d	K, γ	V	d, 2n	280
Mn^{52}	6 d	K, $\beta+$, γ	Cr	d, 2n	400
Mn^{54}	310 d	K, γ	Fe	d, α	5
Fe^{55}	3 y	K	Mn-Cu	d, 2n	10
Fe^{59}	47 d	$\beta-$, γ	Co	d, 2p	2
$Co^{56(57)}$	80d(270d)	K, $\beta+$, γ	Fe	d, 2n(n)	39(10)
Co^{58}	72 d	K, $\beta+$, γ	Zn	d, 2α	0.04
Co^{60}	5.3 y	$\beta-$, γ	Co	d, p	10
Co^{60}	5.3 y	$\beta-$, γ	Cu	d, pα	0.17
Cu^{64}	13 h	K, $\beta+$, $\beta-$, γ	Cu	d, p	23000
Cu^{64}	13 h	K, $\beta+$, $\beta-$, γ	Zn	d, 2p d, α	1200
Cu^{67}	60 h	$\beta-$	Zn	d, 2p	1.5
Zn^{56}	250 d	K, $\beta+$, γ	Cu	d, 2n	20
Ga^{66}	9.4 h	K, $\beta+$, γ	Zn	d, 2n	3600
Ga^{67}	78 h	K, γ	Zn	d, n	700
As^{74}	17 d	$\beta+$, $\beta-$, γ	Ge	d, 2n	80
Br^{82}	35 h	$\beta-$, γ	KBr* K_2CO_3*	d, p	235
Rb^{86}	19 d	$\beta-$, γ	Sr	d, α	20
Sr^{89}	54 d	$\beta-$	Sr	d, p	5
Y^{88}	105 d	K, γ	Sr	d, 2n	55
Cd^{109}	470 d	K, γ	Ag	d, 2n	2.7
In^{114}	50 d	$\beta-$, γ	Cd	d, 2n	16
Au^{198}	65 h	$\beta-$, γ	Au	d, p	1000
Bi^{206}	6.4 d	K, γ	Pb	d, 2n	850

Table I includes a few ($_1D^2$ $_1H^1$) reactions — that is, reactions not producing transmutations — in order to allow of comparison between the relevant production capacity of the Amsterdam cyclotron and that of other cyclotrons and equivalent ($_0n^1$, γ) reactions in nuclear reactors. However, the production of radio-isotopes by transmutation reactions is to be regarded as the special field of application of the cyclotron

Types of target for bombardment in the cyclotron

In the cyclotron the deuterons and alpha particles originate in the centre of the acceleration chamber and, under the influence of the combined magnetic and electric field describe spiral paths under constantly increasing energy until they finally reach the wall of the cylinder. The target carrying the specimen to be bombarded is placed close to this wall.

In direct bombardment with a beam of deuterons the important problem is that of preventing the specimen from melting and evaporating. At the above-mentioned energy of 30 MeV and the average beam current of 30 μA, 900 watts are continuously being converted into heat at the target. Owing to the extremely high vacuum in the chamber of the cyclotron, appreciable evaporation takes place even at relatively low temperatures when the vapour pressure of the material is still low; moreover, in consequence of the vacuum, no cooling occurs by heat transfer to surrounding gas molecules. In many cases, therefore, there is no alternative but to reduce the strength of the deuteron beam in order to avoid damaging the target [9]).

When alpha particles are employed, the cyclotron current is usually lower than with deuterons, and cooling of the target does not present so much difficulty. This lower current, however, is just the reason why bombardment with alpha particles is not so suitable for the preparation of highly active specimens in the cyclotron. The simplest operation is of course the bombardment of materials which are capable of with standing very high temperatures such as tungsten and molybdenum, as these can be cooled by heat radiation. Such cases are, however, exceptional.

A very good and widely used method of cooling consists in soldering the metal plate to be bombarded on to a copper tube which is cooled with running water. Because of the high temperature it is preferable to use silver solder for this purpose and, provided that the soldering is done with care, the heat transfer is quite satisfactory. Metals irradiated in the cyclotron in this way are gold, silver and platinum. Metals which are too brittle, such as mangenese, are not soldered direct to the cooling pipe, but to a copper plate which is in turn silver-soldered to the cooling pipe (*fig. 1*). Low

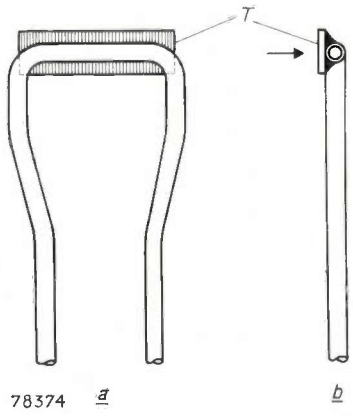


Fig. 1. Target *T* soldered to a copper cooling tube. *a*) back view; *b*) cross-section; the arrow shows the direction in which the ion beam strikes the plate.

[8]) The letter K in the column "radiation" indicates the capture of a K-shell electron. In the column "target", non-metallic substances are marked with an asterisk (see next section). The letter p in the column "reaction" denotes a proton ($_1H^1$), d a deuteron ($_1D^2$), α an alpha particle ($_2He^2$) and n a neutron ($_0n^1$).

[9]) This largely counteracts the main advantage of the classical cyclotron compared with the synchrocyclotron (viz. the continuous beam current).

melting point metals such as tin or germanium are run on to the copper plate by heating to the melting point so as to produce either a uniform layer on the plate or an alloy with it.

The methods described are not restricted to metallic elements, however. Certain non-metals which would be far too volatile for bombardment in the cyclotron can be alloyed with metals and soldered to the target in the form of plates. At Amsterdam, radioactive phosphorus is prepared in this way by bombarding a plate of iron phosphide. Similarly, tellurium can be treated by alloying it with copper and cobalt. After activation, such specimens can be separated from the underlayer by mechanical means before chemical processing is commenced.

There are of course many elements that cannot be soldered in the form of plates or alloys, e.g. the alkali metals, halogens and so on. In these cases it is often possible to produce the required element in the form of a glass; if this can be done, a thin layer, say 0.5 mm, is applied to the water-cooled copper tube, the thermal contact being then so effective that the bombardment can be done without risk of the specimen melting. An advantage of such glasslike specimens is that after bombardment they can in many cases be dissolved in water or dilute acids without the copper tube being seriously attacked and that the radioactivity of impurities in the specimen, originating in the copper, is quite low. It may be added that not only glasslike substances but also certain crystalline salts such as rubidium chloride can be melted on to a copper target.

There are many elements, however, that can be activated, neither in a glasslike form nor as an alloy, and in such cases a target is used which is also suitable for general use, viz. a perforated target. In its simplest form this consists of a metal block securely soldered to a water-cooled copper tube and drilled with a row of parallel holes along the edge. For filling, the block is placed on a closely mating plate so that the holes are closed at that end. The material to be activated, in powder form, is rammed tightly into the holes, and the charged block is then bombarded with deuterons or alpha particles in the cyclotron, after which the radioactive powder is pushed out of the holes with a steel wire probe (which is clamped in a long holder to protect the fingers from burns due to beta radiations from the metal block). Obviously this method is suitable for all compounds which are not too volatile; one disadvantage, however, is that a large part of the beam falls between the holes and the yield of the required isotope is much lower than when the whole target is made of the required material. The smaller the holes the lower the efficiency, but the better the cooling of the powdered material.

Instead of the various metal plates being soldered to the cooling tube they can also be held in a clamp with water-cooled jaws, but the cooling is then usually less efficient and a lower beam current must be employed. At the same time this system is so much easier, quicker and cheaper than the soldered blocks that it is generally given preference. It is accordingly employed for specimens soldered or melted on to target plates and more especially for the perforated type of target (*fig. 2*).

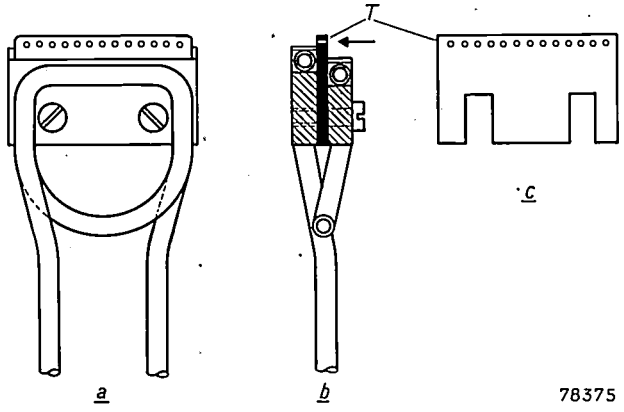


Fig. 2. Perforated target *T* held between the turns of a copper cooling tube. *a*) front view; *b*) cross-section; *c*) separate target plate. The cooling water flows successively through both copper jaws. If a metal plate is to be bombarded in order to render it radioactive the outer edge is set flush with the back copper jaw, as this ensures the most effective cooling. If a perforated target plate is to be used the holes are set just beyond the edge of the back jaw. This facilitates removal of the powder from the holes and prevents the contents of the holes from being contaminated by the face of the back jaw of a holder which is used over and over again.

Sometimes it may be desirable to effect bombardment by means of low-energy deuterons, seeing that this excludes certain types of nuclear reaction and accordingly yields some isotopes in a pure state. Variations of several MeV in the energy with which the particles hit the specimen can be obtained by varying the distance of the target from the centre of the cyclotron. Still wider variations may be produced by retarding the particles in the beam by means of a metal plate (usually of copper) placed in front of the target.

As mentioned above, neutrons are produced in all the nuclear reactions that take place in the cyclotron. Owing to the low beam current obtaining when using alpha particles and the small reaction probability per ion it is preferable when producing neutrons to employ a deuteron beam. The nature

of the target to be bombarded is of minor importance.

For bombardment over short periods with fast neutrons it is usual to employ a vacuum-tight brass tube with a thick water-cooled base, the inside being open to the outer atmosphere. The base is struck by the deuteron beam and neutrons are generated; the specimen to be activated is placed in a glass tube inside the brass tube, close to the base. The advantage of this method is that the specimen can be unloaded very quickly. If activation by slow neutrons is to be avoided the brass tube is lined with cadmium.

In cases where bombardment with fast neutrons for long periods is required, it is economical to pack the specimen in a metal box and attach this to the holder close to another specimen which has to be bombarded with deuterons; in this way use is made of the neutrons which are liberated by the other target.

Slow neutrons are produced, as already noted, by placing blocks of paraffin wax in the path of neutrons leaving the cyclotron. For reasons of geometry, the intensity of these neutrons is always low but, on the other hand, whenever the cyclotron is working these slow neutrons are available without extra trouble or cost. In the Amsterdam cyclotron, therefore, activation of a number of specimens by slow neutrons is usually in progress. It is a complication, however, that the neutron intensity is very dependent on the position outside the cyclotron. When it is necessary to activate various specimens all with the same neutron intensity a rotating paraffin wax cylinder has to be used, the specimens being arranged at equivalent points inside it.

Separation of radio-isotopes

The radio-isotopes obtained as a result of the particular nuclear reaction employed usually differ chemically from the initial material bombarded, and can therefore be separated by chemical and/or simple physical methods.

As long as no stable isotopes of the radio-elements produced are added in the form of salts (for chemical reasons), the radio-isotopes can be obtained free from carrier, that is, in the pure form. This is very important in a varity of applications, particularly in the field of medicine or biology.

In the use of radio-isotopes as tracers (see article referred to in [1]), the radioactive form of one of the components of a chemical or biological system is introduced into that system. It is a great advantage for the quantities so introduced to be effectively weightless, so that the concentration of the particular component and the equilibrium of the system are not disturbed. When it is remembered that 10^{-6} millicurie of most radio isotopes can be measured quite easily, it will be seen that "weightless" additions of these isotopes can indeed serve as tracers.

The half-life and weights per millicurie (see (2)) of a number of radio-isotopes are given in *Table II*.

Table II. Weights of radio-isotopes per millicurie.

Isotope	Half-life	Weight in 10^{-6} mg per mc.
Na^{22}	2.6 y	160
Na^{24}	15 h	0.115
P^{32}	14 d	3.5
Cr^{51}	26 d	10
Mn^{52}	6 d	2.4
Co^{56}	80 d	35
Co^{60}	5.3 y	880
Zn^{65}	250 d	124
Cs^{137}	33 y	12 500
C^{14}	appr. 6000 y	200 000
Ra^{226}	1550 y	1 000 000

The methods of separation employed for isolating carrier-free quantities of radio-isotopes are in general not the same as the ordinary processes of analytical chemistry. Each and every radio-isotope demands a unique sequence of processes and treatments. The most important processes involved are: crystallization, selective reduction, electrolysis, ion-exchange, extraction (solid-liquid and liquid-liquid), distillation, the radioactive colloid process, co-precipitation, paper-chromatography and paper-electrophoresis.

Owing to the exceptionally small concentration, carrier-free radio-isotopes in solution often behave very differently from solutions containing macro-quantities of such elements. Small concentrations, sometimes of not more than 10^{-15} mole per litre, mean that effects such as adsorption on the surfaces of glass vessels, filter papers etc. play a significant part. Thus in the course of a process, the whole quantity of the available radio-isotope may be adsorbed on the sides of a beaker and would be lost, were it not for the fact that a Geiger counter will immediately detect the radioactive area and permit of the recovery of the isotope.

These adsorption phenomena are used in separation by co-precipitation or the radio-colloid process.

It is very important to employ those methods that will yield a final product of outstanding radio-chemical purity. This is not difficult, however, as it is always known what radio-isotopes will be produced in the target specimen by the various

nuclear reactions. In this connection it is of course essential to know in advance the exact degree of purity of the target specimen, so that radio-isotopes having their origin in impurities can be taken into account.

Especially for medical and biological purposes it is essential that radioactive specimens contain no other stable or radioactive elements that may be toxic or that might give rise to undesirable reactions by reason of differences in their radiations or half-life. The addition of non-toxic substances such as sodium chloride or sodium citrate may be necessary to make solutions of radioactive materials isotonic i.e. to give them the same osmotic pressure as that of the body fluids.

An ever-present problem relating to methods of separation as well as to the use of radio-isotopes is that of protecting the operators from the radioactive radiations, but this is a complete study in itself and is beyond the scope of this article. Suffice it to say that all chemical and other manipulations must be carried out in such a way that the operator maintains a safe distance from the specimens, and that many special implements have been designed for this purpose; it is also usual for the apparatus in which the specimens are subjected to chemical processes to be placed behind a wall made of blocks of lead (fig. 4). Lastly, preference is given to methods of separation which take place as far as possible automatically. It must also be remembered that in some cases, where the isotope to be separated has a very short life, rapidity of separation is an all-important factor in the choice of method.

The separating techniques enumerated above will now be discussed in greater detail taking examples from a number of production methods employed in the isotope laboratory at Amsterdam.

Crystallization is often used in order to eliminate the greater part of the target material before final separation of the required isotope. Numerous metal chlorides and nitrates are insoluble in concentrated hydrochloric acid or nitric acid; hence, for example Rb^{86} is easily separated from irradiated strontium by first removing most of the strontium as $Sr(NO_3)_2$ with concentrated HNO_3.

Crystallization is used in particular for the separation of the short-lived carrier-free Na^{24} from irradiated aluminium. The aluminium target is first dissolved in as small as possible a quantity of 8N HCl, after which crystalline $AlCl_3 \cdot 6H_2O$ can be deposited by introducing HCl gas at 0 °C. As micro-quantities of NaCl are not precipitated, all the active material is in the filtrate, the further purification of which is quite straightforward.

A quicker and more convenient method consists in allowing the aluminium to crystallize as $Al(NO_3)_3 \cdot 9H_2O$ with the aid of 90% HNO_3. The precipitate can be filtered very rapidly and all the active Na^{24} appears quantitatively in the filtrate with very little aluminium. Further purification of the Na^{24} is then effected by the ion-exchange method. Separation by precipitation of $Al_2(SO_4)_3$ in H_2SO_4 is not possible, because the carrier-free Na_2SO_4 occurs in the precipitate, probably as the double salt.

Selective reduction is based on that property whereby metals are precipitated from solutions of their salts by other metals lower in the electrochemical series, e.g.,

$$Zn + CuSO_4 \rightarrow ZnSO_4 + Cu.$$

In this way it is also possible to eliminate most or all of the target material before final purification is commenced. Of course, an element must be used for the reduction that can be readily separated from the desired isotope by a simple process.

This method is employed for isolating Mn^{52} from irradiated chromium and Cd^{109} from silver. In the first instance the chromium target is dissolved in hydrochloric acid, the solution is neutralised to a $pH = 3$ and then boiled for about 20 minutes with an excess of zinc powder. This reduces the chromium to metal, leaving the manganese in solution. In the second example the silver target is dissolved in HNO_3, the solution is neutralised to a $pH = 3$ and then boiled for 10-15 minutes with an excess of powdered tin; the silver is precipitated and the Cd^{109} remains in solution.

In both cases the isotopes are further purified by adding ferric salts to the solution and then making the solution strongly alkaline to dissolve the zinc and tin as zincate and stannite, the Mn^{52} and Cd^{109} being co-precipitated with the $Fe(OH)_3$ deposit. This is filtered, washed and dissolved in 8N HCl, after which the $FeCl_3$ is eliminated by extraction with isopropyl ether.

Ion exchange has become a very important means of separation, even of quite complex mixtures of elements. In some instances it is the only means available, as in the separation of the rare earth elements. There are two types of ion exchange, viz. cation and anion exchanges; the latter is very useful for the separation of those elements which have the property of forming stable complexes.

Cation exchanges are used amongst other things for the isolation of Na^{22} from irradiated magnesium and Rb^{86} from strontium. In the latter instance the excess of strontium is first removed by the preci-

pitation of $Sr(NO_3)_2$ with concentrated nitric acid, after which the active yttrium Y^{88}, which is also produced, is eliminated by co-precipitation with $Fe(OH)_3$.

In both cases the next operation is to pass a solution, which has been neutralised as completely as possible, through a column of acid "Dowex-50" resin which adsorbs the cations in the upper layers. Subsequently Na^{22} and Rb^{86} are slowly eluted with 0·2N HCl. Mg, Sr and any other divalent or trivalent ions are retained by the Dowex resin.

Anion exchanges are employed for the separation of Fe^{59} from irradiated cobalt and Fe^{55} from manganese. When a solution of cobalt chloride or manganese chloride + Fe^{55} chloride in 9N HCl is passed through a "Dowex 2" anion column previously treated with 9N HCl at a speed of 0.3 ml/min, the Fe^{59} and a part of the cobalt, or the Fe^{55}, is retained by the resin as complex $FeCl_4$ and $CoCl_3$ ions. Manganese does not exhibit this property.

If 0.3 ml/min of 3-5N HCl is then passed through

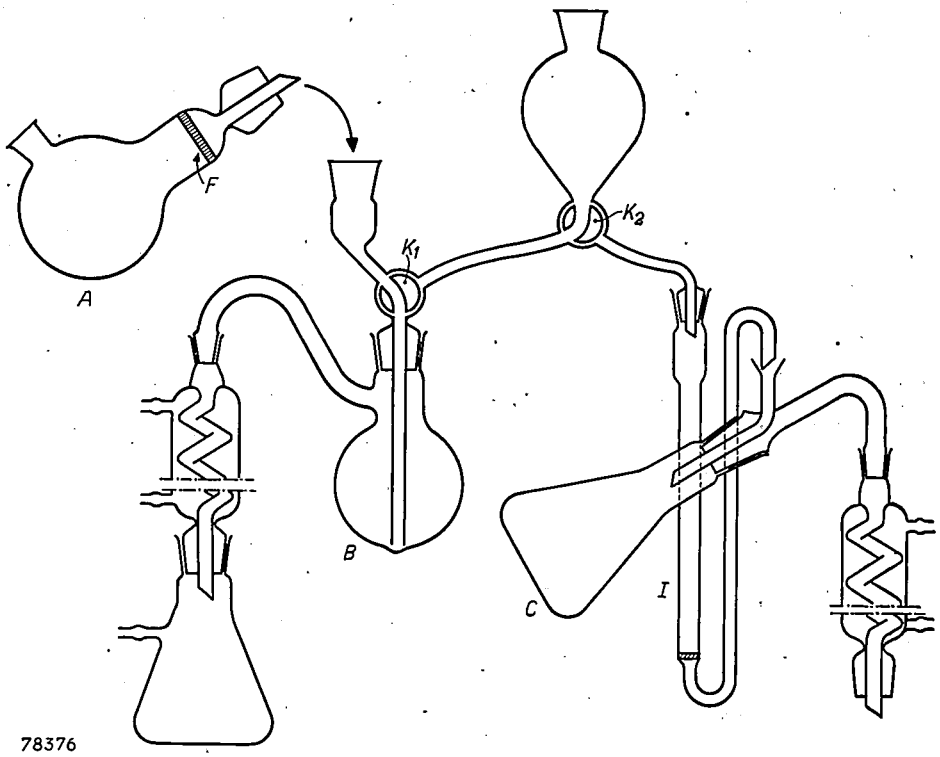


Fig. 3. Apparatus for the scparation of Na^{24} and Al by the method described in the text. *A*, precipitation and filter flask, *F* glass filter, *B* evaporating and diluting flask, *I* ion exchange tube, *C* heated rotating evaporation flask, K_1 and K_2 three-way cocks.

The direct and fairly quick separation of the short-lived Na^{24} from irradiated aluminium is also possible in accordance with this method (*fig. 3*). The Al target is dissolved in the smallest possible quantity of 8N HCl and the solution is then diluted to roughly 0·2N, after which it is passed through a column of "Dowex 50" about 60 cm long and 2.5 cm thick at the rate of 3 to 5 ml/min. Immediately afterwards 0·5N HCl is passed through the column at the same rate; when about 200 ml has flown through, elution of active Na^{24} starts and a further 500 ml is necessary to elute 98% of the Na^{24}. All the aluminium (about 2 gm) and traces of impurities such as Fe and radioactive cobalt and manganese remain in the resin.

the column, the cobalt chloride complex is broken up and $CoCl_2$ is eluted. The Fe chloride complex is more stable and remains at the top of the column. Fe^{59} or Fe^{55} can then be eluted as pure $FeCl_3$ by washing out the column with dilute HCl or water. This method also offers possibilities for other separations, especially when the desired isotope can be obtained as a cation complex and the other elements as anion complexes.

Extraction of metals as complex compounds from an aqueous solution, by means of an organic solvent immiscible in water, is one of the most widely used and convenient methods of separating radio-isotopes quickly. In this way it is possible to separate a certain radioactive element from a mixture,

usually direct and very selectively. It is also possible in this manner first to eliminate the weighable quantities of target material, leaving the radio-isotope in the layer of water.

Extraction can also be employed to remove inactive carrier elements which have been used to isolate the desired isotope by co-precipitation.

The process of extraction is greatly accelerated by the use of a vibrating agitator and, moreover, can thus be effected automatically (*fig. 4*); this reduces the danger to the operator. By means of this apparatus two non-miscible liquids can be mixed to a kind of emulsion in the separating funnel in less than one minute, the contact surface being then quite large. In most cases the liquids separate again very quickly.

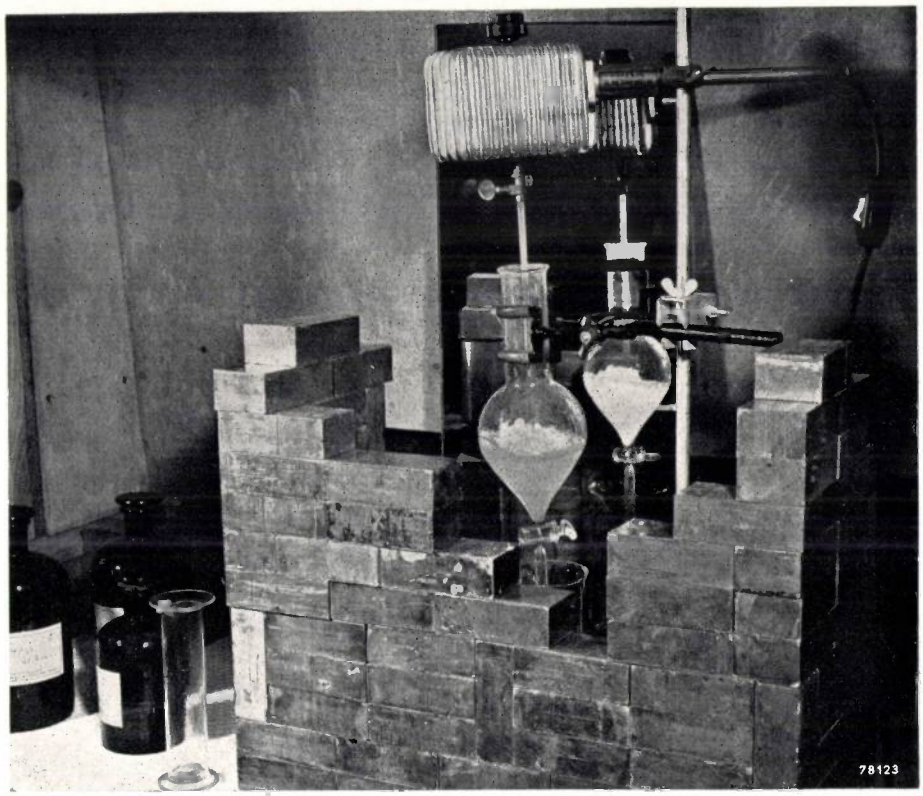


Fig. 4. Extraction apparatus. The non-miscible liquids, one of which contains the required radio-isotope in solution, are emulsified in a separating funnel by means of a vibrating agitator. This greatly accelerates the process. Below the funnel is seen the beaker in which one of the liquids is collected after the emulsion has separated out. A wall of lead blocks is built up round the apparatus to protect the operator from radioactive radiations, and a mirror placed at the back enables the process to be observed. Some of the blocks have been removed for the purposes of the photograph.

By this method Ga^{66} and Ga^{67} can be isolated direct from irradiated zinc, viz. by dissolving in 8N HCl and extracting with isopropyl ether. $GaCl_3$ then passes over into the layer of ether; zinc, Cu^{64}, Cu^{66} and Co^{58} are left in the water. After washing the ether layer with 8N HCl the radio-active gallium is re-extracted with a little water.

Radioactive gold, iron and some other elements can also be extracted in this way in the presence of hydrochloric acid.

Iron as hydroxide is often used for co-precipitation of various carrier-free isotopes; if the $Fe(OH)_3$ deposit is re-dissolved in 8N HCl and extraction then effected with isopropyl ether, the ferric ion can again be eliminated.

Certain nitrates can also be extracted from nitric acid solutions with ether.

Radio-active chromium Cr^{51} can be separated from the vanadate in the form of $Na_2Cr_2O_7$ by extraction from the hydrochloric acid solution with methyl-isobutyl ketone.

Another extraction method applied to chromium is that from a solution of chromate and vanadate in sulphuric acid by means of amyl alcohol, after the addition of H_2O_2. This produces the blue perchromate which is dissolved by the amyl alcohol

in the stable state. For reasons which are not yet understood this method fails completely if the radioactive chromium is carrier-free.

Extraction of an organo-metallic compound takes place in the isolation of Zn^{65} from irradiated copper as dithizonate in chloroform in the presence of an alkali, Co^{58} from irradiated zinc as thiocyanate in amyl alcohol with a weak acid, and Fe^{59} from irradiated cobalt as acetyl acetonate (or another β-diketonate) in xylol, again with a weak acid.

Such extractions of organo-metallic compounds, which are also carried out in ordinary analytical chemistry, make it possible to eliminate in advance the great excess of unwanted target material without loss of the required radio-isotopes.

Many radio-active elements can be separated direct in the pure state by *electrolysis*. This method is employed in the production of carrier-free Cu^{64} from irradiated zinc after preliminary removal by extraction of the highly radio-active gallium. Again, $Co^{56,57}$ can be separated by electrolysis from the Mn^{54} simultaneously produced when iron is bombarded; the excess of iron is first removed by extraction. The peculiar feature of this, however, is that after electrolysis with platinum electrodes only about 95% of the carrier-free $Co^{56,57}$ will dissolve in hydrochloric acid; the remaining 5% is firmly attached to the electrode and can be removed only by anodic solution, whereby some of the platinum is also dissolved.

Electrolysis is the appropriate method for the removal of the excess of copper from a copper target which has been irradiated for the production of Zn^{65}. For the separation of certain products mercury electrodes may also be used, in which the required radio-isotope is dissolved. The mercury is subsequently removed by distillation in vacuo.

Under certain conditions some elements yield compounds which can be separated from their original environment by *evaporation* or *distillation*; established examples of this are osmium oxide OsO_4 and ruthenium oxide RuO_4 which are separated from their solutions in concentrated nitric acid and perchloric acid respectively by distillation. In this respect carrier-free radio-isotopes behave in exactly the same way as macro-quantities of the elements.

Similarly As^{74} can be separated from irradiated germanium. A trace of common arsenic must first be added to prevent loss of the As^{74} due to adsorption on the glass. The As^{74} is first isolated from the germanium and radio-active gallium by twice co-precipitating it as arsenate with a large quantity of $MgNH_4PO_4$. Next, the $AsCl_3$ is distilled from a reducing solution of hydrochloric acid and is collected in an oxidising acid solution to convert it to pentavalent arsenic; the solution is finally evaporated to a small volume without loss of As^{74}.

Germanium, selenium and tin isotopes are also obtained from HCl or HBr solutions rapidly and in a concentrated form by distallation.

Co-precipitation is one of the most widely-employed processes in the isolation of radio-isotopes; it is based on the fact that minute quantities of radioactive elements are often selectivily carried along when a non-radioactive element is precipitated in macro-quantities from a solution.

"True", or isomorphous co-precipitations such as those of radioactive strontium ions with calcium or barium salts, or radioactive iodide ions with silver chloride, or radioactive arsenate ions with $MgNH_4PO_4$ etc are not of great value in practice. Usually, effective separation of chemically similar ions can be achieved with the help of ion exchange resins or — as just mentioned — by distillation. Such methods, however, are sometimes too difficult or take too long.

For this reason wide use is made of the fact that very small quantities of some radioactive elements exhibit considerable adsorption on materials presenting large surfaces. Thus $Fe(OH)_3$ (which can easily be formed in the solution) is an unusually good absorbent medium. Those radioactive elements which in weighable quantities are precipitated as insoluble hydroxides would, in "unweighable" quantities and under the same conditions, remain in solution; in the presence of the $Fe(OH)_3$ precipitate, however, they are adsorbed on to its surface. After filtering and dissolving the precipitate in HCl, the unwanted $Fe(OH)_3$ can be eliminated by extraction with isopropyl ether.

This system is employed in the final purification of Mn^{52}, Mn^{54} and Cd^{109} obtained from irradiated chromium, iron and silver respectively, as well as for the direct isolation of Y^{88} from irradiated strontium and In^{114} from cadmium.

Radio-colloids are sometimes formed by radio-active elements which under normal conditions and in macro-quantities form insoluble compounds, as do most of the hydroxides. As previously mentioned, such colloids are capable of considerable absorption by various materials such as glass, filter paper, etc.

This fact is utilised in the isolation of Be^{7} from from irradiated $LiBO_2$, of Mg^{27} from aluminium and Bi^{206} from lead. In the first of these the target material is dissolved in water and the alkaline solution is drawn through a filter of sintered glass

by suction. In the case of Mg^{27} the aluminium is dissolved in a base, the solution being then drawn through filter paper. With Bi^{206} the lead is first dissolved in nitric acid and the solution decanted into an excess of a base. The resultant plumbate solution is then drawn through filter paper.

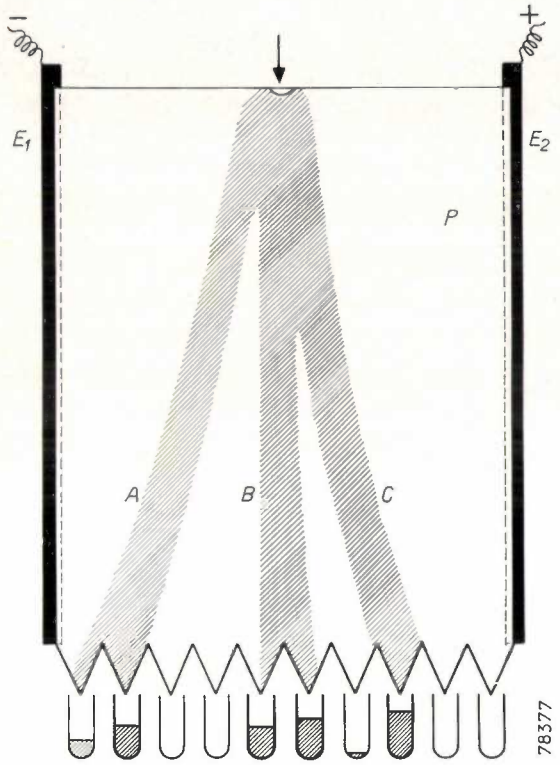


Fig. 5. Paper-electrophoresis arranged as a continuous process. Filter paper P is held vertically between two electrodes E_1 and E_2. The solution containing the various ions (A, B, C) is fed in at the top and the ions (still in solution) are collected separately at the bottom end.

The radioactive colloids Be^7, Mg^{27} and Bi^{206} are adsorbed by the glass or filter paper to the extent of 95-100%. They are subsequently separated from the filter by treating it with a few drops of hydrochloric acid followed by a little water.

For the separation or purification of carrier-free radio-isotopes which by other methods present difficulties, two further methods may be mentioned. These are *paper-chromatography* and *paper-electrophoresis* (*fig. 5*), the latter an especially elegant and selective method.

A solution of the carrier-free radioactive element, mixed with an eluting medium such as a weak solution of hydrochloric or lactic acid, is allowed to drip on to the centre of the top edge of a sheet of filter paper which is moistened with the eluting fluid and clamped between two sheets of glass so that the descending solution runs through the paper and not along it. Electrodes are fitted along the whole length of the sides of the paper and a certain potential difference is maintained between them. The ions do not all pass downwards at the same speed; moreover they are drawn outwards to the left and right by the electric field. In this way the individual radio-isotopes are collected in a very pure state at the bottom. The great advantage is that the whole system can be made to operate continuously and automatically.

By the process of ordinary paper-chromatography, and notably by the "ascending" method, mixtures of $Fe^{55,59}$, $Co^{56,57}$ and Mn^{54} can be effectively and sharply separated.

Summary. Some introductory remarks concerning nuclear reactions are followed by a comprehensive review of the various methods of producing radio-isotopes. The nuclear reactions used in the cyclotron at Amsterdam are then considered in detail, with a survey of the strength of some of the products thus obtained. Some information is also given about the design of the target employed at Amsterdam. In conclusion the various methods adopted for separating the radio-isotopes in the pure state are considered in detail.

ELECTRONIC FLASH-TUBES

by N. W. ROBINSON *).

621.327.4:771.447.4

*Gas discharge or "electronic" flash-tubes have in recent years found widening fields of application. They have largely replaced the air-spark as an ultra rapid flash for scientific photography, and in certain fields of press and night photography they open up interesting possibilities. A versatile family of electronic flash-tubes has been developed, ranging from a "microsecond" flash tube to a tube which dissipates 10,000 joules in a flash lasting a few milliseconds. Some of the principles of design and properties of these tubes are discussed in this article **).*

The air spark has long been used as a flash illuminant in scientific photography. As long ago as 1892, C.V. Boys made spark exposures of the order of one microsecond [1]). When certain circuit conditions are fulfilled, flashes of duration even less than a microsecond are possible.

Efforts to improve on the air spark lead naturally to the study of the spark discharge in other gases. Apart from the higher luminous efficiencies obtainable in other gases, a sealed discharge tube has the important advantage that the electrodes can be protected from oxidation effects, thus ensuring better reproducibility and more reliable triggering. Furthermore, in a sealed tube the discharge path is constant.

Earlier work on electrical conduction in gases was already considerable and its specific application to the spark discharge as a light source was largely initiated in the U.S.A. by Edgerton and his co-workers [2]). A number of flash tubes appeared on the market, but it was not until the 1939-45 war that the most significant advances were made. During the war an addition fillip was given to work on spark discharges tubes by the requirements of ballistics research. One result of this was the development of the Arditron [3]), a three-electrode tube filled with argon, capable of giving flashes of the order of a microsecond. Such a tube was manufactured by Mullard during the war, under the designation LSD. 2 (*fig. 1*).

The spark discharge as a light-source has, however, far wider potentialities than in scientific research alone. Because of its excellent properties for stopping motion, its economy in use and its multiple life, the flash discharge tube is also attractive for conventional photographic purposes, especially press photography. For the latter, portability of the entire equipment (including the energy source) is essential. For such portable equipment, tubes of high luminous efficiency and low operating voltage were required. By 1947 a number of tubes had been marketed giving about 30 lumen seconds per joule and operating at a few kilovolts. With the new developments in high voltage condensers of small dimensions, these tubes made portable flash equipment a practical proposition. Further work in this direction has led to even lower operating voltages [4]).

Parallel to the development of flash tubes for portable equipments the Mullard laboratory at Salfords has developed a series of higher rated tubes suitable for studio and stage work, and tubes for stroboscopic operation. A selection of these tubes is depicted in *fig. 2*. The design and characteristics of some of these tubes will now be discussed [5]).

Basic design of flash discharge tubes

The flash-tube consists basically of two electrodes sealed into a glass tube containing a gas of predetermined composition at a relatively low pressure. A trigger electrode, which may be internal or external, serves to initiate the discharge. A length of wire running along the glass envelope forms the trigger electrode of the tube shown in *fig. 3*, which also shows the basic circuit. The condenser C is charged to a high D.C. potential through a current-limiting resistance R, whilst a second condenser C_T is charged to a lower voltage through r. On closing the switch S, a current surge in the primary of the transformer T produces a high voltage pulse in the secondary which is applied to the trigger electrode.

*) Mullard Research Laboratories, Salfords, Surrey, England.

**) Attention should be drawn to a number of articles published earlier in this Review dealing with various special types of electronic flash-tubes: S. L. de Bruin, An apparatus for stroboscopic observation, Philips tech. Rev. 8, 25-32, 1946; N. Warmoltz and A. M. C. Helmer, A flash lamp for illuminating vapour tracks in the Wilson cloud chamber, Philips tech. Rev. 10, 178-187, 1948; J. E. Winkelman and N. Warmoltz, Photography of the eye with the aid of electronic flash-tubes, Philips tech. Rev. 15, 342, 1953/54 (No. 12).

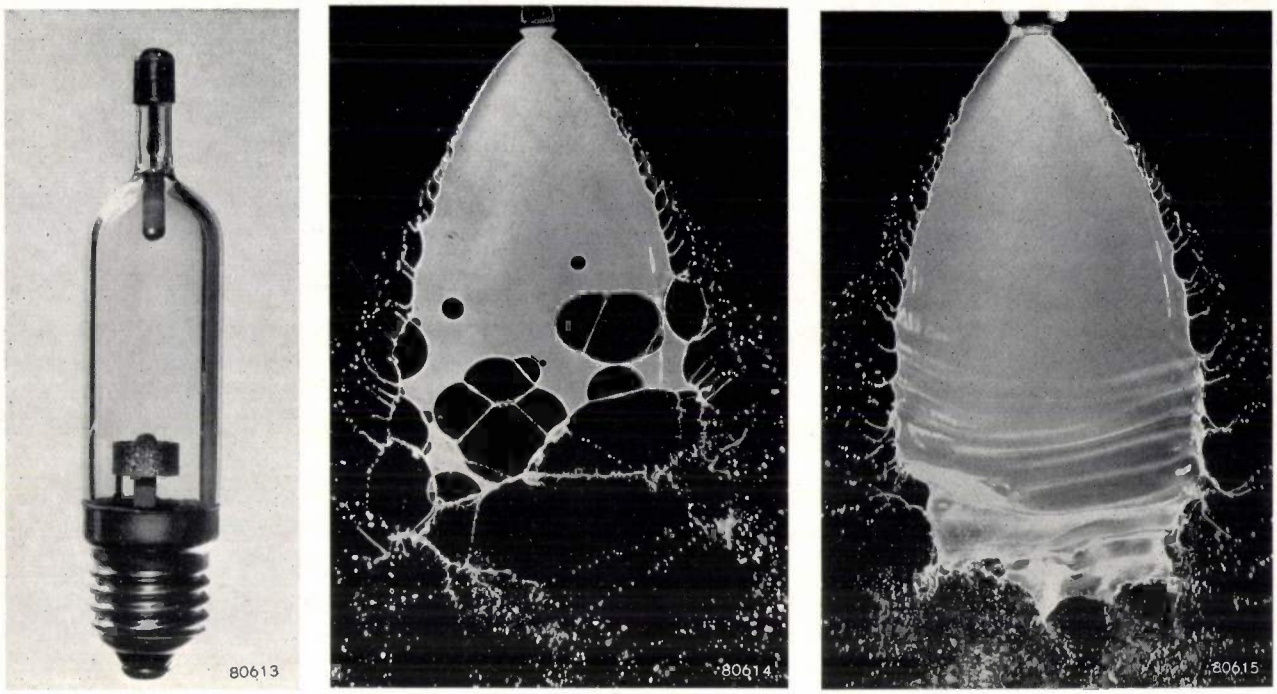
1) C. V. Boys, Proc. Roy. Soc. 47, 415 and 440, 1893. See also Worthington, Proc. Roy, Soc. 59, 250, 1895.

2) H. E. Edgerton, J. K. Germeshausen and H. E. Grier, J. appl. Phys. 8, 2-9, 1937.

3) J. W. Mitchel, Trans. Illum. Eng. Soc (Lond.) 14, 91-104, 1949.

4) An article describing new developments in low voltage flash tubes (500V and under) will appear shortly in this Review.

5) See also G. Knott, High-intensity flash-tubes, Photographic J. 89B, 46-50, 1949.



a *b* *c*

Fig. 1. *a*) The LSD. 2 "Arditron" flash tube. This tube can give a flash of peak intensity 100 megalumens, and of duration of the order of one microsecond. *b*) Photograph taken with LSD. 2 showing the break-up of a plane liquid jet. *c*) Same as *b*) but with a higher spraying pressure, illustrating a somewhat different mechanism of break-up. Two LSD. 2 tubes connected in series were used for each of these photographs. With a series connection, simultaneous firing of the tubes is ensured; connected in parrallel, there may be a small interval between the peaks of the flashes, even through triggered by the same pulse. (Photos (*b*) and (*c*) by courtesy of N. Dombrowski, Jet Research Laboratory, Imperial College, London.)

The sudden rise in the electric field breaks down the gap between the electrodes, and the main condenser discharges, giving an intense burst of light.

The tubes themselves assume a number of different shapes (fig. 2). A compact source is desirable for most purposes and this is most easily achieved, with a long tube, by winding it in the form of a helix. Some tubes are provided with protective glass covers; tubes for stroboscopic operation however, are generally not so protected and the helix is often wound more loosely to provide better heat dissipation. The LSD. 2 (microsecond) tube needs a short discharge path: a short fat tube is therefore used. The LSD. 24 also has a relatively short discharge path: in this case a simple U-tube form is used.

Gas filling

The rare gas xenon is used as a filling for all the Mullard flash tubes with the exception of the LSD.2 (microsecond) flash tube. A rare gas is chosen because it is chemically inert and because a relatively high light output is obtainable for a given discharge energy. Xenon, among the rare gases, gives the best performance in this respect. The relative light outputs of Xe:Kr:A:Ne:He are approximately in the ratio of 100:70:50:18:6. Another important reason for the choice of xenon is that the spectral distribution of its flash discharge approximates to that of mean noon daylight (*fig. 4*). This is an important quality for photographic light sources, since good colour rendering can then be obtained. Figure 4 also shows the curve for argon, which is used as a filling in the LSD.2 tube. Here, the red deficiency of the light is no disadvantage, since this tube is used exclusively for the photography of ultra-rapid events, where colour rendering is of only secondary importance.

The pressure of the gas filling is an important parameter since it determines the breakdown potential of the tube and hence the working voltage. For a working voltage of 2.5 kV the tubes are filled with xenon at a pressure of about 10 cm, which gives a breakdown voltage some 500 to 1000 volts above the working voltage. In the LSD. 2 tube, where an operating voltage of 10 kV is used, the necessary breakdown voltage of $10\frac{1}{2}$-11 kV is obtained by increasing the gas pressure to one atmosphere. A xenon filling at this pressure would be rather expensive: for this reason, an argon filling is nor-

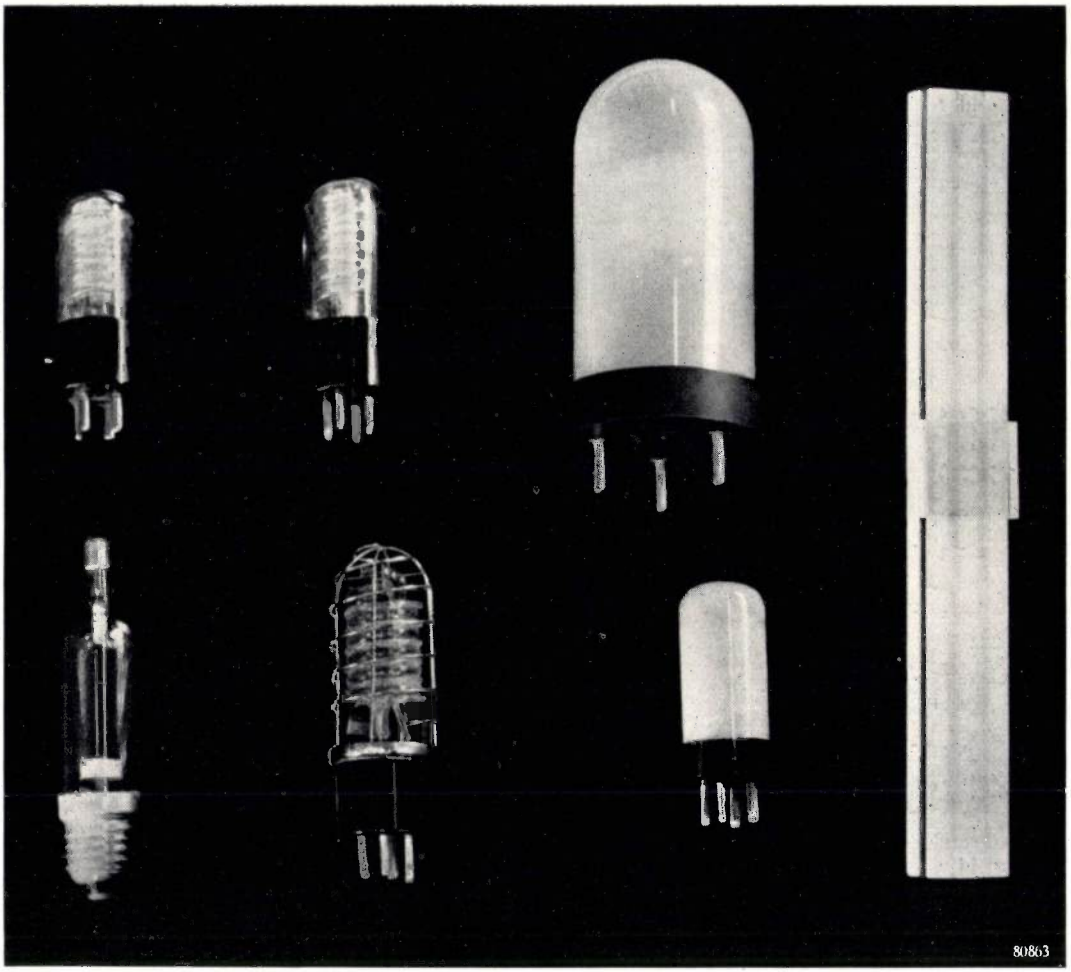


Fig. 2. Photograph of some typical Mullard flash tubes. Upper row, left to right: LSD. 3 (100 joules dissipation), LSD. 7 (200 joules), LSD. 5 (1000 joules). Lower row, left to right: LSD. 2 (35 joules), LSD. 8 strobe tube (dissipation 20 watts at 500 c/s), LSD. 24 (100 joules).

mally used in this tube. Xenon-filled LSD.2 tubes are, however, made for special applications, where the increased light output or better colour rendering justify the extra cost.

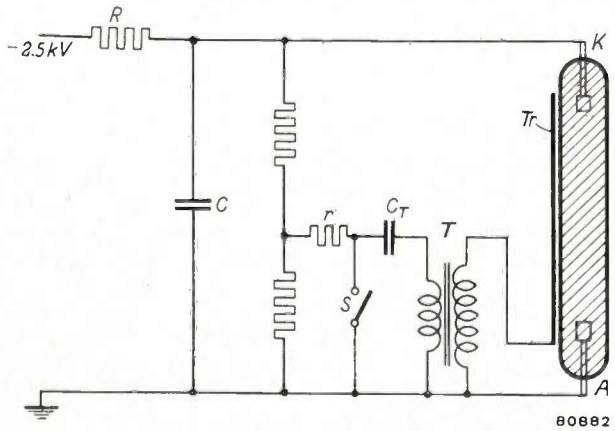


Fig. 3. Basic circuit of a flash tube. R charging resistance, C main condenser, C_T condenser for trigger pulse, r charging resistance for C_T, S initiating contact, T transformer, K cathode. In the usual arrangement (as shown here), the anode A is earthed and the trigger electrode Tr is held at the same potential. The trigger pulse is then a positive surge.

Some characteristics of the discharge

The light output of a flash tube consists of a large number of emission lines superimposed on a continuum which stretches from the infra-red to the ultra-violet (fig. 4 shows only the continuum in the visible region). The continuum accounts for the

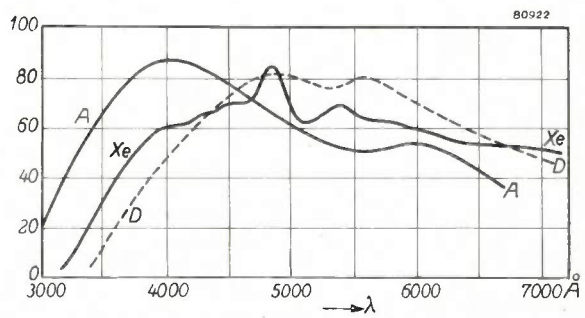


Fig. 4. Spectral distribution curves of xenon and argon discharges, compared with that of mean noon daylight (D). It may be seen that the curve for Xe is a fairly near approximation to that of daylight. The curves represent substantially only the continuous radiation of the discharge, which is in general more intense than the spark or arc lines.

major part of the intensity, although the quantity of light originating in the emission lines may vary considerably, depending on the discharge conditions.

Early measurements of the time variation of the light output indicated that light at all frequencies was not emitted simultaneously. Olsen and Huxford[6]) have made a time analysis of the light output from argon and neon flash tubes and found that there are spark lines, continuum and arc lines, which can be linked to different stages in the life of the plasma. When the discharge is first initiated, high energy electrons are produced because of the high applied field, and ionization proceeds at a very rapid rate. The resulting excited positive ions radiate the characteristic spark lines. The current through the tube increases rapidly and the electrons and positive ions are collected at the electrodes and neutralize the charge on the condenser until the applied field is reduced to zero. Some ions and electrons are lost by diffusion to the walls and by recombination but, due to the high energy electrons produced in the early stages, to thermal ionization and possibly also to metastable atoms, the plasma continues to build up to a maximum which is reached a few microseconds after the peak current. After this the ion density declines. The rise and fall of the continuum coincides very closely with that of the ion concentration and Olson and Huxford attribute it primarily to the retardation of electrons moving in trajectories about ions (*bremsstrahlung* radiation). The rate of decay of this bremsstrahlung radiation depends upon the rate of decrease of ion density and the decay of electron temperature. (The latter may reach a value of 7000 °C or more during the peak of the discharge [7]). The arc lines are observed in the afterglow and are attributed to radiation from the excited atoms formed by recombining electrons and ions.

Light output and luminous efficiency

The light output required for general flash photography at distances of 20-30 feet is of the order of 5000 lumen seconds. The tubes designed for portable flash equipments are desigend around this figure so that with luminous efficiencies of 30-40 lumen second/joule, the tubes must dissipate between 100 and 200 joules. For studio and stage photography, and for research applications, compactness of the energy source is no longer a limitation and tubes dissipating up to 10,000 joules have been made.

Light output is governed primarily by the energy of the discharge, although it is also dependent on the nature and pressure of the gas filling, the geometry of the discharge tube, and to some extent on the working voltage of the tube.

The choice of the working voltage of the tube is a question of compromise between a number of conflicting factors. For greatest efficiency in the conversion of electrical energy to luminous energy, the electric field should be as high as possible. On the other hand, the voltage should be kept as low as possible to minimize leakage losses in the condensers and for reasons of safety — especially in portable equipment. The working voltage adopted for the lamps described here is 2.5 kV, with the exception of the LSD. 2 (10 kV) and the LSD. 24 (1000 V).

In spite of the lower working voltage of the LSD. 24, a luminous efficiency has been achieved better than that of the tubes operating at 2.5 kV. The design of the LSD. 24 reflects the present tendency to use lower operating voltages, especially in tubes for portable flash equipments. This tendency was initiated by the introduction, in recent years, of reliable electrolytic condensers capable of withstanding 500 V. A considerable saving of space is thereby achieved. Furthermore, the weight of the electrolytic condenser is small compared with the corresponding paper condensers, and this alone is sufficient to counteract the disadvantages of their relatively large leakage current. For the LSD. 24, banks of condensers are used in series-parallel to allow operation at 1000 V.

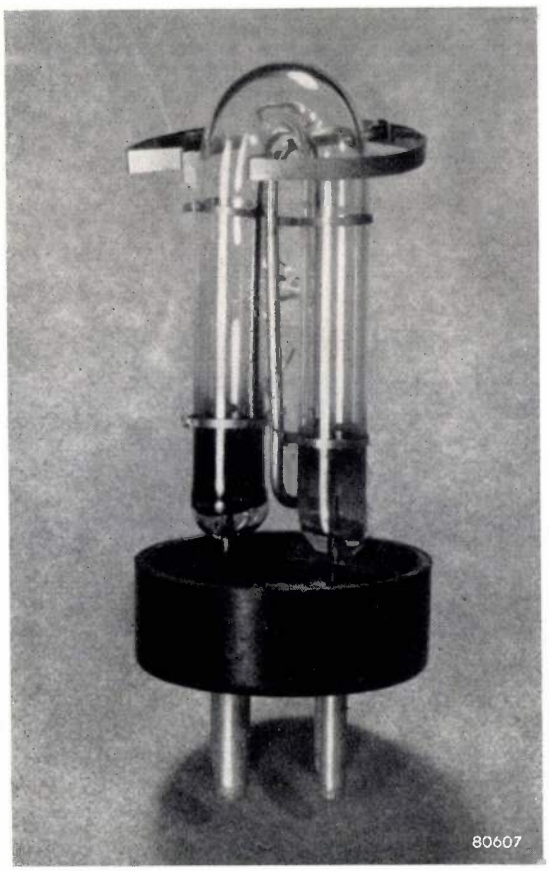


Fig. 5. Photograph showing the construction of the LSD. 24 (1000 V) tube.

[6]) H. N. Olsen and W.S. Huxford, Dynamic characteristics of the plasma in discharges through rare gases, Phys. Rev. **87**, 922-930, 1952.

[7]) The fundamental processes occurring in the spark discharge have been studied in hydrogen and argon by J. D. Craggs and J. M. Meek, The emission of light from spark discharges, Proc. Roy. Soc. **A 186**, 241, 1946.

To obtain the lower operating voltage, the LSD. 24 is considerably shorter than the other flash-tubes, and takes the form of a simple U-tube (*fig. 5*). At the same time the diameter of the discharge tube is made larger in order to carry the heavier currents, which may reach peak values of the order of 1000 amps. Sturdy electrodes are necessary to withstand the dissipation. The light output at different loadings up to 100 joules is given in *fig. 6*. It is seen that there is a considerable improvement compared with with the LSD. 3.

The light outputs of a number of other tubes are also shown in figure 6. It will be seen that the luminous efficiency of a given tube increases with the energyof the discharge. The energy cannot be increased indefinitely, however, owing to the onset of sintering (see below). Under operating conditions, the luminous efficiency of most standard tubes is in the region of 30-40 lumens per watt, although higher values are obtainable with some forms of tube.

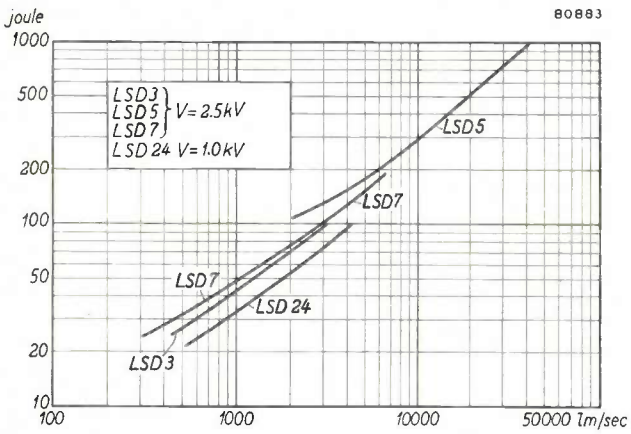


Fig. 6. Luminous outputs of a number of flash tubes as functions of the initial energy in the condensers. In general, the luminous efficiency of a given tube increases as the energy of the discharge is increased.

Current loading

There is a limit to the power rating of a discharge tube of given dimensions, set by the maximum current which can be passed without overheating of the glass bulb. The intense heat generated during the discharge appears to melt the inner surface of the glass and the strains produced on subsequent cooling give rise to minute hair-like cracks (*fig. 7*). This "crazing" or "sintering" of the glass is therefore a thermal shock effect. Although slight sintering does not weaken the tube mechanically, it may cause an increase in the minimum operating voltage, due to contamination by gases evolved from the walls. For this reason the tube must be so constructed that sintering does not occur when flashed at its maximum rated loading.

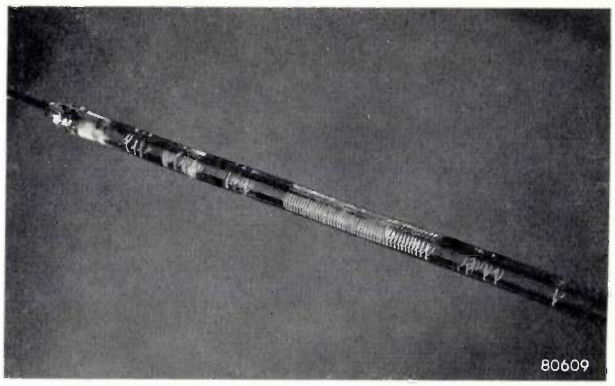


Fig. 7. Photograph of a "sintered "tube. The characteristic crazing of the inner surface of the glass has in this case shown itself in a very regular helical pattern.

The sintering "cracks" often develop some time after the discharge, and only in the case of severe overloading do they appear immediately. Sintering is prone to occur in the neighbourhood of the electrodes and also near sharp bends or constrictions in the discharge path. The cracks do not generally penetrate far into the glass but there is a tendency for them to spread over the surface. In severe overloading, however, the surface cracks may extend sufficiently to cause fracture of the tube.

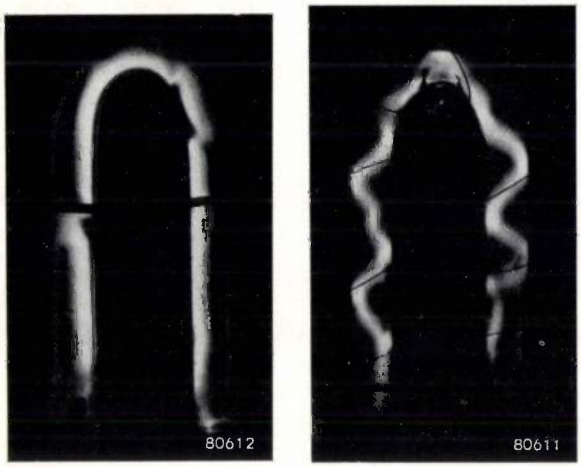
By a special treatment of the glass surface, sintering can be minimized, so that the dimensions of a tube to dissipate a given load may be somewhat reduced.

Initiation of the discharge

The auxiliary electrode by means of which the discharge is triggered, may be internal or external; see for example, fig. 1*a* (internal trigger) and fig. 5 (external trigger).

The trigger electrode is held at anode potential which is normally earth, and the cathode is at a high negative potential (fig. 3).

In the absence of the trigger electrode, there would be an almost uniform potential field between the anode and cathode. With the trigger electrode in position, the field is almost all concentrated between the cathode and the tip of the trigger. Under these circumstances the tube is nevertheless stable. If now a positive surge is applied to the trigger electrode, the field between the cathode and the trigger electrode increases so much that any electrons in the neighbourhood are accelerated up to the ionization potential of the gas. Hence an ionized region is created and at the same time secondary electrons will be produced from the glass and photons will be generated. The ionization will spread in the direction of the anode at a very rapid pace until finally the gap between the anode and the cathode is bridged, and a track is made for the main discharge. The photographs of *fig. 8* show that initially the main discharge follows very closely the path taken by the trigger electrode.



a *b*

Fig. 8. Photograph of the discharge of an LSD. 24 tube, showing that the discharge follows the path closest to the trigger electrode.
a) LSD.24 with ribbon-shaped strip of metal as trigger electrode. The dark bands across each arm of the tube are securing bands.
b) LSD.24 with helically wound wire as trigger electrode.
These photographs were taken at a discharge energy well below the rated value; in this way the discharge does not fill the whole tube and the preferred path is thus rendered visible.

When fully developed, the discharge spreads across the whole cross-section of the tube. The tube is here an LSD. 24: in the one case the trigger electrode is a spirally wound wire, whilst in the other it is a ribbon shaped strip of metal laid along the tube, with two securing loops round each arm of the tube.

To determine the most effective position for the trigger electrode, the tube is operated under continuous glow-discharge conditions, the current being limited by a resistor. It is found that the discharge path is clearly defined and is usually the shortest distance between the electrodes. If the trigger electrode is now laid as closely as possible to this preferred path, the trigger voltage required to initiate the discharge is a minimum.

The value of the triggering voltage required is dependent on its rate of rise and on its sign. The trigger surge is usually in the form of a highly damped oscillation, and it is found that the amplitude required is lowest when the frequency of oscillation is high. The first peak of the oscillation, being the greatest, is the most significant from the point of view of triggering, although the subsequent peaks are of some importance in all except the ultra-rapid flash tubes. For the most effective triggering, therefore, the rise time of the first peak must be as short as possible and, with the usual arrangement of an earthed anode (see fig. 3), it should be a positive peak.

The trigger voltages applied to the tubes operating at 2500 V are of the order of 3000 V (so that at the moment of triggering, a voltage of about 5500 V exists between cathode and trigger electrode). To prevent misfires, the applied trigger pulse must obviously be greater than the lowest voltage required to produce *certain* flashing. In fact, there is a large spread in the triggering voltages required for different tubes of the same type and even between successive firings of the same tube. The factors governing this spread of triggering voltage are similar to those governing the spread of the self-breakdown potential, discussed below.

It is not possible to increase the trigger voltage indefinitely owing to the risk of tracking across the base or on the surface of the glass, but the necessary voltage can be reduced somewhat by using a cathode activated with an alkali or alkaline earth metal. (This also has the result of slightly lowering the self-breakdown voltage of the tube, but this can be restored by a small increase in gas pressure.) Barium is commonly used for activating the cathode, and is evaporated on to it in vacuo. The barium also fulfils another important function by acting as a "getter" for impurities present in the gas or liberated during the discharge. With activated cathodes the triggering voltage required may often be less than 1000 V, but to allow for statistical variations, trigger voltages of 3000 V are recommended.

Self-breakdown

The potential at which a flash-tube will discharge without any surge on the trigger electrode is known as the self-breakdown voltage. As in the case of the triggering voltage, this is by no means a fixed value. *Figs. 9a* and *b* are records of the number of times a given flash-tube has broken down at the voltages shown on the axes.

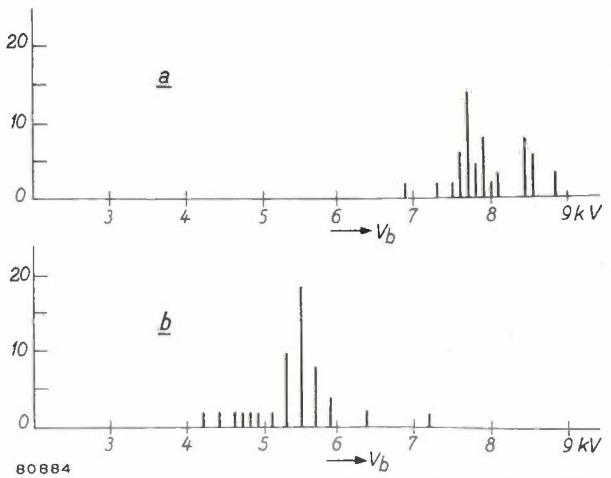


Fig. 9. Statistical variation of the self-breakdown voltage V_b of an experimental LSD. 3 tube.
a) Trigger electrode held at cathode potential.
b) Trigger electrode held at anode potential.

The self-breakdown voltage is considerable dependent on the previous history of the tube. Under this heading, the effects of wall charges on the inner surface of the glass are probably the most important cause of the wide variations in breakdown. Wall charges will depend upon previous flashes, temperature, the nature of the glass surfaces, and any external fields. Another source of variation in the self-breakdown voltage is the presence of sources of ionization in the neighbourhood; for example, ionization due to ambient light, cosmic rays etc.

As a result of these variations, the tubes must be constructed with a mean self-breakdown voltage very much greater than the operating potential across the tube.

It should be noted that the self-breakdown voltage of a particular tube is also influenced by the rate of rise of the potential applied across the tube. The greater the rate of rise, the lower will be the self-breakdown potential. In practice, premature self-breakdown from this cause is prevented by a current limiting resistor before the condenser (fig. 3).

Operation

The behaviour of flash-tubes in practice is to a large extent governed by the external circuit. Of the energy stored in the charged condenser, only a portion is used effectively and converted into radiation. The rest is distributed between the residual charge in the condenser after the flash, circuit losses, and ohmic losses in the tube itself.

Duration of the flash

The concept of a "resistance" is applicable to flash-tubes [8]). This may be adduced from the voltage and current curves of *fig. 10*. These measurements were made on an experimental tube of the LSD. 3 type. After peak current is reached, the current and voltage curves both decay exponentially at such a rate that their ratio is roughly constant, i.e. the effective impedance of the tube is purely resistive [9]).

The value of this ratio (ϱ) is of the order of a few ohms in most flash tubes, but is dependent on the operating conditions, e.g. the voltage across the condenser and its capacity. It has a particularly high value for small capacitances, i.e. for small energy dissipation: this is due to the discharge not occupying the whole cross-section of the tube. The product of the tube resistance (ϱ) and the capacitance (C) of the condenser, gives an approximate time constant $C\varrho$ for the discharge. To the value so obtained, a few microseconds must be added to allow for the initiation of the discharge.

From fig. 10 it may be seen that the development and decay of the light output follows closely the current curve, except that the peak light output persists somewhat longer than the current peak. Thus the time-constant $C\varrho$ may be taken as a rough estimate of the minimum duration of the flash.

This method of calculating the duration is not applicable to the LSD. 2 tube. This has a short direct discharge path whose impedance is very low and has an inductive component which gives rise to an oscillatory discharge. The duration of the LSD. 2 flash under normal operating conditions is of the order of 5 μ sec, which is considerably greater than that obtained if the above calculation were applied. Normal operation implies a maximum dissipation *in the tube* of 35 joules : this is usually obtained from a condenser of 2 μF charged to 10 kV, assuming circuit losses of about 65%. Photographically an effective duration of 1-2 μsec can be obtained by a careful control of the intensity of light falling on the plate and by judicious development.

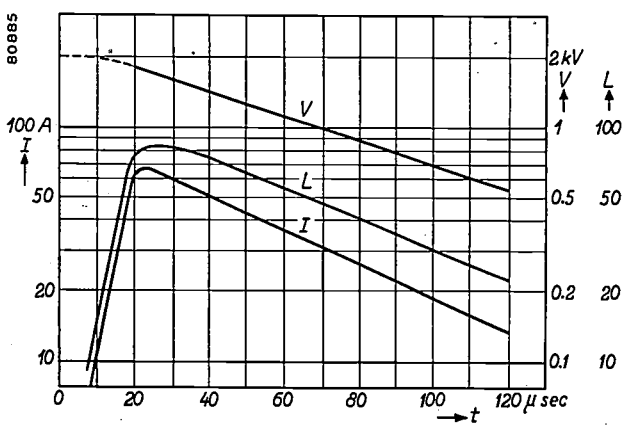


Fig. 10. Curves of voltage (V), current (I) and luminous output (L) plotted logarithmically as functions of time (t), for an LSD.3 experimental tube. During the decay period, the current is very nearly proportional to the voltage. The light output follows the current curve very closely, save that the peak of the former persists somewhat longer than that of the current.

Energy losses

One of the chief sources of energy loss in flash tubes is the heat loss in the tube itself. Calorimetric measurements show that these may be as high as 60%. An incidental but important practical con-

[8]) This was first shown by M. Laporte, Etude de la décharge d'un condensateur à travers un tube à gaz, J. de Physique 8, 332, 1937. See also P. M. Murphy and H. E. Edgerton. Electrical characteristics of strocoscopic flash lamps, J. appl. Phys. **12**, 849, 1941.

[9]) Prior to the development of peak current, the ratio of voltage to current is very high, and is of course infinite before the discharge is initiated. After a finite decay time, the voltage across the tube becomes insufficient to maintain the discharge, so that the exponential decay of current gives place to a more rapid drop to zero. The resistance of the tube then becomes infinite once more.

sequence of the very high rate of heat liberation at the very high currents which flow through the electrodes, is, that good welds between the latter and the lead-in wires are essential.

Losses in the external circuit are due to dielectric losses in the condenser and to ohmic losses in the wires, which are considerable owing to the large currents.

To minimize these losses and also to keep the flash duration as short as possible, low leakage, non-inductive condensers are desirable and the connecting wires should be kept as short as possible. Depending on the arrangement, circuit losses may account for between 10 and 50% of the total energy stored in the condenser. In the case of the LSD. 2 tube, circuit losses may be as high as 70%; by mounting the tube directly on the condenser, this may be reduced to 60%. Connection in this manner also makes for the shortest flash duration.

One further source of energy loss is the residual charge left in the condenser. This arises from the fact that when the voltage drops to a certain value, the discharge can no longer be sustained. Unless a flash is followed immediately by recharging and a further flash, the charge corresponding to this voltage will leak away and thus represent a loss of energy. The extinction voltage of a flash tube depends on the geometry of the tube, the gas filling and the operating voltage. For a given tube, with the condenser initially charged to a few kilovolts, the residual charge corresponds to a voltage V' of a few hundred volts, being larger the larger the initial charging voltage. Neglecting V' for the moment, for a specified total energy dissipation, the capacitance of the condenser to be used for the various tubes depends inversely on the square of the operating voltage. The residual voltage (V'), however, for a given tube depends only very slightly on the operating voltage so that the residual energy $^1/_2\, CV'^2$ becomes a smaller fraction of the whole as the operating voltage increases. Thus, for example, at 100 joule operation of a tube designed for 2.5 kV, the residual energy in the condenser when working at this voltage ($C \approx 30$ μF) is about 1.4 joules, whilst at 1 kV ($C = 200$ μF) it is about 9 joules For the tubes *designed* for operation at lower voltages, however, (e.g. LSD. 24) the extinction voltage is proportionately lower, so that the energy loss due to the residual charge in the condenser is still only a few per cent.

Life of flash-tubes

There is no theoretical limit to the number of times a flash-tube may be fired. Samples of the various types of tube are normally given life tests of 10 000 flashes but one tube has been tested to 100 000 flashes without any electrical deterioration; some of the barium from the cathode had sputtered into the glass envelope but without seriously reducing the light emitted.

Some practical problems of construction

A series of flash-tubes for general and specialist requirements have been developed from the principles and design data described above. Photographs of the standard tubes are given in fig.2. Performance curves (light output v. energy dissipated) of a number of tubes are plotted in fig. 6.

Some of the problems encountered during the development of flash-tubes may now be mentioned.

Heat dissipation

The problem of heat dissipation is accentuated in the higher rated flash-tubes and in the stroboscopic tubes. Attention has to be paid to the sinter-resisting properties of the glass, and the electrical resistivity, both of which decrease with temperature.

As a result of the conductivity of the glass at higher temperatures, power is absorbed from the trigger circuit, and the possibility arises that the trigger voltage might decrease to a value insufficient to produce a discharge. If the trigger voltage is maintained at the original voltage required when the tube is cold, then there is some danger that the trigger discharge will penetrate the heated glass and cause puncture. In the LSD. 5 tube, therefore, which runs hot (300 °C) as a result of a 50 watt filament modelling lamp mounted inside the helix, the trigger wire is *separately* covered in a glass helix which remains sufficiently cool to retain its insulating properties (see *fig. 11*). In this way a trigger voltage as high as 6 kV can be used and efficient operation at loadings of 200-1000 joules is possible. An example of a multiple flash photograph taken with this type of tube is shown in *fig. 12*.

A similar type of trigger is used on the LSD. 8 stroboscope tube which runs hot when strobing at its rated loading of 20 watts (0.04 joules per flash at 500 c/s). Without the protected trigger, firing becomes uncertain when the temperature of the helix reaches about 300 °C.

Work was done on a flash-tube capable of operating at a much higher temperature, in which a helix of silica was used in place of glass. In addition, a nickel cage was used instead of a protective glass cover; this not only allows a faster dissipation of the heat, but also has the advantages that the ultra-violet radiation is not absorbed. Such a tube dissipated up to 1000

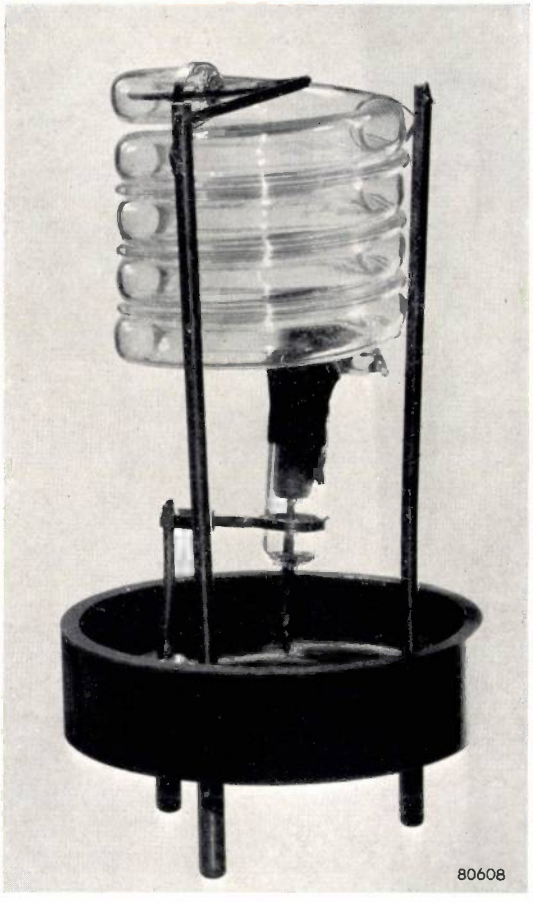


Fig. 11. The LSD. 5 tube, of 1000 joules dissipation. The subsidiary glass helix containing the trigger wire can be seen lying between the turns of the main helix.

joules although it was no larger than the 200 joule LSD. 7 tube (which uses a glass helix).

Another experimental tube was constructed which was capable of dissipating 10 000 joules. This was made from glass tubing, the operating temperature in this case being kept sufficiently low by spacing the turns of the helix.

Sputtering

In stroboscopic tubes, some of the material is sputtered from the cathode during each flash, and would very soon build up a deposit on the walls of the tube and reduce the light output. Precautions have to be taken in the design of the tube to prevent this deposition in the light path. The cylindrical cathode is therefore made large in diameter so that only a small annular space is left between the cathode and the envelope, and also the cavity behind the cathode is made as small as possible. The reason for this construction follows from a consideration of the effect of a discharge on the movement of the gas in the tube. The gas in the discharge path is heated and the pressure increases to many times its static value. A shock wave is generated. and the gas stream forces the sputtered particles into the annular region. Since the cathode almost fills the cross section of the tube, the particles now have only a short distance to travel before reaching the envelope and have a chance to adhere to the glass before the reflected wave catches them and carries them into the main discharge space. Further-

Fig. 12. Multiple flash photograph taken with LSD. 5 tube. (Photo by Ronald Startup, courtesy of Picture Post.)

more, there is a tendency for sputtered particles to be deposited when the velocity of the transporting gas is low: if the cavity behind the cathode is small, the gas stream comes to rest momentarily within the length of the cathode and any particles caught by the initial wave front have a further chance to be deposited before the rebound. If the cavity behind the cathode is large, the shock wave tends to be broken into vortex motion and the gas-borne particles are swept back into the discharge path were deposition is likely to take place.

This feature of design has been particularly useful in high power strobe tubes, as sputtered material from the cathode, which is considerable at heavy loadings, is then confined to a narrow ring on the glass near the tip of the cathode and none penetrates into the discharge space.

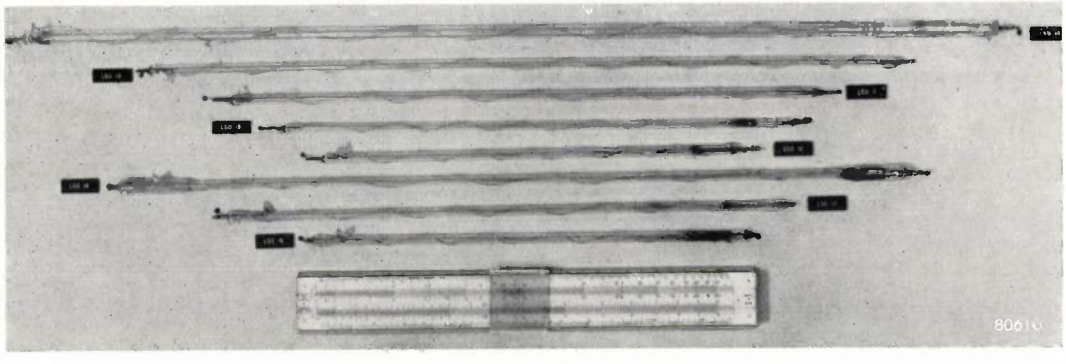


Fig. 13. Linear flash tubes. These tubes are used chiefly for photography in Wilson cloud chambers and for medical research. Another application is in photo-printing and copying, where their speed and absence of heating are used to advantage. The efficiency of the linear tubes is high, but their shape and size make them unsuitable for general use.

Special tubes

In addition to the tubes already referred to, a group of linear tubes have been designed (*fig. 13*). Besides these tubes, a number of flash-tubes have been developed for special applications. Some of these tubes are non-standard only in respect of their special shape. Such tubes are designed for photography in awkward or confined spaces, such as in oil boreholes.

A special model of the LSD.2 was developed, filled with xenon instead of argon. This had a considerably greater light output than the standard LSD.2 but retained its special feature of a very short duration flash. A difficulty in the use of xenon, however, arises from its greater atomic weight. As mentioned above, the discharge gives rise to a travelling shock wave, which is reflected at the ends of the tube, and the impact of the heavier atoms of xenon on the glass envelope is sometimes sufficient to cause shattering. As a result, certain regions of the glass envelope had to be considerably thickened.

A number of multiple-electrode tubes have been made. One of these, designed for a special research application, took the form of two coaxial cylindrical tubes joined at one end and the inner one open at the other (*fig. 14a*). Each tube had a cathode at the closed end and shared a common anode at the other. A continuous discharge could be maintained in the centre tube, whilst the outer tube was flashed at intervals. Another multiple tube for research purposes was designed to give a succession of four flashes on the application of one triggering voltage. The tube was circular in form (fig. 14*b*) and had three intermediate electrodes in addition to the cathode and the final anode. The condensers were connected either between each successive pair of electrodes, or between each electrode and a common cathode. A discharge initiated in one section caused

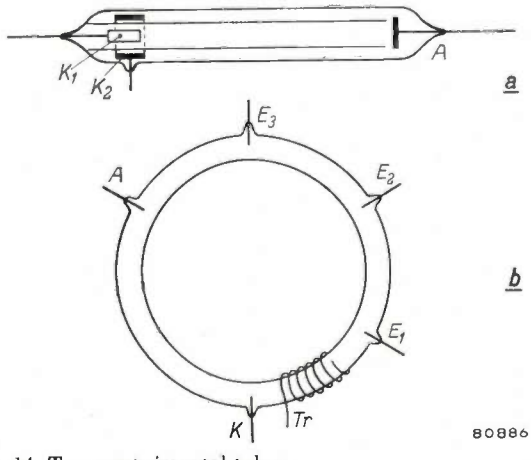


Fig. 14. Two experimental tubes.
a) Dual co-axial tube in which the outer tube may be independently flashed whilst a continuous discharge is maintained in the centre.
b) Multi-electrode tube which gives four successive flashes on the application of one triggering pulse.

expansion of the gas and a consequent rapid transport of electrons to the next section which was thereby triggered. The same process caused triggering of the third and fourth sections in succession. The interval between the discharge of the various sections remains constant for a given tube and given operating conditions.

Other multiple flash tubes have been made for stroboscopic purposes. One of the simplest of these consist of a U-tube with cathodes at the end of each limb. and a common anode at the bend. The two limbs may then be flashed alternately, so that the loading of each limb is halved at a given flash frequency.

By virtue of their special electrical characteristics flash tubes are sometimes used as circuit elements. An example is the use of a flash tube to simulate a magnetron as the load for a pulse forming network. Other flash tubes are used to provide current pulses for the excitation of magnetostriction oscillators for under-water echo sounding. In these applications the light output is, of course, incidental and unimportant. Argon is normally used as the gas filling.

Electronic flash-tubes have now established a firm footing in the photographic world. A better insight into their operation, backed by practical experience in their design, is likely to lead to further developments, particularly with regard to reduction in size, reduction in operating voltages and improvement of luminous efficiencies.

Summary: The principles of operation and the basic design of electronic flash-tubes are outlined. The gas filling and the nature of the discharge are discussed and the various parameters such as light output, operating voltage and current loading, which determine the design of a tube, are set forth. The mechanism of triggering is dealt with, followed by discussions on the influence of the circuit on the behavior of flash-tubes, the "resistance" of flash-tubes, flash duration and energy losses. Finally, some practical problems encountered during design and some unusual experimental tubes are briefly described.

AN EXPERIMENTAL PHOTOCONDUCTIVE CAMERA TUBE FOR TELEVISION

by L. HEIJNE, P. SCHAGEN and H. BRUINING.

621.383.4:621.385.832:621.397.611

The photographs shown here relate to a new type of television-camera tube [1]). Development work on this type of tube is in progress at the Philips Research Laboratories in Eindhoven.

Unlike conventional camera tubes, in which light impinging upon a light-sensitive layer causes it to emit electrons (photo-emission), the new tube contains a sensitive layer that becomes conductive under the influence of light. If the surface of the layer serving as the image electrode initially has a uniformly distributed charge, incident light will

[1]) P. K. Weimer, S. V. Forgue and R. R. Goodrich, The Vidicon photoconductive camera tube, R.C.A. Rev. **12**, 306-313, 1951.

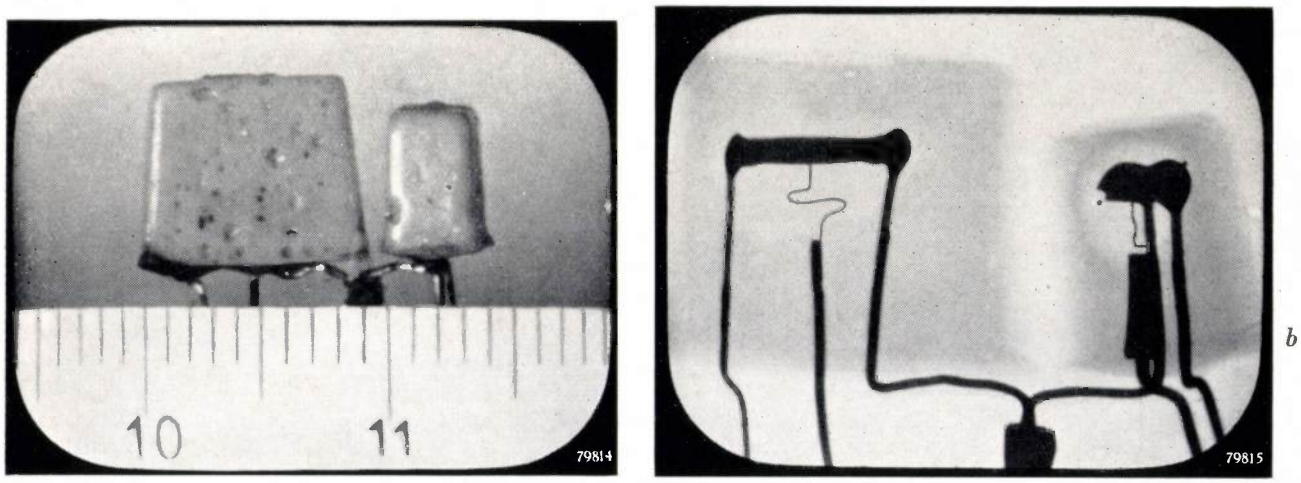


Fig. 1. Photographs of two transistors, made with the aid of a camera tube using a photoconductive sensitive element. *a*) Picture obtained with ordinary light. *b*) Picture obtained with X-rays; both are photographs of the actual image produced on the screen of the picture tube. (These photographs have already appeared in Nature **173**, 220, 30 Jan. 1954.)

expansion of the gas and a consequent rapid transport of electrons to the next section which was thereby triggered. The same process caused triggering of the third and fourth sections in succession. The interval between the discharge of the various sections remains constant for a given tube and given operating conditions.

Other multiple flash tubes have been made for stroboscopic purposes. One of the simplest of these consist of a U-tube with cathodes at the end of each limb. and a common anode at the bend. The two limbs may then be flashed alternately, so that the loading of each limb is halved at a given flash frequency.

By virtue of their special electrical characteristics flash tubes are sometimes used as circuit elements. An example is the use of a flash tube to simulate a magnetron as the load for a pulse forming network. Other flash tubes are used to provide current pulses for the excitation of magnetostriction oscillators for under-water echo sounding. In these applications the light output is, of course, incidental and unimportant. Argon is normally used as the gas filling.

Electronic flash-tubes have now established a firm footing in the photographic world. A better insight into their operation, backed by practical experience in their design, is likely to lead to further developments, particularly with regard to reduction in size, reduction in operating voltages and improvement of luminous efficiencies.

Summary: The principles of operation and the basic design of electronic flash-tubes are outlined. The gas filling and the nature of the discharge are discussed and the various parameters such as light output, operating voltage and current loading, which determine the design of a tube, are set forth. The mechanism of triggering is dealt with, followed by discussions on the influence of the circuit on the behavior of flash-tubes, the "resistance" of flash-tubes, flash duration and energy losses. Finally, some practical problems encountered during design and some unusual experimental tubes are briefly described.

AN EXPERIMENTAL PHOTOCONDUCTIVE CAMERA TUBE FOR TELEVISION

by L. HEIJNE, P. SCHAGEN and H. BRUINING.

621.383.4:621.385.832:621.397.611

The photographs shown here relate to a new type of television-camera tube [1]). Development work on this type of tube is in progress at the Philips Research Laboratories in Eindhoven.

Unlike conventional camera tubes, in which light impinging upon a light-sensitive layer causes it to emit electrons (photo-emission), the new tube contains a sensitive layer that becomes conductive under the influence of light. If the surface of the layer serving as the image electrode initially has a uniformly distributed charge, incident light will

[1]) P. K. Weimer, S. V. Forgue and R. R. Goodrich, The Vidicon photoconductive camera tube, R.C.A. Rev. **12**, 306-313, 1951.

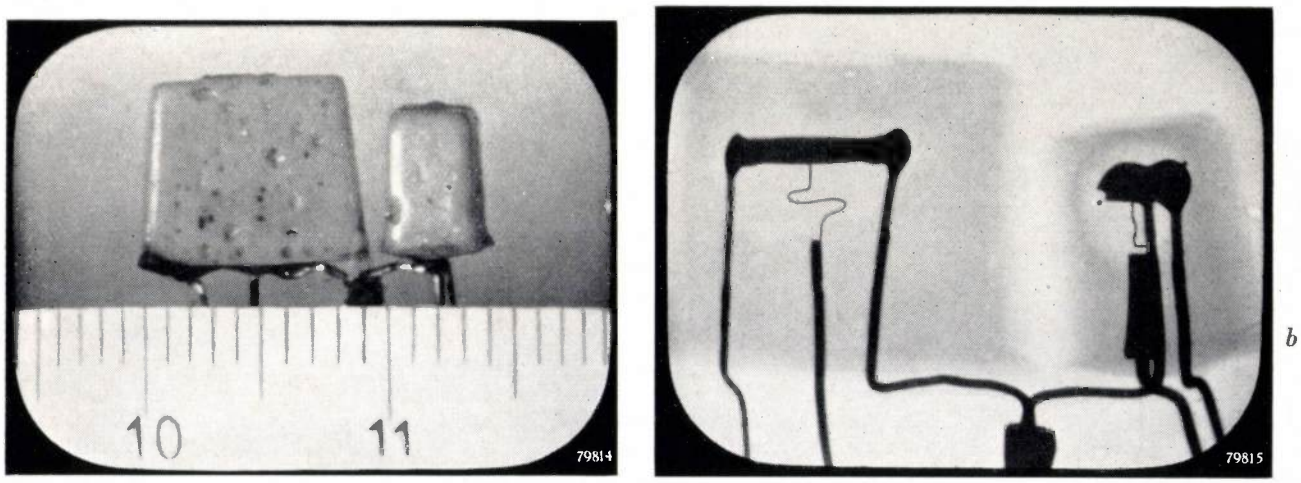


Fig. 1. Photographs of two transistors, made with the aid of a camera tube using a photoconductive sensitive element. *a*) Picture obtained with ordinary light. *b*) Picture obtained with X-rays; both are photographs of the actual image produced on the screen of the picture tube. (These photographs have already appeared in Nature **173**, 220, 30 Jan. 1954.)

cause this charge to leak away through the layer. The speed at which this occurs depends on the intensity of illumination in this region, i.e. in the "white" parts of the picture more charge will flow away through the layer than at the dark parts.

materials used in other tubes of this type, this substance contains one of the heavy elements, viz. lead, and has, therefore, the property of absorbing X-rays to a considerable extent. Even the thin layer (5 μ) in the tube absorbs a sufficient number of

Fig. 2. Set-up for making the photograph shown in fig. 1*b*. *CAM* television camera with the new camera tube. Lens (*L*) has been removed and the object to be irradiated (*O*, in this case two transistors) has been placed directly in front of the tube. *R* is the X-ray tube ("Practix"). On the screen of the picture tube, to the left, the greatly enlarged X-ray shadow image of the two transistors is visible.

In this way a charge pattern is produced on the image electrode. This pattern is scanned by a beam of low velocity electrons. The scanning replenishes the charge and at the same time gives rise to the TV-signal.

The light-sensitive material used in he present tube is a specially prepared lead oxide. Unlike the

X-ray quanta to produce a useful photo-conduction current. It is thus possible to convert an X-ray image directly into a television signal which is subsequently made visible on a picture tube.

Fig. 1a shows example of a picture made with ordinary light, and fig. 1*b* of one of an X-ray image. Both are photographs of the actual picture pro-

duced on the picture-tube. The object here consists of two transistors and a centimetre scale. For fig. 1*a* an optical image was produced on the picture electrode by a lens in the normal way. Fig. 1*b* shows the image obtained when the object is placed directly in front of the window of the camera tube and irradiated by X-rays. *Fig. 2* shows the set-up with which the latter picture was made. The X-ray source is a "Practix" apparatus, placed at a distance of 70 cm from the object. The lens has been taken off the camera and is replaced by the object (the two transistors), which are just visible in the photograph. More details can be seen in *fig. 3*, which also shows some interior components of the camera. From fig. 2 it can be seen that a considerably enlarged X-ray image is formed; if a picture tube with a 35 cm diagonal is used, the magnification is 17×, since the active area of the camera tube has a diameter of 2 cm.

Fig. 4 shows the camera tube itself. Compared with the image iconoscope type 5854, the new tube has a considerably simpler construction and smaller dimensions.

The sensitivity of the new tube to incandescent lighting of colour temperature 2600 °K amounts to about 100 to 200 μA per lumen. This is about 3 times greater than the sensitivity of the transparent photo-emission type cathode, consisting of antimony and caesium, used in other types of camera tubes. The spectral distribution of the sensitivity is about the same for both types of photosensitive layer.

So as not to impair the X-ray sensitivity, a special glass has been used for the envelope, which contains mainly light elements. The result is that the 1.2 mm-thick window absorbs only about 10% of the X-rays. The X-rays used for this demonstration are generated in a tube operated at 70 kV direct voltage, and filtered by 5 mm of aluminium. The 5 μ layer of lead oxide absorbs 5 % of these rays. It should be noted that the X-ray sensitivity is insufficient for medical applications; this would require a prohibitive radiation dose from the point of view of the patient. The X-ray applications are thus confined to the industrial field.

Fig. 3. The camera, showing the camera tube (*P*) and the focussing and deflection coils (*FD*). Above these is the video preamplifier. On the front of the camera, fixed over the lens aperture, are the two transistors (*O*), which served as the object in the set-up shown in fig. 2.

A special problem inherent in photoconductive layers — as opposed to photo-emitting layers — is their inertial effect. The response-time, however, is sufficiently short for most industrial purposes. It is hoped to reduce this time sufficiently to permit the use of this tube for broadcast television.

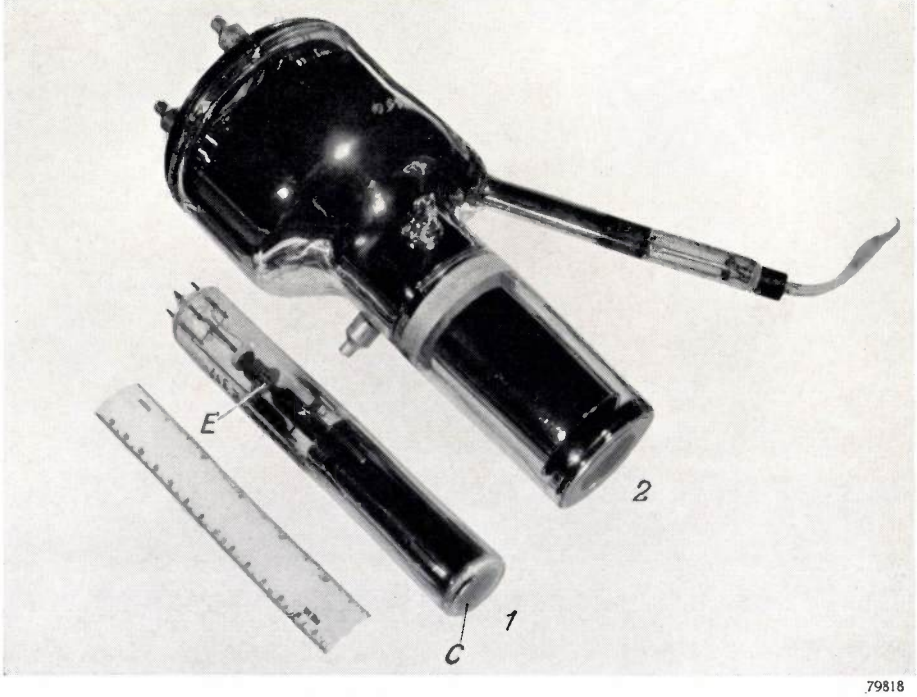


Fig. 4. Camera tube (*1*) with photo-conductive electrode (*C*) and electron-gun (*E*), compared with an image iconoscope (*2*), type 5854.

SECONDARY EMISSION FROM THE SCREEN OF A PICTURE-TUBE

by J. de GIER, A. C. KLEISMA and J. PEPER.

537.533.8:535.371.07:621.397.62

The emission of secondary electrons by the screen of a television picture-tube warrants a closer study to examine its relationship (and that of the associated screen potential) to the undesirable phenomenon of ion burn on the screen due to positive ions, and to the effect of the screen potential on the brightness of the picture. Measurements of the screen potential carried out on finished tubes demonstrate the behaviour of the screen during the life of a tube and facilitate the search for means of improving the screen properties.

The fluorescent screen is one of the most important elements of a picture-tube, since the televised image is formed on it; any faults that may occur in the screen produce immediately noticeable defects in the image. These defects are closely related to the secondary electron emission of the screen when struck by primary electrons.

The screen, itself a fairly good insulator, is completely isolated on the inner surface of the glass window in the picture tube; the net effect of the primary electrons and the secondary emission cause it to assume a certain potential, which must be maintained within given limits to ensure proper tube operation.

An investigation into the secondary emission and the potential of the screen is described in this article; the relationship between these two quantities will first be considered.

Secondary emission

When a stream of primary electrons (I_p) impinges on a substance, in this case the fluorescent screen of a picture-tube, a stream of secondary electrons (I_{s0}) is released. The ratio of the two currents ($I_{s0}/I_p = \delta$) is described as the secondary emission factor, the value of which depends upon the nature of the particular substance and the energy of the primary electrons. The relationship between δ and the primary electron energy can be shown graphically as in *fig. 1*, from which it will be seen that secondary emission is zero at relatively low energy levels, and increases with the energy. The secondary emission factor δ passes unity at a primary energy level corresponding to a certain acceleration potential V_1 and reaches a maximum (δ_{max}) at a higher level corresponding to an accelerating potential V_{max}. Any increase in the primary electron energy beyond this level is accompanied by a gradual decrease in δ, which again passes unity at an energy level corresponding to the accelerating potential V_2. In the case of a fluorescent screen, the value of V_{max} is about 1000 V, the exact value depending upon the method of application and the nature of the screen.

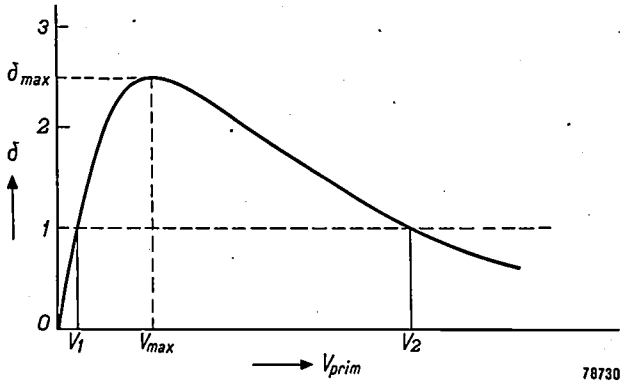


Fig. 1. Variation of the secondary emission factor δ of a solid as a function of the accelerating potential V_{prim} of the primary electron beam. $\delta = 1$ when $V_{prim} = V_1$ and $V_{prim} = V_2$, and it reaches a maximum ($\delta = \delta_{max}$) between the two, at $V_{prim} = V_{max}$.

The variation of δ as a function of the primary electron energy may be explained in the following manner. In passing through a crystal primary electrons impart their energy to electrons of the crystal. Some of the crystal electrons which thus acquire a small extra energy (roughly half of them) pass to the surface of the crystal and are able to leave the latter by virtue of their increased energy; these, then, are the secondary electrons. The secondary emission increases with the energy of the primary electrons because of the continual increase in the number of crystal electrons to which extra energy is imparted. As the energy of the primary electrons increases they are enabled to penetrate more deeply into the crystal; hence some of the crystal electrons released lie so deep within the crystal that they either lose their extra energy on their way to the surface, or are trapped at capture centres. Accordingly, as the primary electron energy continues to increase, the secondary emission will finally reach a saturation value.

The gradual decrease in secondary emission that accompanies any increase in the primary energy beyond this point is due to the fact that the number of secondary electrons released per unit path length diminishes with further increase in the primary electron energy; thus the number of secondary electrons derived from the surface layer, which is the principal source of secondary electrons, gradually decreases. The average energy of secondary electrons is a few electron volts,

the maximum value being some tens of electron volts. Spurious secondary electrons, that is primary electrons reflected either from the crystal surface or from within also occur; reflection takes place with or without loss of energy, so that these electrons have energies comparable with the primary electrons (in this case a few keV). These spurious secondary electrons are not considered in this article.

Let us now consider the case of a tube of which the accelerating voltage $V_a - V_k$ (anode potential V_a, and cathode potential V_k) lies between V_1 and V_2; hence $\delta > 1$. The quantity of secondary electrons released will then exceed the number of primary electrons striking the surface of the screen; thus the potential V_s of the screen which, let us assume, is at first highly negative with respect to the anode, will increase, whereas $|V_a - V_s|$ will decrease. Consequently, some of the secondary electrons then fail to reach the anode and therefore return to the screen. The net secondary emission (I_s) is therefore smaller than I_{s0}. A state of equilibrium is reached when $I_s = I_p$; the screen is then usually slightly negative with respect to the anode [1]). The same applies when the screen is initially positive with respect to the anode ($V_s > V_a$); the charge imparted to the screen is then predominantly negative, the screen becomes less and less positive with respect to the anode and the proportion of secondary electrons reaching the anode increases until $I_s = I_p$.

When $V_a - V_k > V_2$ ($\delta < 1$), the situation is that an excess of negative charge is imparted to the screen regardless of the original screen potential, and V_s becomes more and more negative with respect to the anode potential V_a. Accordingly, the energy $e(V_s - V_k)$ of the primary electrons reaching the screen is less than the energy $e(V_a - V_k)$ imparted to them in the electron gun; hence $V_s - V_k$ decreases until it is equal to V_2, when a stable condition is reached. The secondary emission factor δ is then unity and all the secondary electrons (i.e. a number equal to the number of primary electrons striking the screen) reach the anode [2]), i.e. $I_s = I_{s0} = I_p$.

When $V_a - V_k > V_2$, whereby the screen acquires a potential V_s appreciably lower than V_a, the performance of the tube is seriously affected; firstly, electrons striking the screen with an energy of $e(V_s - V_k) < e(V_a - V_k)$ reduce its brightness below the level consistent with the applied voltage. Secondly, the secondary electrons travelling from screen to anode are accelerated and thus acquire such energies that they may easily ionize molecules of the residual gases present in the tube [3]); the positive ions thus formed are attracted to the negatively charged screen and, striking this with high energy, cause destruction of the luminescing surface and so produce dark patches on the screen.

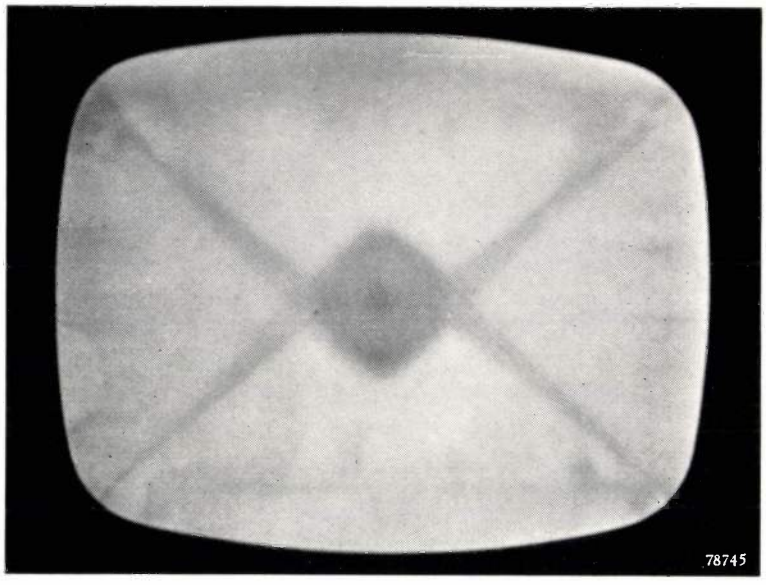


Fig. 2. Photograph of the fluorescent screen of a television picture-tube, showing ion burn caused by positive ions.

A round patch is formed at the centre of the screen by positive ions produced in the neck of the tube. Ions produced elsewhere in the tube are so affected by electric fields (particularly in tubes with rectangular screens) that they form a cross, or butterfly-shaped pattern on the screen (*fig. 2*). After a certain period, which varies somewhat from tube to tube, the luminescent properties of the powder in the areas bombarded by positive ions are noticeably impaired. This effect is especially noticeable when the acceleration potential is temporarily reduced, since the primary electrons then

[1]) The screen potential may slightly exceed the anode potential in certain cases, depending upon the arrangement of the electrodes, the current intensity and the value of δ. The argument contained in this article is based on the premise that the screen becomes slightly negative relative to the anode.

[2]) This effect is described in the literature as "sticking", since $V_s - V_k$ "sticks" at V_2.

[3]) The number of ions formed by an electron per unit path length through a dilute gas depends on the energy of the particular electron in accordance with a curve similar in shape to that of fig. 1. In the case considered here, maximum ionization takes place at energy levels of the order of 100 eV, so that the secondary electrons, having traversed a potential difference of 100 to 200 V, will ionize far more readily than the fast primary electrons.

penetrate only a short distance into the screen and are stopped within the outer layer, where the crystals have been most affected by the positive ion bombardment.

To avoid this, then, it is necessary to ensure that the potential does not surpass V_2 (the "threshold potential") when the tube is in operation; accordingly V_2 must exceed the rated voltage of the tube. Investigations have therefore been made to learn something about the value of V_2 and the conditions affecting it.

Measurement of the threshold potential

The most obvious method of obtaining information about the screen potential is to plot the brightness of the screen as a function of the accelerating potential. *Fig. 3* shows the luminance L of a

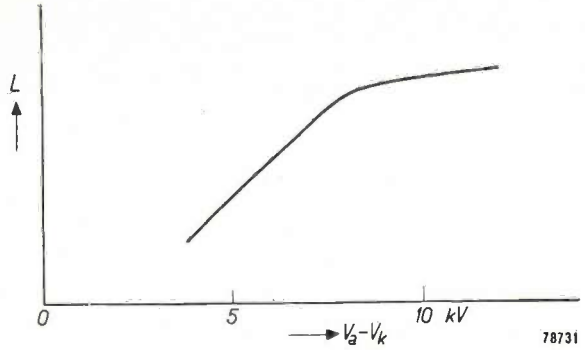


Fig. 3. Luminance L of the screen (on relative scale) as a function of the accelerating potential $V_a - V_k$. The threshold potential (at which the increase in luminance practically ceases owing to the charge on the screen) is here about 8 kV.

screen plotted against the accelerating potential. It is seen that little or no increase in L takes place beyond a certain voltage, which means that $V_s - V_k$ has reached V_2.

This method necessitates certain precautions and is not suitable for large-scale application. The desired information can be obtained more quickly by a purely electrical measurement, in the following manner. A plate B (*fig. 4*) is placed in contact with the window of the picture-tube and connected to the plate C of an air condenser CD. A thin wire

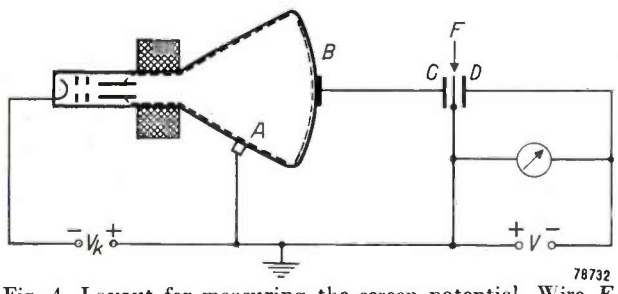


Fig. 4. Layout for measuring the screen potential. Wire F, and the anode A, are earthed. When $V = V_D = V_B = V_s$, the wire is not deflected.

F, which together with the anode A is earthed, is stretched exactly midway between C and D, and a voltage V, the value of which can be read from a voltmeter, is applied to D.

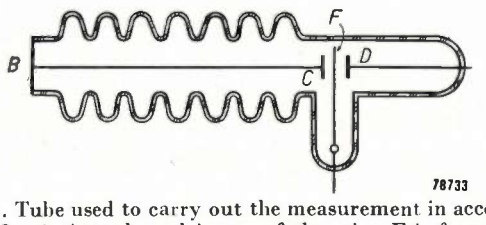


Fig. 5. Tube used to carry out the measurement in accordance with fig. 4. An enlarged image of the wire F is formed on a scale.

Plate B, together with the screen, comprise one of two capacitors in series, the other being formed by C and D. As long as $V_C - V_D \neq 0$, a field occurs around the wire F which deflects it. When $V = V_D$ is varied for zero deflection,

$$V_s = V_B = V_C = V_D = V;$$

hence the potential difference $V_s - V_a$ to be measured can be read direct from the voltmeter ($V_a = 0$).

The condenser system BCD used for this measurement is mounted in a glass tube (*fig. 5*); the plate B is sealed into the end and the tube is

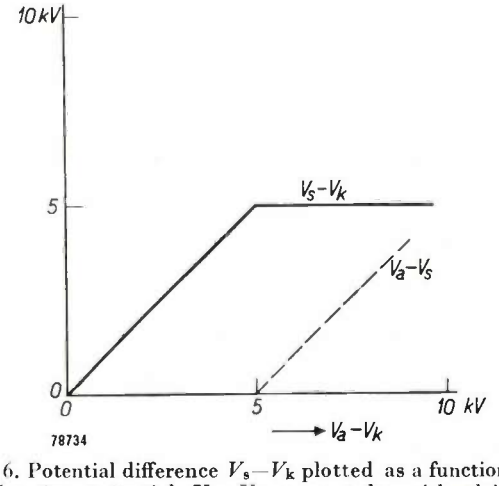


Fig. 6. Potential difference $V_s - V_k$ plotted as a function of the accelerating potential $V_a - V_k$ on a tube with plain glass envelope (no screen). The threshold potential (5 kV) is clearly defined. Dotted line: $V_a - V_s$.

corrugated between B and C to ensure high insulation resistance and thus prevent leakage or tracking between B and F. A lamp and lens are used to project an image of F on to a screen provided with a scale.

Fig. 6 shows the value of $V_s - V_k$ thus measured, plotted as a function of $V_a - V_k$, for a tube without fluorescent screen; in this case, then, the potential of the glass wall was measured. It will be seen that

the increase in accelerating potential beyond $V_a - V_k = 5$ kV is not accompanied by any increase in $V_s - V_k$; hence the threshold potential $V_s = 5$ kV.

A similar curve for a tube with screen is shown in *fig. 7*; as in fig. 6 there is a clearly defined kink, but here $V_s - V_k$ continues to increase gradually beyond the kink instead of remaining constant. Although this continued increase suggests that the secondary emission is not the same for all the crystals, and in all the layers of the screen, the general shape of the curve is the same as that shown in fig. 6 and its resemblance to fig. 3 is evident.

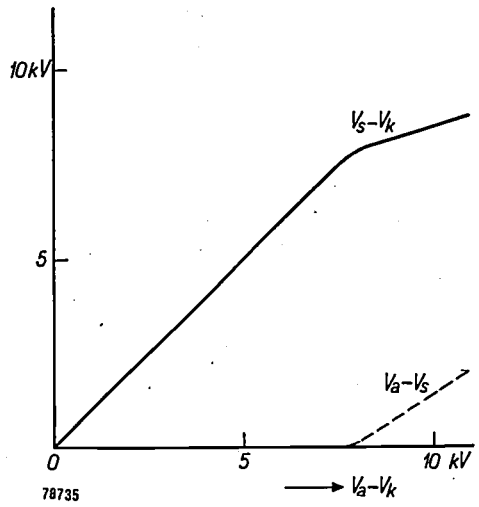


Fig. 7. Potential difference $V_s - V_k$ as a function of the accelerating potential $V_a - V_k$ of a tube containing a screen. The threshold potential in this case is about 7 kV, but is not clearly defined. Dotted line: $V_a - V_s$.

Figures 6 and 7 also show $V_a - V_s$ plotted as a function of $V_a - V_k$, from which it is seen that the screen becomes more and more negative with respect to the anode according as the tube voltage increases beyond the threshold potential.

Since δ and the threshold potential tend to decrease during the life of the tube, it is necessary to ensure that V_2 in a new tube is appreciably higher than the rated voltage. Because of this requirement however it is difficult to measure the threshold potential by the method just described, since this would necessitate the temporary application of voltages far higher than that for which the tube is designed; ultimately therefore a different method of obtaining the required information regarding screen properties was employed.

Direct measurement of δ

By the method now to be described, the value of δ is measured direct. As in the previous method a plate B is placed in contact with the glass tube wall (*fig. 8*), but in this case the plate is earthed

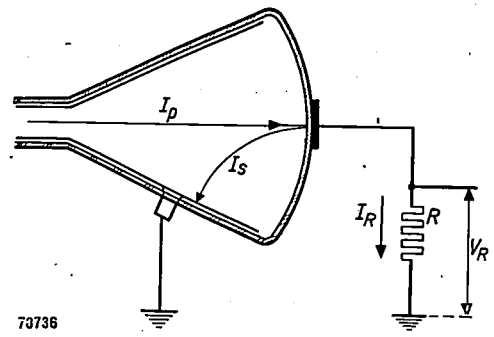


Fig. 8. Method of determining δ by means of a current measurement.

across a resistor and is thus at earth potential in the stable condition.

Let us assume that initially $V_a = 0$ and $\delta > 1$, so that the screen is slightly negative. If a positive voltage ΔV_a (e.g. 200 V) is applied to the anode, the secondary electron current I_s (which, in the stable condition, is equal to I_p) will suddenly increase until it is equal to I_{s0}; hence a positive charge will be imparted to the screen by a current $I_{s0} - I_p = (\delta - 1)I_p$. The process will continue until the potential difference between anode and screen is so small that some of the secondary electrons start to return to the screen, after which the charge imparted to the screen per second will decrease steadily to zero; the stable condition is thus restored. Since the screen induces in plate B a charge which can only be dissipated through R, the above process can be observed by measuring either the current I_R in the resistor, or the potential difference $I_R R$ across the resistor, as shown in *fig. 9a*.

Conversely, if the anode voltage is reduced to zero (whereby the screen acquires a positive potential with respect to the anode) all the secondary

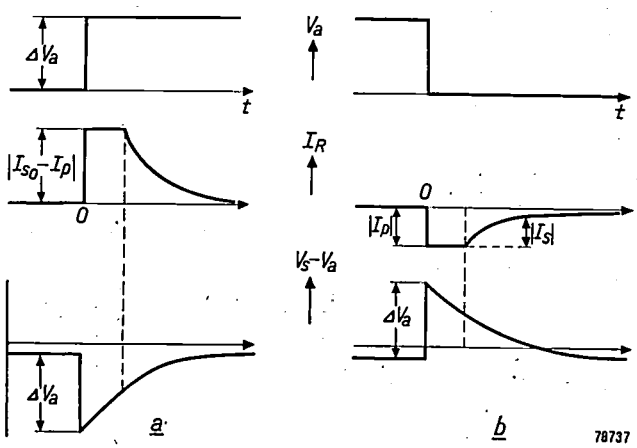


Fig. 9. *a*) Form of the screen current and potential difference $V_s - V_a$ when a positive pulse ΔV_a is superimposed on the anode potential.
b) The same when the anode potential is reduced from ΔV_a to zero.

electrons are temporarily prevented from reaching the anode. The screen is then charged by current I_p and the potential decreases accordingly. After a time, the escape of secondary electrons to the anode recommences, and the charging current decreases until the original stable condition is restored.

In both cases, then, curves representing current I_R plotted as a function of time will exhibit a flat peak, the absolute value of which corresponds in the first case (fig. 9*a*) to $I_{s_0}-I_p = (\delta-1)I_p$, and in the second (fig. 9*b*) to I_p; the value of δ can be derived direct from these two curves.

Method of measurement

Practical measurements are carried out in the following manner. Positive voltage pulses, each of 100 μsec duration, are applied to the anode at the rate of 2500 per second; at the same time the primary electron current is suppressed by the application of a negative voltage to the grid of the tube.

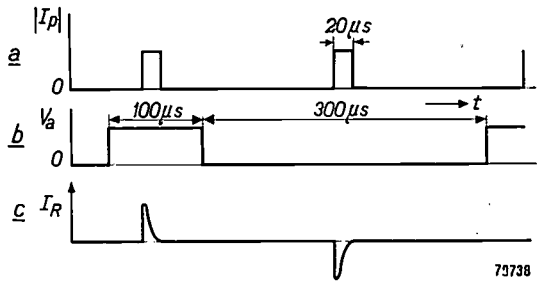


Fig. 10. *a*) Beam current pulses caused by positive voltage pulses on the grid of the tube. *b*) voltage pulses applied to the anode. *c*) form of the screen current I_R.

Half-way through each pulse period of 100 μ sec., a positive voltage pulse of 20 μsec. duration is applied to the grid, whereby the primary electron current is temporarily restored (*fig. 10*); a similar pulse is applied to the grid halfway between each

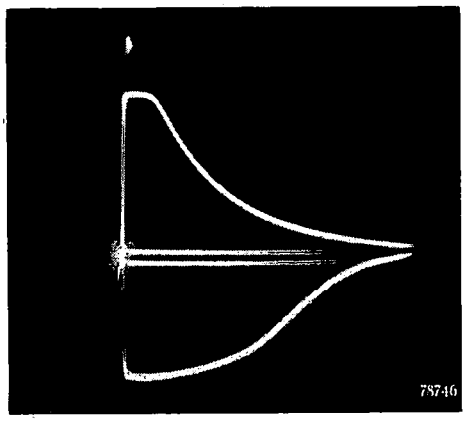


Fig. 11. Oscillogram of voltage $I_R R$, with linear time base synchronized with the grid pulses.

300 μsec interval between successive anode pulses, i.e. at moments when the anode voltage is zero. The situation, then, is slightly different from that demonstrated in fig. 9, but likewise produces char-

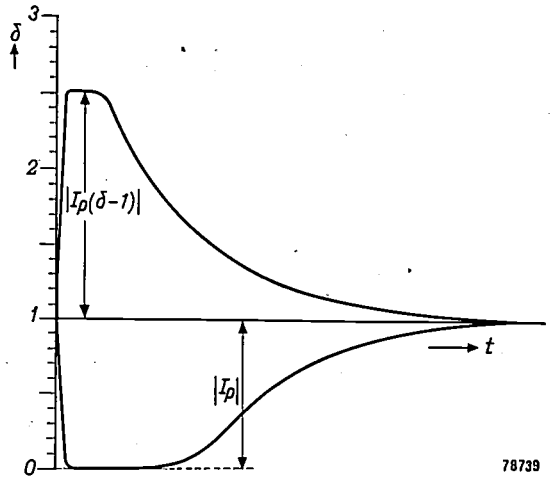


Fig. 12. Method of reading δ direct from the oscillogram.

ging currents exhibiting flat maxima corresponding to $(\delta-1)I_p$ and I_p. The voltage $I_R R$ is amplified and applied to a cathode ray tube, the time base of which is synchronized with the grid pulses. Current curves similar to those shown in figures 9*a* and 9*b* are then described alternately on the screen of the cathode ray tube at a rapid repetition rate so that they form a stationary pattern (*fig. 11*). The

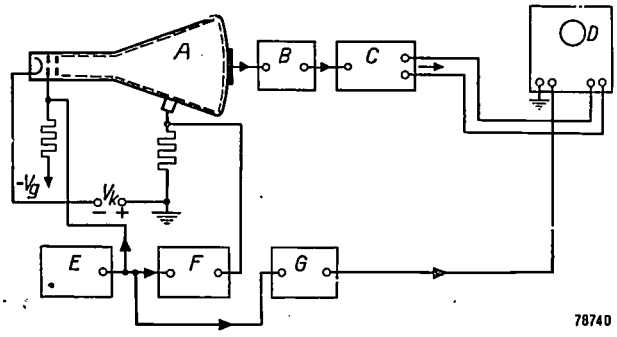


Fig. 13. Block diagram of the electronic equipment used to determine δ. *A* picture-tube, *B* pre-amplifier, *C* output amplifier, *D* cathode ray oscilloscope, *E* generator for grid pulses, *F* the same for anode pulses, *G* time base.

screen is provided with a scale. The amplitude of the current curves is varied so that the maximum of the lower curve is unity; then with the scale set as shown in *fig. 12*, the reading of the upper peak gives the value of δ.

Fig. 13 is a block diagram of the electrical equipment used for generating the pulses; the voltage pulses ΔV_a employed should not be too small, since it is necessary to ensure that all the secondary electrons do, in fact, reach the anode when this is

positive with respect to the screen; in other words, that none of these electrons are forced back to the screen by their space charge.

Effect of the space charge of secondary electrons on the screen potential

The most likely conclusion to be drawn from the above arguments viz. that a value of δ greater than unity is enough to ensure proper tube performance (screen potential almost equal to anode potential; no ion spot produced by positive ions), requires some correction. The combination of screen and anode is comparable to a diode. The screen emits secondary electrons corresponding in current density to the maximum emission current I_{s_0}; the energy of these electrons (a few electron volts) corresponds to a certain apparent temperature of the screen, considered as the cathode of the diode. Although virtually no variation in the energy distribution and therefore in the "temperature" of the secondary electrons, results from variation of the primary electron energy or current, the latter quantities affect I_{s_0} and thus the space charge [4]). The voltage $V_a - V_s$ required to produce a current of secondary electrons $I_s = I_p$ on the anode increases as I_{s_0} (and therefore δ) decreases; hence the potential difference $V_a - V_s$ between anode and screen in the stable condition will be appreciable, even when $\delta > 1$ (e.g. $\delta = 1.1$). The screen thus becomes strongly negative with respect to the anode (resulting in considerable ionization of residual gases and the formation of an ion spot due to ion bombardment). This is demonstrated in *fig. 14*, which shows $V_a - V_s$ versus δ for three values of I_p; it will be seen that a value of $\delta > 1.3$ is required to ensure a sufficiently small potential difference between the screen and the anode.

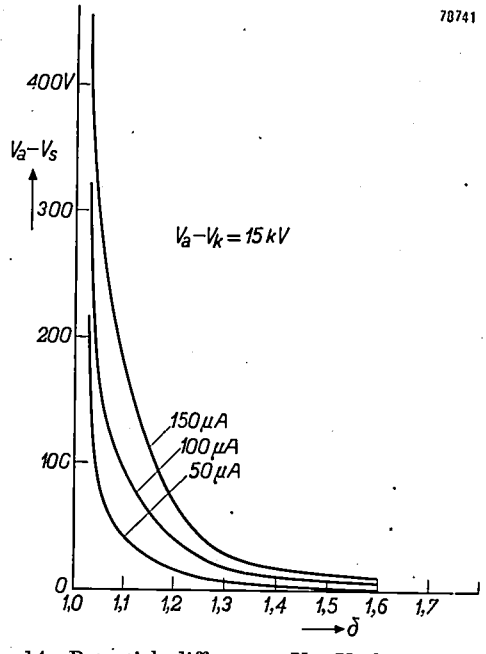


Fig. 14. Potential difference $V_a - V_s$ between anode and screen as a function of δ, for $I_p = 50$ μA, 100 μA and 150 μA. Accelerating potential 15 kV.

The potential that may be acquired by the portion of the glass envelope between the screen and the inner coating of the tube, connected to the anode, is important as well as the potential of the screen itself. In a highly negative state, this glass wall will similarly prevent secondary electrons from reaching the anode; this is seen from *fig. 15*, which is based on measurements carried out on a tube whose screen is covered by a thin layer of metal (of which the potential can be measured), with an insulated metal ring between the screen and the inner coating of the envelope. In this figure, the potential difference $V_a - V_s$ is again plotted as a function of δ for a constant primary electron current, but for different values of the ring potential. The reason for the relative rarity of ion spot in tubes with metal cones will now be seen [5]). The screen is virtually connected to the conductive portion of the tube wall; hence there is no ring-shaped zone of glass capable of acquiring a negative charge.

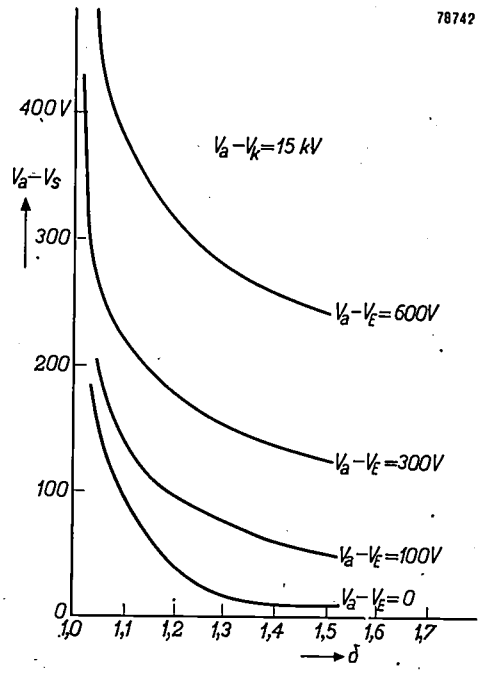


Fig. 15. Potential difference $V_a - V_s$ plotted as in fig. 14 for $I_p = 100$ μA, $V_a - V_k = 15$ kV, for a tube containing a ring-shaped electrode E between the screen and the inner coating, at different values of the potential difference $V_a - V_E$.

[4]) This refers particularly to an electron beam which is sharply focused on the screen, since in these circumstances zones of very high current density and therefore with heavy space charge occur. In the case of a defocused electron beam, the system more closely resembles a plane diode, so that the effect of space charge is much less noticeable. (see fig. 16).

[5]) See Philips tech. Rev. **14**, 281, 1952-53.

Factors affecting the secondary emission

It will be seen, then, that in order to be satisfactory, a tube must contain a screen whose secondary emission factor is appreciably higher than unity (e.g. $\delta = 1.5$—2) at the rated voltage, and cannot fall below a certain critical value during the life of the tube.

There is no question of freedom in the choice of screen material, since this choice is governed entirely by specifications regarding such things as the colour of the fluorescent light emitted and the efficiency of the fluorescence. Moreover it is found that the secondary emission depends upon the binding agent used in applying the screen to the glass wall, as well as upon the composition of the actual screen material. It should be borne in mind that the binder is not confined to the glass and to the phosphor crystals in contact with the glass, but also spreads over the crystals on the anode-facing side of the screen. It is also probable that *changes* may take place in this thin film of binder residue when the tube is in operation, which will cause the secondary emission factor to vary as a function of time (usually to decrease); since it is difficult to assess the manner in which this variation affects the operating efficiency of the tube, it is necessary to find means of improving the secondary emission. This is done by applying a highly emissive material to the screen.

A material suitable for this purpose is magnesium oxide (MgO), this being applied to the screen in the form of an extremely thin film, so thin in fact that the energy of the primary electrons and therefore the efficiency of the fluorescence are not appreciably affected. The secondary emission factor of screens freshly treated in this way exceeds 2, and does not fall below the critical value during the life of the tube.

Fig. 16 shows the potential difference $V_a - V_s$ plotted against the primary electron current for a screen without MgO ($\delta = 1.03$), and with MgO ($\delta = 2.6$); in the former case $V_a - V_s$ may increase to as much as 500 V, whereas in the latter it does not exceed 10 V (at 15 kV accelerating potential and $I_p = 100$ μA),

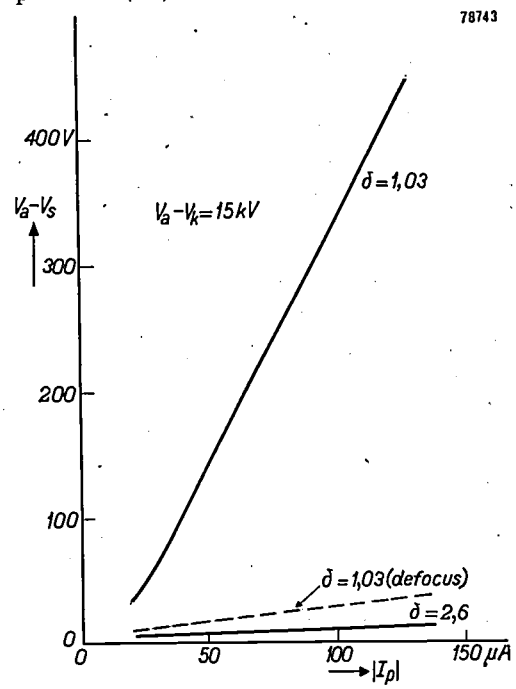


Fig. 16. Potential difference $V_a - V_s$ as a function of the screen current I_p (at $V_a - V_k = 15$ kV) for screens with $\delta = 1.03$ and 2.6; the electron beam is sharply focused. The dotted line represents the corresponding variation for $\delta = 1.03$ in the case of a defocused electron beam (see note [4]).

Summary. It is demonstrated that the secondary emission factor δ of the screen in a picture-tube affects the potential acquired by the screen. When $\delta < 1$ the screen may become highly negative with respect to the anode, thus losing part of its brightness and causing positive ions to be produced, which cause ion burn. Some methods of measuring both the screen potential and δ under operating conditions are described. Negative screen charges and the formation of an ion spot may also take place when $\delta > 1$; a secondary emission factor $\delta > 1.3$ is required to ensure satisfactory tube performance and a complete absence of ion burn. A suitable value of the secondary emission δ has been attained by applying an extremely thin film of MgO to the screen. In tubes processed in this way δ is initially > 2 and does not fall below the critical value during the life of the tube; the deleterious effects caused by inadequate secondary emission, are then avoided.

VOL. 16 No. 2, pp. 33-68 AUGUST 1954

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

A UNIVERSAL APPARATUS FOR X-RAY THERAPY WITH MOVING FIELD IRRADIATION

by H. VERSE *).

621.386.1:615.849

Deep therapy with X-rays, that is, the treatment of tumours sited deep within the body, is one of the most difficult and least rewarding branches of medical practice. Attempts are continually being made to better the conditions in this phase of therapeutics by the application of new technical aids. The apparatus for moving field irradiation described in this article constitutes a further advance in this direction. Of course, the moving field irradiation technique, which with this unit is of quite general application, is by no means capable of effecting a cure in every case. Experience has shown however that with this technique the complicating subsidiary effects of the treatment can be largely eliminated, and that a higher percentage of cures in many kinds of cases can be obtained.

Deep therapy and moving field irradiation

A problem long associated with the X-ray treatment of lesions (tumours) sited deep within the body is that of administering a suitable dose of radiation to the morbid tissues without damaging the surrounding healthy tissues, especially those near the surface of the body. Considered superficially, this is apparently impossible when the irradiation is from an outside source; the dosage rate is always higher on the skin than in the lesion, owing to the usual decrease in radiation intensity with the square of the distance, and to the roughly exponential attenuation of the radiation with increasing depth of penetration. The ratio of the lesion dose to the skin dose can be increased considerably in two ways: firstly by decreasing the relative differences in distance by using a relatively long source-skin distance, or by "compressing" the patient; secondly by employing hard X-rays (high tube voltage and heavy filter) and so obtaining a more gradual decrease in the dose with increasing depth. However, the percentage values of the lesion dose/skin dose ratio, i.e. the *relative depth dose*, obtainable in this way do not exceed about 40% (when the lesion is 10 cm beneath the skin). Hence the success of the treatment depends entirely on how far the radiation *tolerance* of the healthy tissue exceeds that of the morbid tissue. Even if parts of the body outside the lesion suffer no permanent injury, the patient generally takes quite a long time to recover from the heavy load imposed on these parts.

The use of extremely hard radiation (equivalent tube voltage of several million volts) is more favourable from a physical point of view. Owing to the directional effect of the secondary X-rays and electrons generated within the body by the hard rays, a *dosage maximum* is produced below the surface of the body. The dosage maximum can be made to coincide with the site of the lesion.

There is another — long known and inherently simple — way of avoiding the above problem, that is, by irradiating the lesion from several directions (using normal tube voltages of 200 to 250 kV) and so rendering the skin dose innocuous. Multiple field irradiation, cross-fire irradiation and moving field irradiation [1]) (with a moving tube) function in this way.

The continual progress made during the last decade in the development of X-ray tubes and high tension shields, especially with regard to their

*) C. H. F. Müller A. G., Hamburg-Fuhlsbüttel.

[1]) See: G. F. Haenisch and H. Holthusen, Einführung in die Röntgenologie, G. Thieme, Stuttgart 1951.

decrease in size and weight, has led to an entirely new approach to the mechanical problem of moving field irradiation, under far more favourable conditions than before. In particular, these developments have widened the possibilities for a fuller exploitation of the advantages associated with a moving tube. With this in view, the X-ray works of C.H.F. Müller in Hamburg have produced a new constructional solution to the problem of moving field irradiation, namely, the TU 1 apparatus [2]) illustrated in *fig. 1*, which will now be described.

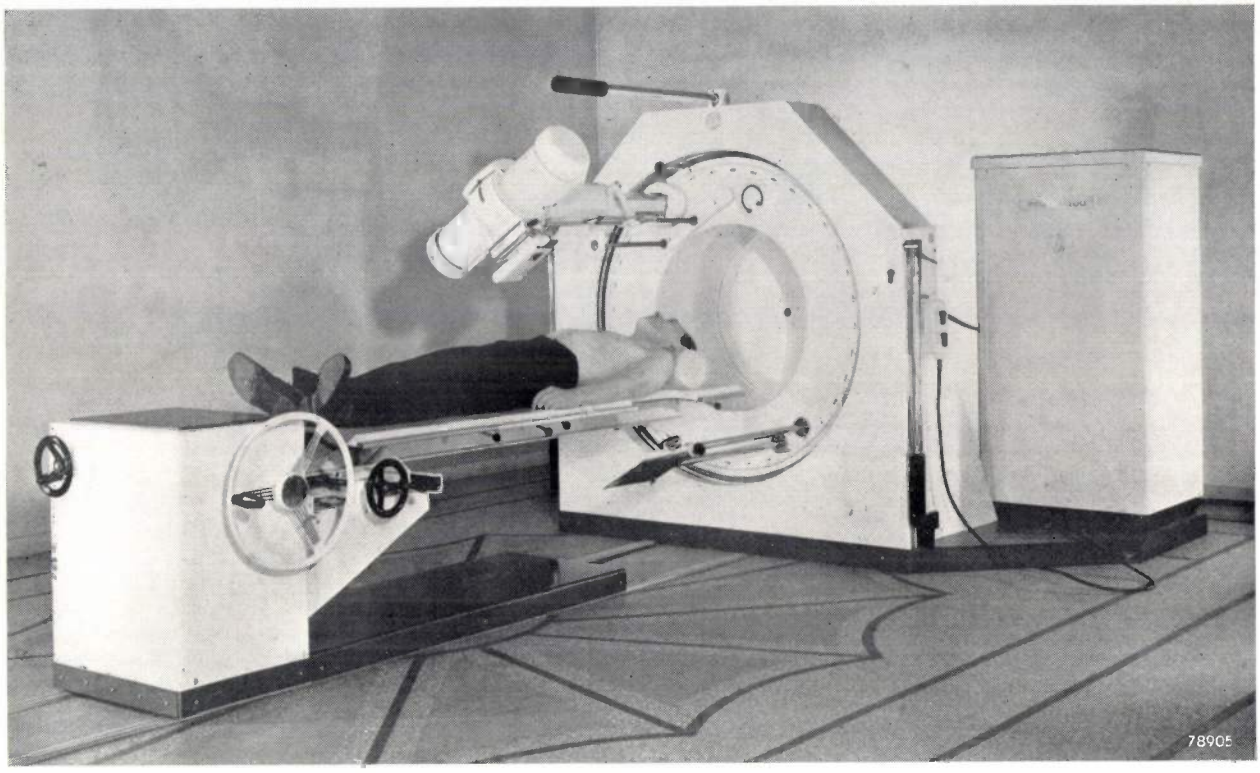


Fig. 1. View of the Müller TU 1 equipment for moving field X-ray irradiation. The treatment-table, running on rails, is seen on the left, the mounting for the X-ray tube and shield at the centre, and the H.T. generator (Müller RT 200) for the X-ray tube on the right of the photograph. The control desk is placed behind a lead glass screen in an adjoining room.

Principles of the design

Briefly, the principle of moving tube irradiation is that during irradiation the X-ray tube is moved in a specific path around the body of the patient, so that, while the cone of X-rays is always directed at the tumour, the region of entry of the rays moves continually from one area of skin to another.

There are many possibilities as regards the shape of the path and the geometrical details; the choice of these features is governed by medical requirements and by constructional limitations. The TU 1 is based on a *horizontal* positioning of the patient (see fig. 1) and a movement of the tube over a circular arc about his horizontal axis [3]). The design of this apparatus was considerably influenced, however by the fact that irradiation over a single circular arc is in some cases not sufficient. This is borne out by the following observations concerning the dosage distribution in the body during moving tube irradiation [4]).

[2]) H. Verse, Einige gerätetechnische Überlegungen zur Röntgenbewegungsbestrahlung, Fortschritte Röntgenstrahlen u. Röntgenpraxis 77, 362-367, 1952.

[3]) According to a method evolved elsewhere, the patient is irradiated in a sitting position (R. Du Mesnil de Rochemont and H. Fiebelkorn, Strahlentherapie **88**, 198-205, 1952). The X-ray tube is then fixed and the patient is rotated about a vertical axis with the aid of a revolving chair during irradiation. We prefer a horizontal attitude of the patient in view of the combination of rotational and traversing movements used. Indeed, to produce the desired concentration of X-rays at precisely the correct point in the body, such a combination of movements can hardly be achieved in any other way than with the patient horizontal and completely immobilized. Moreover, this position is of course more suitable for patients who are gravely ill.

[4]) The data here employed are derived mainly from the investigations of Howard Nielsen; see his book: Rotations Bestraaling, Munksgaard, Copenhagen 1948, and the publication: Rotary Irradiation, Acta Radiol. **37**, 318-328, 1952.

Fig. 2a shows the computed dosage distribution in a cylindrical paraffin-wax phantom 30 cm in diameter irradiated with a stationary tube focus. If no filter is placed in the path of radiation, the dosage at the centre of the cylindrical body, that is, 15 cm beneath the "skin", is only 11% of the surface dose [5]). When the focus is moved so that it describes a full circle about the phantom, "isodose" curves of radial extremely sensitive organs which must not come within range of the X-ray beam, and not useful if the rays in passing through a particular angular region must penetrate heavy bones and thus undergo considerable attenuation before reaching the tumour. The isodoses produced in the above phantom when the tube movement is restricted to particular arcs of the circle are shown in figures

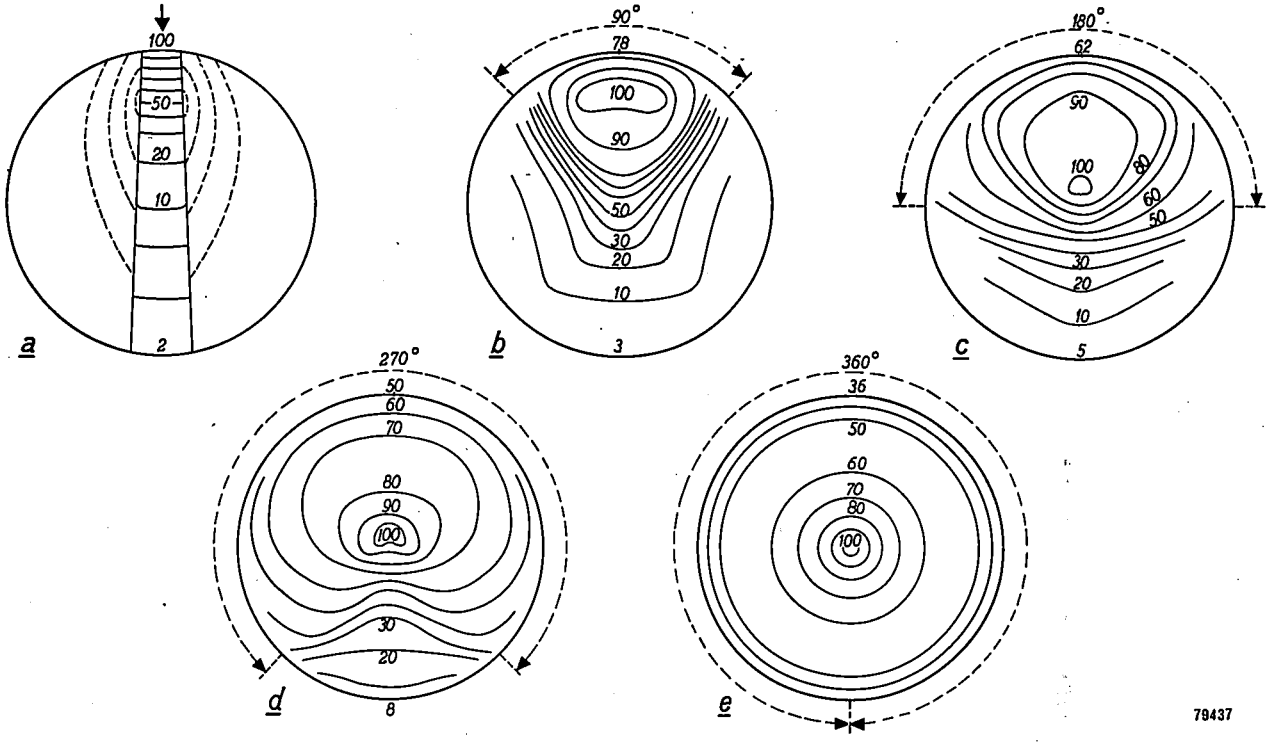


Fig. 2. Isodoses in a cylindrical phantom of paraffin wax 30 cm in diameter, when a field 5 cm wide and 15 cm along the axis of the cylinder is irradiated at a focus-lesion distance of 50 cm, by radiation of half-value thickness 0.7 mm Cu (reproduced from page 47 of the book by H. Nielsen referred to in note [4])).
a) Stationary tube irradiation. The envelope of scattered radiation around the direct X-ray beam is indicated by dotted lines. The lesion dose is 11% of the surface dose.
b) Rotation of X-ray tube through an angle of 90° about the axis of the cylinder. The dosage maximum (100%) is now within the body.
c) Angle of rotation 180°.
d) Angle of rotation 270°
e) Angle of rotation 360°. The isodoses here become concentric circles. The dosage maximum is at the axis of the cylinder; the skin dose is everywhere only 36% of the dosage maximum.

symmetry are obtained (fig. 2*e*). The centre dose is then 280% of that at every point on the surface of the cylinder; this demonstrates the great advantage of moving tube, compared with stationary tube irradiation. In some cases, however, it is neither feasible nor useful to move the tube through the whole 360° angle about the patient: not feasible if a particular region of this angular field includes 2*b*, *c* and *d*. The relative dosage distributions along the diameter of the phantom which passes through the dosage maximum are shown in *fig. 3* for all these cases. It will be seen that the relative depth dose decreases appreciably as the angle of rotation is made smaller; instead of 280%, it is only 130% at the dosage maximum when the angle of rotation is 90°. (The patient is of course so positioned that the dosage maximum lies in the tumour. It may well happen that the disposition of the tumour and the portion of the skin used as a port of entry are such that the pivoting point of the required circular tube movement is not on the axis of the

[5]) The relative depth dose mentioned at the beginning of this article, viz. 40% at a depth of 10 cm, applies to the irradiation of a large field with the aid of a heavy filter, this being the method usually employed in normal deep therapy to increase the depth dose; such filtration naturally necessitates the use of a far more powerful X-ray tube.

patient. As a consequence the dosage distribution may differ considerably from those of fig. 2. This does not affect the qualitative validity of the argument however.)

Thorough investigations have demonstrated [6]) that it is impossible to improve matters either by increasing the focus-skin distance or by using harder rays (employing a filter). A very considerable relative depth dose, however, can still be obtained with a comparatively small angular rotation, if the region of entry of the rays is made to describe *a number* of parallel bands on the skin of the patient, the cone of X-rays being always directed at the same point in the tumour throughout the process (see *fig. 4*). This is termed convergent irradiation. To accomplish it the tube must of course perform in addition another kind of movement such that the rotation about the patient takes place in a succession of different planes.

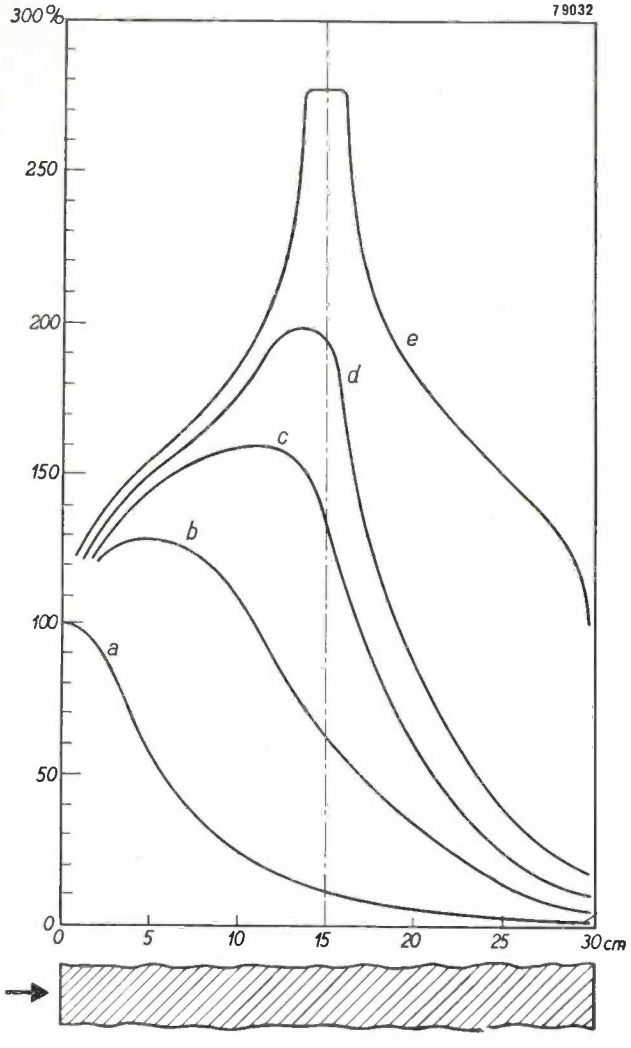


Fig. 3. Dosage distribution along the diameter running through the dosage maximum, of the cylindrical paraffin wax phantom shown in figures 2*a* to 2*e* inclusive. The arrow indicates the direction of the X-rays.

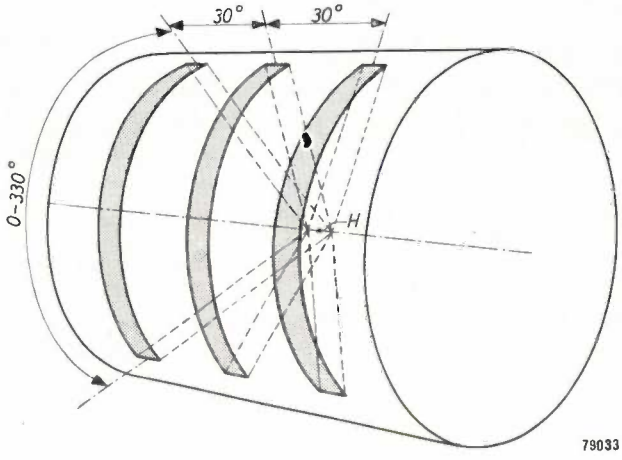


Fig. 4. Convergent irradiation of the cylindrical phantom. The port of entry of the X-ray beam describes parallel bands on the surface of the skin. The cone of rays remains always directed at the lesion *H*.

Such a spatial movement can be obtained in several ways; in the TU 1 it is done by imparting to the X-ray tube a traverse at right-angles to the rotation, which can be controlled independently of the latter. The advantages associated with the above solution and the manner of its application in the practical design will become apparent from the description of the apparatus given below.

Design of the irradiation apparatus

The X-ray generator used in the irradiation apparatus TU 1 is the standard Müller "RT 200" deep therapy unit, the X-ray tube of which is housed in an oil-filled shield. The Philips "250/25" deep therapy unit can be used as an alternative.

Rotational movement

The irradiation apparatus comprises a vertical disc in a fixed frame, mounted so that the disc can be rotated about its axis by an electric motor (*fig. 5*). Mounted near the periphery of the disc is a horizontal arm, to which the X-ray tube is attached in such a way that the X-ray beam emitted is directed towards the axis of the disc. The high tension cables and the oil ducts of the cooling system pass through the (hollow) arm to the shield of the X-ray tube. Since the tubes of both the Müller RT 200 and the Philips 250/25 units mentioned above use a D.C. voltage supply (maximum tube voltages 200 and 250 kV respectively), relatively thin, flexible H.T.

[6]) R. Du Mesnil de Rochemont, Die Dosierungsgrundlagen der Rotationsbestrahlung, Strahlentherapie **60**, 648-674, 1937. M. Nakaidzumi and A. Miyakawa, Über die räumliche Dosisverteilung der Röntgenstrahlen bei der Rotationsbestrahlung, Strahlentherapie **66**, 583-592, 1939.

cables are employed; hence there are no difficulties when the tube is moved.

To enable the patient to be so positioned in the cone of rays that the lesion is at the correct place (that is, in the region of the dosage maximum), the apparatus is equipped with a special treatment-table which runs on rails set in the floor parallel to the axis of the rotating disc; the table top, on which the patient lies, can be adjusted vertically and laterally.

The diameter of the circular path described by the focus of the X-ray tube is 1 metre; hence the focus is (at most) 50 cm from the lesion and, on an average, 30 to 40 cm from the skin of the patient.

In order that the tube shall not strike the table during the rotational movement, the table top is divided into several interchangeable plates one of which, to be placed in the appropriate position, is cut away to allow the tube to pass (see fig. 1).

Two end contacts (fig. 5, E_1, E_2), which can be positioned round the rotary disc, are used to vary the angle of rotation; as soon as either of these contacts touches a corresponding fixed contact on the frame, the direction of rotation of the electric motor driving the disc is reversed. The maximum angular range is 330 °; the angle of 30° not covered includes the heavy iron table girder used to ensure perfect positional stability of the patient.

The rotational movement takes place at a rate of 6° per second.

Traversing movement

The traverse of the X-ray tube is achieved by a movement of the arm (T in fig. 5) parallel to the axis of the disc in a bush (B). This movement is actuated by a second electric motor. Unlike the rotational drive which is mounted on the fixed frame of the apparatus and actuates the disc through a simple chain transmission, the traversing drive is mounted on the rear face of the rotary disc (*fig. 6*) and is connected electrically via a flexible cable to the control desk from which the movement is controlled. To produce the traversing movement of the X-ray tube, a screwed collet mounted in bearings on the disc, and driven by the motor, propels a threaded shaft attached to the tube shield. At the same time another threaded shaft, moving in the same direction as the first but more slowly, acts upon a lever mechanism which changes the position

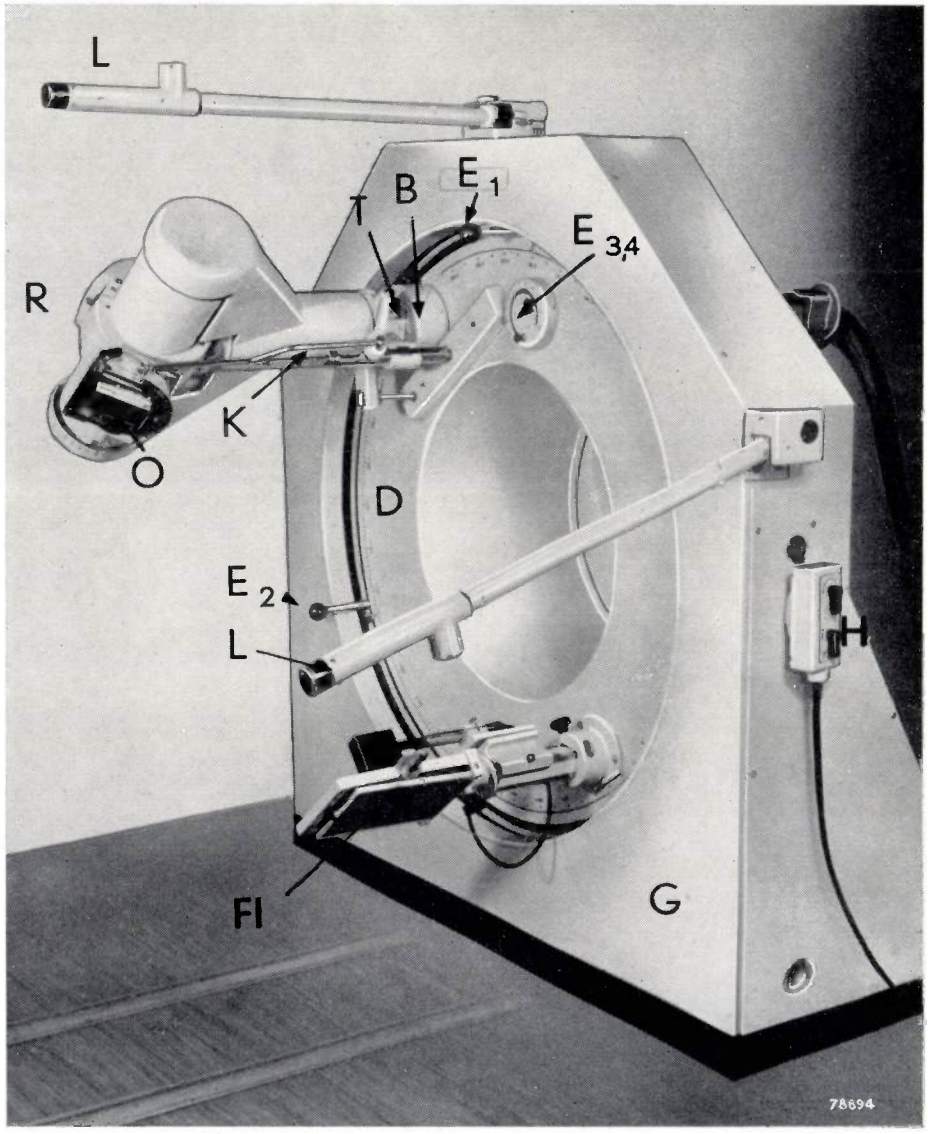


Fig. 5. Mounting for the X-ray tube. G fixed frame, D rotary disc in peripheral roller-bearing on frame, T arm carrying X-ray tube, R X-ray tube with shield and beam aperture O, K coupling rod, B bush in which arm T moves. E_1, E_2 end contacts for rotation, E_3, E_4 adjusting knobs for end contacts of traverse, H hand switch, Fl fluorescent screen, L two of the sources of the light beams described later. (The arrangement of the beam aperture of the apparatus shown in fig. 1 differs in some respects from that of the latest model shown here.)

of the beam aperture in the tube shield (see fig. 5 and *fig.* 7). The pitches of the two threads are such that their relative speeds maintain the projected cone of X-rays always directed at a given point on the axis of the disc.

The range of traverse is controlled with the aid of two end contacts, these being adjusted by means of two knobs attached to the front face of the disc (see *fig. 8*, which also shows other features of the design). The maximum traverse obtainable is 60 cm; this requires a corresponding rotation of the beam aperture of the tube through an angle of 60°. To treat an elongated, say, a more or less cylindrical lesion, different sections of which are to be exposed successively to rotational irradiation, the coupling rod can be removed from the lever mechanism; the X-ray beam then remains aligned in the same direction throughout the traversing movement of the tube.

This rod is likewise uncoupled when the apparatus is to be used for stationary tube irradiation (sometimes combined with compression, which is of course neither practicable nor necessary in the case of rotational irradiation). When uncoupled, the direction of the beam can be adjusted through an arc of 120°.

The normal speed of the traverse is 2/3 cm per second; hence the tube takes 90 seconds to travel from one end of the traverse to the other.

Apart from convergent irradiation along parallel bands, the TU 1 apparatus can be used for irradiation of another kind, which may be termed *oscillatory convergent irradiation* [7]). In this process, the to-and-fro rotational movement of the tube takes place *simultaneously* with a gradual traversing

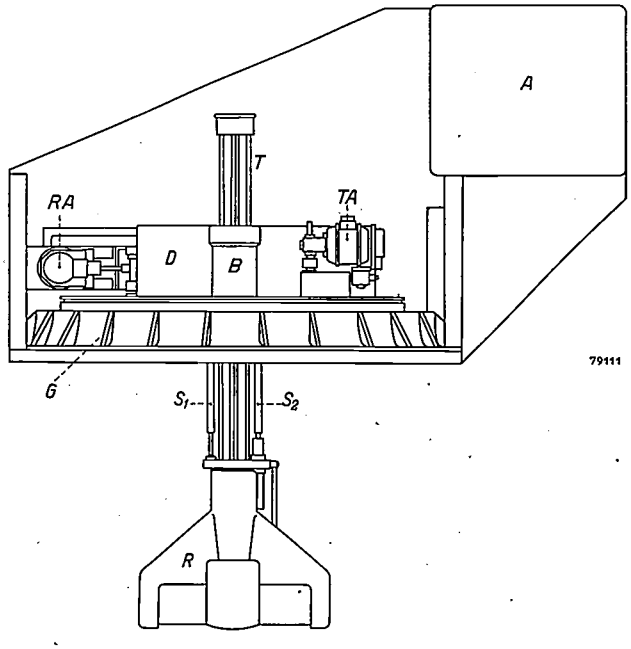


Fig. 6. Top view of main frame and X-ray tube mount with cover plates removed, *G* fixed frame, mounted on the same base as the H.T. generator *A*. *RA* motor unit for rotation, *TA* motor unit for traverse, mounted on rotary disc. S_1, S_2 threaded shafts, other letters as in fig. 5.

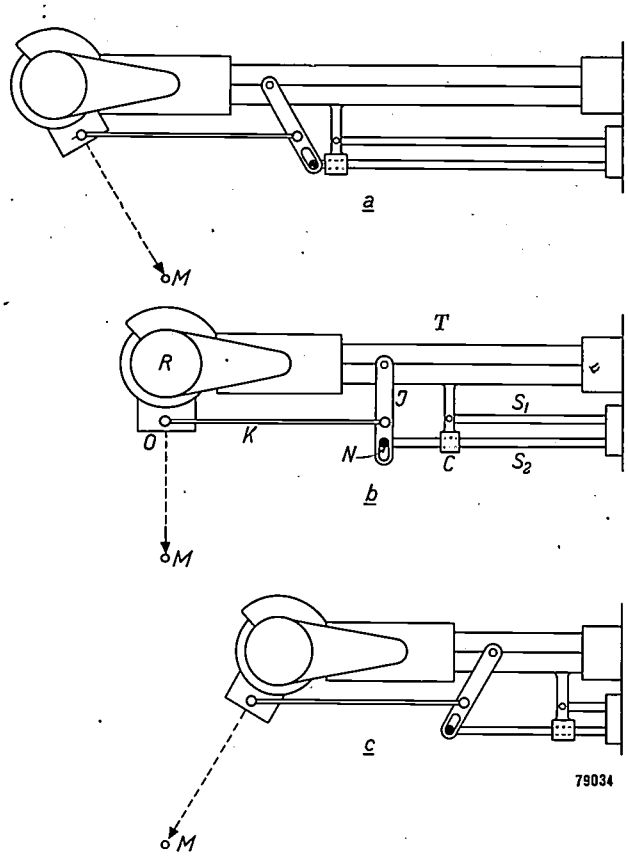


Fig. 7. Mechanism for aligning X-ray beam during traversing movement of X-ray tube *R*. The screwed shaft S_1 (here shown slightly displaced for clarity) is connected to carrier arm *T* and actuates the traverse. The threaded shaft S_2, which moves in slide bearing *C* and is moved more slowly but in the same direction as S_1, governs the position of lever *J* through a pin *N*. This lever (via the coupling rod *K*) moves the beam aperture *O* in the tube-shield through an angle (here shown in the centre and in the two extreme positions of the traversing movement, *a*, *b*, *c*). The central ray of the beam is thus always maintained in alignment with a fixed point *M*, unless coupling rod *K* is removed.

movement at 1/6 of the normal speed referred to in the above; hence the port of entry of the X-rays describes on the skin a zig-zag pattern, as shown in *fig. 9*. It is seen that by virtue of this gradual traverse the coverage is almost completely uniform over the entire area of skin available for the entry of X-rays. Hence this method of irradiation is particularly valuable as a means of effecting adequate distribution of the total skin dose, even in cases where, for medical reasons, the angular range of the rotation is

[7]) H. Wichmann, Physikalisch-technische Bemerkungen zur Bewegungsbestrahlung, Röntgenstrahlen — Geschichte u. Gegenwart (published by C. H. F. Müller A, G.) 3, 72-79, 1953.

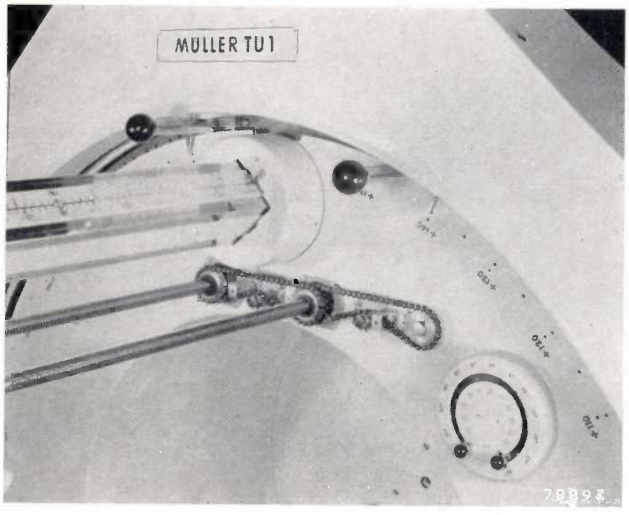


Fig. 8. Close-up of part of the rotary disc. The arm carrying the X-ray tube is seen on the left; beneath it are the two screwed collets, mounted in bearings on the disc and driven by a chain transmission, and the two threaded shafts. The two end contacts for the rotation are seen at the top, and the adjusting knobs for the traverse end contacts in the bottom right-hand corner.

restricted. If the full traversing range of 60 cm is used, the time required for complete irradiation by this method is 9 min. The dose accumulated in this period is sufficient for most cases occurring in practical X-ray therapy.

Preparations and procedure for an irradiation treatment

Positioning of the patient

The first task in preparing for the treatment is to position the patient correctly, so that the lesion (or other anatomically defined site within the body) will be at the pivoting point of the tube movement. Since this pivoting point is on the axis of the rotating disc and is thus fixed in relation to the frame of the apparatus (it is here assumed that the coupling rod K of the lever mechanism shown in fig. 7 is in position), it can be indicated in space with the aid of a system of light beams, fixed in relation to the frame. The light-beams are provided by three light sources carried by hinged arms attached to the frame. A fourth light beam coinciding exactly with the axis of rotation, is produced by a projector at the centre of the rotating disc (*fig. 10*). By moving the table with the patient upon it towards or away from the disc, and adjusting the table top vertically and laterally, all these light-beams are brought to bear upon marks previously made on the skin of the patient. Correct positioning of the patient is thus ensured. A fifth light source can be mounted on the beam aperture of the X-ray tube to supply a light-beam to coincide with the central ray of the X-ray cone. Provided that the patient is correctly aligned, this light beam will always point exactly at the lesion. The three arms referred to above are provided with safety switches which, in the event of excessive deformation of any one of them, switch off all the light sources and thus prevent incorrect positioning of the patient.

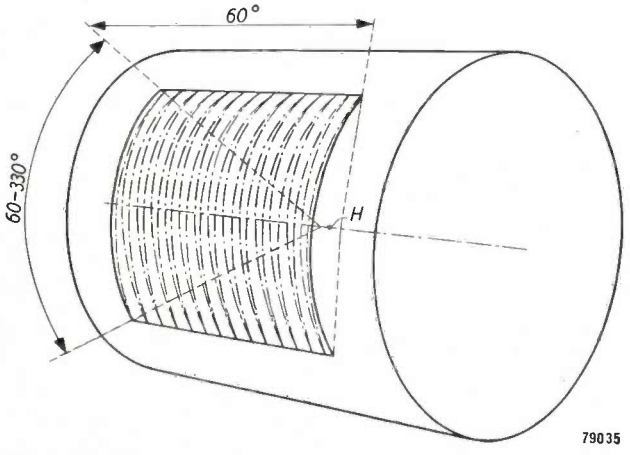


Fig. 9. Zig-zag pattern described by the port of entry of therays on the skin of a patient undergoing oscillatory convergent irradiation. The beam is maintained in constant alignment mith point H.

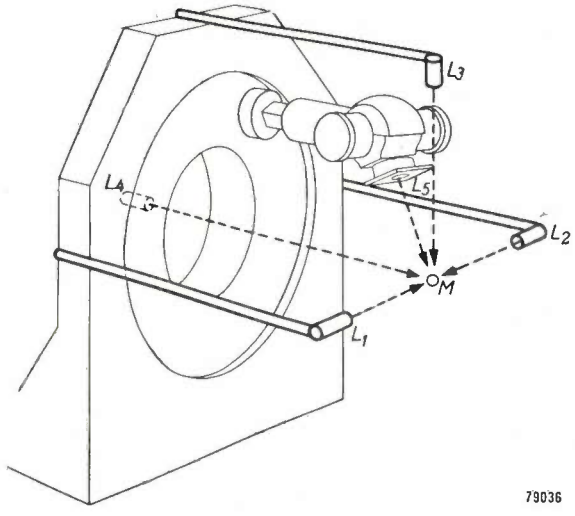


Fig. 10. Method of positioning the patient with the aid of light-beams L_1-L_4. A light source fitted to the beam aperture of the X-ray tube produces the light beam L_5, which coincides with the central ray of the X-ray beam subsequently emitted. The lesion (or other specific point in the body of the patient) should be positioned at the point of intersection M of all these light beams.

The position of the patient can be checked immediately before irradiation with the aid of a small fluorescent screen attached to the rotating disc at a point directly opposite the X-ray tube. Provision for fluoroscopic examination while the beam is vertical (frequently desirable in the case

of a recumbent patient) is made by placing the iron supporting girder off-centre beneath the table, as shown in *fig. 11*. This diagram also shows the different margins of adjustment allowed for positioning the patient. To give as much clearance as possible for moving the patient along the axis of rotation, the centre portion of the rotating disc is deeply recessed (see figures 1 and 5).

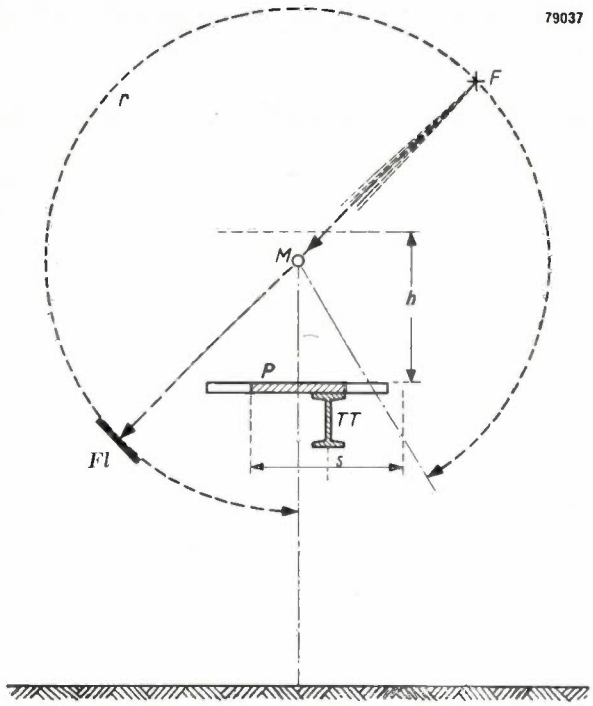


Fig. 11. The top plate *P* of the treatment-table can be moved vertically through a distance *h* and laterally over a distance *s*. *F* is the focus of the X-ray tube and *Fl* the fluorescent screen opposite the tube, with the aid of which the patient can be screen-tested in any position within the angular range *r* of the rotational movement. The iron girder *TT* of the table is placed off-centre to permit the use of the fluorescent screen whilst the X-ray beam is vertical.

Control and limitation of the movement

Since the apparatus is equipped with two separate motor units for the rotational and traversing movements, both of which can be controlled independently through a system of relays, a great diversity of movements can be made to take place automatically. To ensure the utmost simplicity and clarity in operation, only three of the many possibilities have been selected; these correspond to the methods of stationary tube irradiation, rotational irradiation and oscillatory convergent irradiation as outlined above. Experience has shown that most of the cases occurring in practice can be covered by these three methods; moreover, these methods are precisely those for which the most complete therapeutic experience and dosage data are available.

The movement of the tube during irradiation is governed by a controller (*fig. 12*) mounted on a desk, placed behind a screen of lead glass outside the irradiation room. The three positions of the selector seen at the top of the controller correspond to the three methods of irradiation mentioned above. When stationary tube irradiation is selected, all the mechanisms for the various movements of the X-ray tube are locked, and the only remaining control is to switch on the radiation for the required period at the control desk. When the selector is turned to the position for rotational irradiation, the tube starts its movement automatically as soon as the H.T. is switched on at the control desk, and continues to travel to and fro between the two pre-adjusted end contacts until the expiry of the period set by the time switch of the apparatus. The tube then stops and the tube voltage is switched off. If several parallel bands are to be irradiated, the master switch seen in fig. 12 below the method selector can now be used to shift the tube in the direction of traverse to the next irradiation band. Provided that the end contact for the traverse is pre-set to the required distance the tube may be allowed to proceed to the contact, its arrival there being indicated by a signal lamp; the rotational irradiation described in the above can then be repeated. On completion of this process, the master switch is turned to the other position and the tube then proceeds to the oppositie pre-set end contact of the traverse; when the signal lamp indicates that

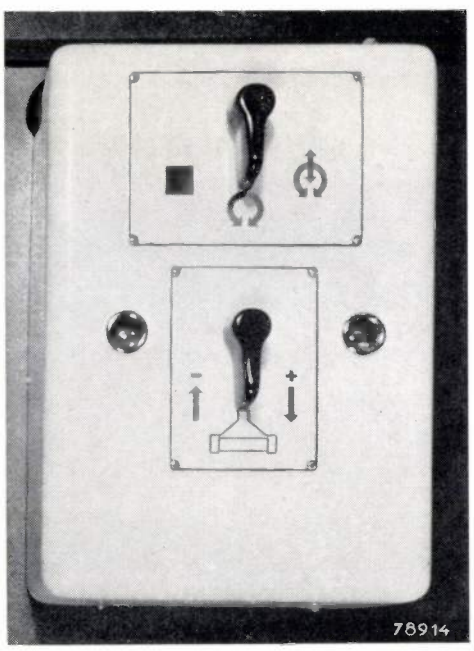


Fig. 12. Controller for automatic moving field irradiation. The selector for the three methods of irradiation is seen above, and the master switch for the traversing movement below.

this contact has been reached, the third band is irradiated [8]). A locking mechanism is provided as a safeguard against accidental manipulation of the master switch during the rotational movement. To prevent accidental incorrect irradiation of the patient, this motion is governed by a centrifugal switch on the appropriate electric motor; no H.T. can be applied to the X-ray tube, and hence no radiation can be produced, unless the motor is running. Should one of the end contacts, or one of the associated change-over relays for the motor fail to operate, the entire apparatus is put out of action automatically by special safety end contacts so that the patient cannot be harmed, or the apparatus damaged by such failure.

Finally, let us consider the third position of the selector, that is, oscillatory convergent irradiation. When this method is selected the speed of the traverse motor is reduced by a switch-over to 1/6 of the normal speed. As soon as the tube voltage is switched on, the tube starts its combined traversing and reciprocating movement. The time switch of the X-ray apparatus is set to the exact time required to cover the range of traverse between the particular positions of the end contacts; hence the radiation and the movement of the tube are switched off simultaneously as soon as the tube completes this movement. The traversing movement is also governed by a centrifugal switch on the motor and by safety end contacts in the manner described above for the rotational movement.

With the patient correctly positioned, the end contacts limiting the movement of the tube are adjusted, using a hand switch (*H* in fig. 5) to move the tube. Removing this hand switch from its hook automatically locks the tube voltage switch and thus safeguards the operator performing the adjustment against accidental exposure to X-radiation. The X-ray tube can be shifted at will in the directions of rotation and traverse by means of two levers on the hand switch. The tube positions corresponding to the limits of the desired movement can be located quite simply with the aid of the light-beam concident with the direction of the X-rays, and the end contacts can then be set to these positions. In critical cases the entire pattern of the irradiation can be checked before it is administered, using the hand switch.

On completion of these preparations, the arms carrying the light-pointers are swung back to the frame, the light source is removed from the beam aperture and a lead diaphragm is inserted in its place in the holder. Initially this diaphragm leaves an aperture corresponding to the size of the floures-cent screen used for the preliminary fluoroscopic check. Before irradiation is actually started, however, the diaphragm is pushed further into the holder so that it limits the cone of X-rays to the cross-section corresponding to the size of the lesion to be treated. If necessary a filter can be inserted in a holder in front of the diaphragm (as already mentioned, it is in principle unnecessary to use a heavy filter in moving field irradiation). The patient is then left alone in the irradiation room and the pattern of irradiation is regulated from the adjoining control room in the manner described.

We shall finally refer very briefly to a problem which is extremely important in this and every other method of deep therapy, viz. the determination of the lesion dose. In general, the lesion dose is measured with the aid of small ionization chambers, placed either in the immediate vicinity of the lesion during irradiation or in an equivalent position in a phantom previously exposed to a trial irradiation. In the case of rotational irradiation involving an almost completely circular movement of the X-ray tube it is possible to employ a simpler procedure, which consists in measuring simultaneously the dosage rate of the tube and the dose transmitted through the patient [9]). The dosage rate of the tube is measured continuously by means of a large, flat ionization chamber, mounted between the filter holder and the diaphragm holder on the beam aperture in the shield (*fig. 13*), and connected to an indicating instrument on the control desk; the transmitted dose is determined by an integrating measurement with a sensitive ionization chamber mounted in the central ray of the X-ray beam behind the patient, on a movable mount attached

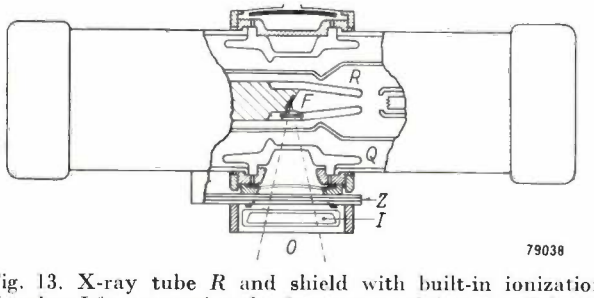


Fig. 13. X-ray tube *R* and shield with built-in ionization chamber *I* for measuring the dosage rate of the tube. *F* focus, *O* beam aperture, *Z* filter mount, *Q* oil-bath for insulation and cooling.

[8]) The simple method of construction described here is of course possible only by reason of the fact that no more than three irradiation bands are employed, the middle band being irradiated first.

[9]) W. Neumann and F. Wachsmann, Strahlentherapie **71**, 438-449, 1942.

to the fluorescent screen holder (see fig. 5). Although this method, as already noted, is applicable only to angles of rotation in the region of 360°, a new method has recently been developed whereby the dose absorbed by the lesion can be determined for other angles of rotation, however small, without the aid of phantom tests [10]).

Conclusion

In conclusion it may be worth mentioning once again the great scope in the choice of movement available to the user of the TU 1. This is primarily attributable to the fact that the apparatus is equipped with two separate, electrically controlled units, one for each of the two degrees of freedom of movement. This arrangement provides a method that is in a high degree automatic and largely foolproof against possible errors, without appreciable curtailment of the therapeutic possibilities. In fact there is every reason to suppose from the experience gained so far, that this apparatus gives a general solution to the problems associated with normal X-ray deep therapy.

[10]) H. Wichmann, Tabellen zur Dosierung bei Bewegungsbestrahlung, to be published shortly.

Summary. During X-irradiation of deep sited lesions in the body, the harmful effects of the rays upon adjacent organs and in particular upon the areas of skin exposed to the rays, can be obviated by moving the X-ray tube around the patient in a suitable manner during the process of irradiation. Owing to the relatively small size and weight of modern X-ray therapy units, this movement can now be made to take place automatically. The Müller TU 1 apparatus is designed for this purpose. Here the X-ray tube is capable of rotation about the horizontal patient, and can perform a traversing movement parallel to his longitudinal axis. Separate motors are used to actuate the two movements. Control of the motors is such that a very wide variety of motions can be accomplished automatically. In the TU 1 apparatus three automatically controlled methods of irradiation are provided, i.e. stationary tube, rotational and oscillatory convergent irradiation. This article includes a description of the equipment and a brief account of the provisions made to ensure correct positioning of the patient and to limit the movement of the tube to the desired range.

CHEATER CIRCUITS FOR THE TESTING OF THYRATRONS

I. MEASUREMENT OF GRID CURRENT

by M. W. BROOKER *) and D. G. WARE **).

621.387:621.385.38

The range of thyratrons now being produced commercially includes certain types which operate at high anode voltages and draw high currents. The conventional ways of thoroughly testing these valves under their full-load operating conditions would make excessive demands on the power supply. Attempts have therefore been made to devise "cheater circuits" which simulate full-load operation at only a fraction of the power load and cost of energy.

In Part I of their article the authors describe a cheater circuit which allows the measurement of grid current just before the valve strikes. The pre-striking grid current is an important check on valve quality. Part II will deal with another form of cheater circuit, designed for the life-testing of thyratrons.

In a high-vacuum triode with unsaturated emission the flow of anode current is limited mainly by a space charge of electrons. This in turn can be controlled at will by variation of the negative potential of the grid. In a gas-filled triode (thyratron) there is no such control and the grid acts only as a kind of trigger which determines when the valve shall fire. This, too, means a certain controllability, though of a slightly different nature from that in a vacuum tube. When the voltage applied to the grid is made less negative and ultimately reaches a certain critical level, anode current commences to flow; the positive ions produced neutralize the electronic space charge and the anode current rises rapidly to the value prescribed by the circuit parameters. This process is known as the firing or striking of the valve. Once a gas-filled valve has struck, the grid has no control over the current, as the effect of its potential is masked by a sheath of positive ions.

When a gas-filled valve is in the non-conducting state with the grid biassed negatively, a grid current will flow. This is due to a number of effects, which will be explained later. The pre-striking grid current flows through the grid resistor and may disturb the accuracy of control of the valve. Indeed, if the grid current becomes relatively large and the grid resistance is high, it may be quite impossible to control the valve. It is therefore necessary to have an accurate method of measuring the pre-striking grid current, both as an aid in design and as a check on quality.

This article first reviews some known methods of measuring grid current in thyratrons and then describes better methods, which are simple to use and which measure grid current under conditions identical with the full-load valve operating conditions.

*) Mullard Radio Valve Co., Ltd., Mitcham, England.
**) Formerly with Mullard Radio Valve Co. Ltd.

Conventional methods of measuring grid current

The grid current is a thyratron before it strikes is actually the sum of several currents flowing in the same sense. It has components due to:

1) electrical leakage between the electrodes,
2) the flow to the grid of positive ions produced by electrons escaping control by the grid ("uncontrolled cathode emission"),
3) primary thermionic emission from the grid itself, and perhaps
4) photo-emission from the grid.

It is possible to measure some of these components individually. Thus the leakage current alone can be found by normal insulation measuring methods. The current due to uncontrolled cathode emission can be measured by putting a micrometer in the grid circuit and applying the rated anode voltage; the grid current-grid voltage may then be plotted. (The microammeter should be protected against overload, as the grid current rises to a high value when the valve strikes.) But the real problem is to measure it under actual operating conditions. It is in securing these conditions that the main difficulty lies.

The basic idea employed by most of the known methods of measuring the pre-striking grid current of gas-filled valves are given below.

The switch method

This method is very simple to set up and simple to use, but it only gives an approximate idea of the grid current under working conditions.

The procedure is to operate the valve with a resistive load at its rated mean anode current for a time sufficient to raise all the electrodes to their working temperatures. A switch is then operated which disconnects the load and connects anode to cathode and at the same time connects a negative voltage of, say, 100 V to the grid via a microammeter. The current read on this meter is called "the" grid current.

The advantages of this method can be listed as follows:

1) The method can be applied very easily to any type of thyratron.
2) The equipment is cheap.

The disadvantages are equally obvious, thus:

1) The current recorded decreases due to the valve electrodes cooling.
2) Leakage current (if present) is incorrectly measured owing to anode and grid voltages being abnormal.
3) Uncontrolled cathode emission is absent.
4) Ionization due to emission from other electrodes such as screens and anode is absent.

The cut-off method

A method generally used in the United States, known as the cut-off method, consists in operating the valve from an A.C. power supply, which may be the 220 volt mains, through a resistor which is adjusted to give the maximum rated mean anode current when the valve conducts for complete half cycles. After running like this for a long enough period for the valve to reach temperature equilibrium, the grid voltage is made more negative until the valve just fails to re-strike, and the voltage V_{g1} at the input side of the grid resistor is noted at the cut-off point. This is done with a low and a high grid resistance (R_{g1}, R_{g2}). If the critical grid current is represented by $I_{g\text{crit}}$, the critical voltage at the grid itself is given by $V_{g1} - I_{g\text{crit}} R_{g1} = V_{g2} - I_{g\text{crit}} R_{g2}$, V_{g1} and V_{g2} being the voltmeter readings. From this we find:

$$I_{g\,\text{crit}} = \frac{V_{g2} - V_{g1}}{R_{g2} - R_{g1}}. \quad \ldots\ldots \quad (1)$$

This method has a number of advantages over the previous one, viz.:

1) The grid voltage at the point at which the measurement is made is the normal critical grid voltage. Thus the electrical leakage current included in the measuring results is correct.
2) Any uncontrolled emission will be measured, but as this depends upon the applied anode voltage, it is necessary to perform the test at the full rated anode voltage if any accurate result is to be obtained.
3) Grid emission is also measured at the proper voltage.

However, the method does not allow the full rated *peak* anode current to be drawn, as for a half-sine waveform, the peak-to-mean anode current ratio hardly exceeds 3, while modern thyratrons are rated for peak-to-mean ratios of 10 or even more. This disadvantage is important, because in general the valve voltage drop is higher when high peak currents are drawn; therefore the valve runs at a higher temperature, and the grid current is greater under such conditions.

Another disadvantage is that as the grid voltage is made more negative toward the cut-off point, the original half-sine waves diminish to quarter sine waves and the mean anode current correspondingly decreases to half its original value. Now in order to obtain an accurate reading of cut-off voltage, it is necessary to raise the negative grid voltage rather slowly, with the result that the valve cools appreciably during the time taken for the measurement. This of course means that the measured grid current is lower than the actual working grid current.

On the other hand, on switching from a low grid resistance to a high one, striking will occur earlier in the cycle, so the mean anode current will increase, making the valve hotter and this in turn increases the grid current. It was found that in some cases the grid current changed so rapidly that it was not possible to obtain any steady operating condition with a high grid resistance. This, of course, occurred with defective valves where the total grid current which flows is high, but for test purposes it is essential to be able to measure grid currents up to extremely high values.

The modified cut-off method with feedback

The drawbacks of the standard cut-off method have led to the design of a modified method. Here, the difficulties mentioned have been avoided by means of negative feedback. This method, which was the first to be developed by the authors, proved of considerable value in early thyratron development work by Mullard.

The basic method is shown in *fig. 1*. The anode voltage of the test valve is supplied from 220 V A.C. mains. The grid circuit includes a stabilized D.C. supply adjustable by a potentiometer supplying a positive voltage to the grid. A variable load resistor R_1 is in series with the cathode, and a group

of resistors R_g can be switched, one by one, into the grid circuit.

When an anode current flows, a rectified half-wave voltage appears across the load resistor R_1. This voltage is smoothed by a filter L_1-C_1-L_2-C_2 and appears as a negative voltage at the grid of the valve

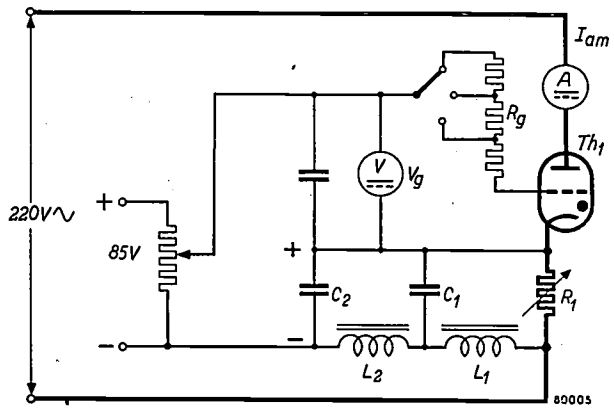


Fig. 1. Cut-off method for measuring the grid current of a thyratron (Th_1), with negative feedback. The D.C. voltage developed across the load resistor R_1 is smoothed by a filter L_1-C_1-L_2-C_2 and fed back with negative polarity to the grid of Th_1, in opposition to an adjustable D.C. voltage from a stabilized supply (85 V). The resulting grid voltage V_g is read on a voltmeter, the mean anode current I_{am} on an ammeter. By means of a switch, various resistors R_g can be inserted into the grid circuit.

This negative voltage is partially counterbalanced by the stabilized supply, which acts in the opposite sense. Thus the striking of the valve is controlled by a (negative) grid voltage, which increases linearly with the mean anode current.

The self-compensating action of the circuit is best understood by reference to *fig. 2*. Consider the effect of a rise in grid voltage from level *1* to level *2* caused by a change in current flowing through the grid resistor. The valve will strike earlier in the cycle, the "firing angle" being reduced from α_1 to α_2.

The mean anode current and the mean voltage across R_1 will therefore increase. A greater negative D.C. voltage is therefore fed back to the grid of the valve, thus tending to keep the anode current steady.

This self-compensating effect ensures stable running at a steady anode current, providing no attempt is made to fire the valve near the peak of the anode voltage curve. With firing angles greater than about 80°, instability and oscillation may set in because of the time delay in the filter circuit.

Using this new method, the variations of the mean anode current are reduced to a much lower level and it is easy to maintain a steady anode current, which is set by adjusting the load resistor.

The resultant grid voltage applied to the thyratron is read on the voltmeter V_g of fig. 1, and the grid currents may then be measured with a firing angle of, say, 45°, as described above under the heading *The cut-off method.*

The chief advantage of the method just described is that measurements of grid current can be made without affecting the working conditions. However, it still suffers from the disadvantage that the peak anode current cannot exceed 5 times the mean anode current, owing to the firing angle being limited to about 80° [1]). It is possible to overcome this disadvantage by superimposing on the D.C. bias an A.C. voltage (sinusoidal, or better still, pulse-shaped), leaving the rest of the circuit unaltered. By phase-shifting the A.C. voltage, firing angles up to 180° and smooth anode current control down to zero can then be obtained.

In the next section a method of measurement will be described which makes use of pulse firing to obtain stable operation at high peak-to-mean current ratios.

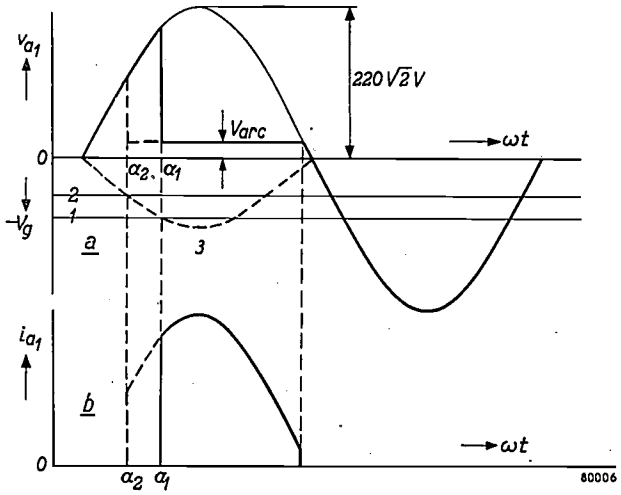


Fig. 2. Explanation of stabilizing effect of feedback in fig. 1. A rise in grid voltage V_g from level *1* to level *2* decreases the firing angle (determined by the point of intersection with the dotted critical grid voltage curve, *3*) from α_1 to α_2, thereby increasing the mean anode current I_{am}, i.e. diminishing V_g. This tends to keep I_{am} steady. V_{a_1} voltage across Th_1; V_{arc} arc voltage; i_{a_1} anode current; $\omega = 2\pi \times$ mains frequency; t time.

Cheater circuit

A disadvantage common to all methods mentioned above — so far as high-power thyratrons are concerned — is that the testing makes an excessive demand on the power supply, especially when voltages much higher than 220 V r.m.s. are to be used. Attempts have therefore been made to devise special circuits which simulate full-load operation at only a fraction of the power.

[1]) As can be shown by a simple calculation, the peak-to-mean anode current ratio for firing angles α up to 90° is given by $2\pi/(1+\cos\alpha)$, and for α between 90° and 180° by $2\pi \sin\alpha/(1+\cos\alpha)$, see *fig. 3*. The latter expression tends to infinity when α approaches 180°.

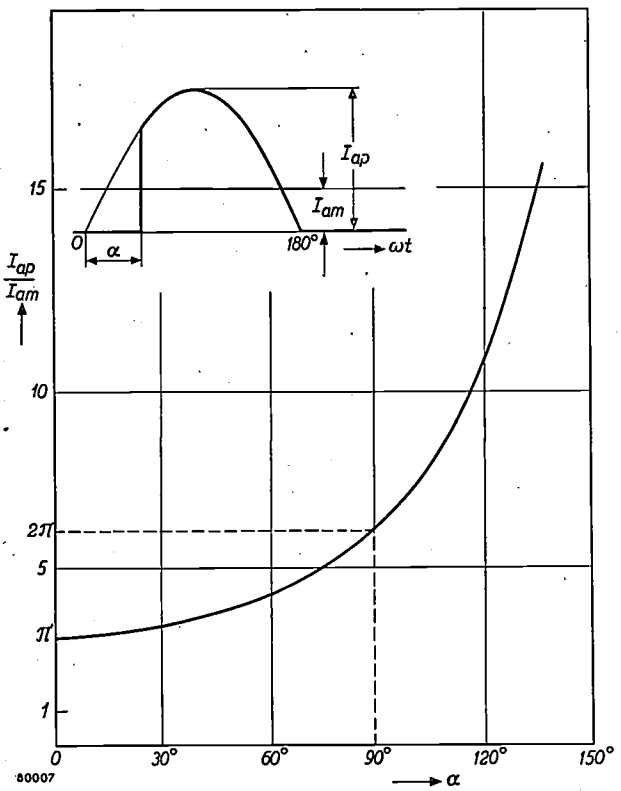


Fig. 3. Ratio of peak anode current I_{ap} to mean anode current I_{am} as a function of the firing angle α.

A characteristic fundamental to all thyratrons makes possible such circuits: the fact that once a valve has struck, the voltage across it drops to a low level, which is almost independent of the anode current. Although the peak anode voltage before striking may be, say, 1500 V, yet when carrying its full rated anode current the anode voltage V_{arc} is less than 20 V, just sufficient to maintain ionization (fig. 2). Providing that the voltage applied to a valve does not fall below this maintaining potential (arc voltage), the valve continues to conduct. If, then, a circuit can be devised which switches automatically from a high to a low voltage supply immediately after the valves has struck, only a low surplus voltage will have to be taken up by the load, and the power economy will be considerable. Such a circuit is known — for obvious reasons — as a *cheater circuit*.

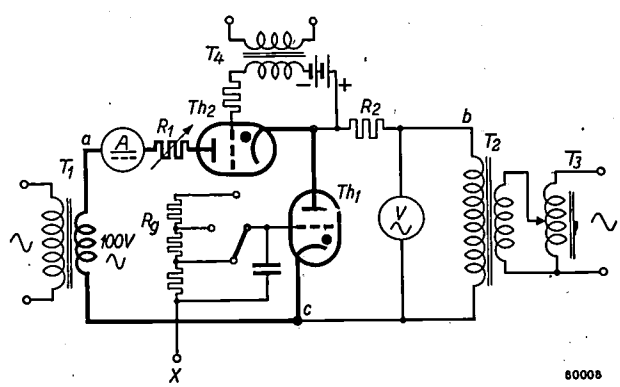


Fig. 4. Cheater circuit for measuring grid current under full operating conditions. Th_1 thyratron under test. Th_2 auxiliary thyratron. T_1 low voltage transformer (100 V). T_2 high voltage transformer, fed from variable autotransformer T_3. R_1 load resistor. R_2 current limiting resistor in H.T. circuit. R_g bank of grid resistors. Firing pulses are applied at X and also to transformer T_4 (see fig. 6).

To provide a versatile circuit, useful in certain other tests, a special cheater circuit was devised. Its mode of operation is based on the same principles as the methods previously described for the cut-off method: the grid current is deduced from the change observed in the critical grid voltage at a given anode voltage when a resistor is switched into the grid circuit. Once again a D.C. bias is used for control and measurement, but the system of firing the valve is changed. A pulse triggering circuit is introduced, which ensures reliable firing at any point in the positive anode voltage half-cycle, and also plays an important part in the cheater action.

The main circuit of the new apparatus is shown is *fig. 4*. The valve under test Th_1 is connected to a 100 V A.C. supply, in series with a load resistor R_1 and an auxiliary thyratron Th_2. It is connected also, via a large series resistor R_2, to a high voltage A.C. supply. The high and low voltage supplies are connected so as to be in phase. In the grid circuit of Th_1 there is the usual bank of resistors R_g used in the actual measurement of grid current.

The cheater action of this circuit is best explained diagrammatically. *Figs. 5a, b* and *c* show the waveforms of the voltages appearing across Th_1 and Th_2 and the current through Th_1. Consider first Th_1. In its non-conducting state, the voltage across it is the high one supplied by the high-tension transformer T_2. If now the valve is caused by a pulse to strike at P, this voltage drops to about 10 V; because of R_2, however, the current which flows through the test valve is small, say 10 mA. Considering now Th_2, during the first part of the high-voltage cycle, the cathode of Th_2 is driven positive, which is, of course, equivalent to the negative voltage shown in fig. 5*c*. This negative voltage is much greater than the 100 V supply to the anode of Th_2, which is therefore non-conductive. When Th_1 has struck, however, the large negative voltage on the anode of Th_2 disappears: it has then a positive anode voltage from the 100 V supply and it too can be fired by the same pulse through Th_1 and Th_2. This current can be the full load current of the test valve Th_1, which is thus being fired from a high-voltage supply (say 1500 V), but subsequently drawing its current from a low-voltage source (100 V).

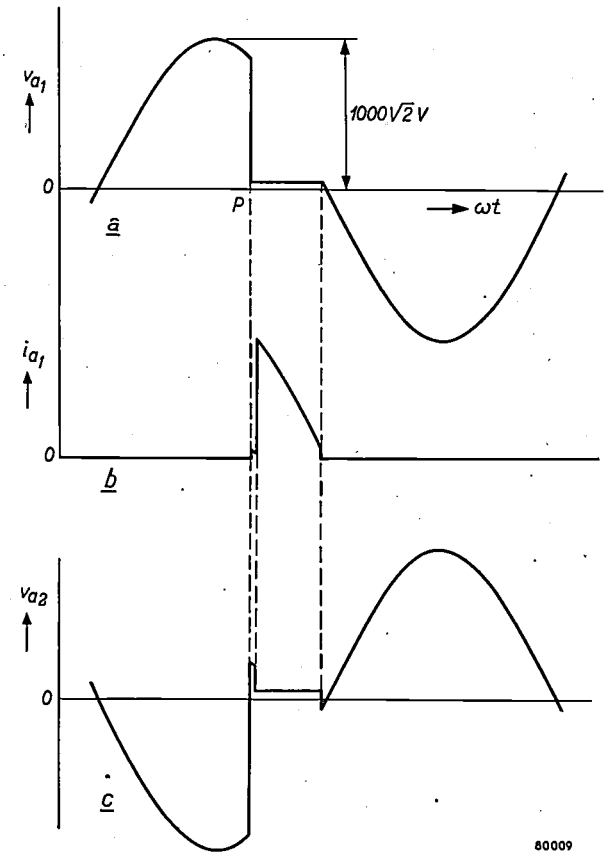


Fig. 5. Explanation of cheater circuit of fig. 4.
a) Voltage v_{a_1} across valve under test, due to source of high tension (here supposed to be 1000 V r.m.s.).
b) Current i_{a_1} through test valve.
c) Voltage v_{a_2} across auxiliary valve.
Immediately before striking of test valve (at P), v_{a_1} is high. Power consumption is low, owing to current being mainly drawn from low voltage source. (Firing of auxiliary valve is shown slightly later than striking of test valve.)

To provide a grid voltage for Th_1, the circuit shown in *fig. 6* is employed, which supplies both the D.C. component and the triggering pulse. The negative D.C. bias, which is variable up to —100 V, is drawn from a full-wave rectifier (not shown). The trigger pulse, which has an amplitude of 300 V and a duration of 50 microseconds, is generated as follows. An A.C. voltage of 50 V is applied to the control grid of a high-slope pentode Pe_1. This A.C. voltage can be phase-shifted by means of a variable resistor R_3. The valve is overloaded by such a voltage and therefore produces a square-wave anode current. This waveform is differentiated (C_4-R_4) and the result applied to the grid of a Mullard 2D21 thyratron (Th_3), which forms a low-impedance output stage. The valve Th_3 is therefore caused to strike and so initiates the discharge of a capacitor C_5 via a choke L_5 and a resistor R_5. The peak-shaped voltage which then appears across L_5-R_5 is fed, via the capacitor C_6, to the grid of the valve under test.

The grid of the auxiliary valve (Th_2, fig. 4) is connected to a permanent D.C. bias in series with a transformer through which is fed the trigger pulse generated by the circuit of fig. 6.

Measurement of pre-striking grid current

Suppose a valve is set up in the above apparatus (fig. 4) and fired by the pulse at some time after the occurrence of the peak anode voltage (firing angle $\alpha_1 > 90°$). The pulse is also fed to the auxiliary valve Th_2, so it too fires and allows full load current to flow through Th_1 and Th_2 in the way previously described. If now the negative grid bias of Th_1 is gradually reduced, the valve first continues to strike at the angle α_1 determined by the grid pulse, until the D.C. bias reaches a value equal to the critical grid voltage corresponding to the H.T. peak anode voltage. In the H.T. circuit the firing of Th_1 is then suddenly advanced from α_1 (>90°) to 90°. The grid voltage $V_{g,90}$ at which this happens can be read on a voltmeter. If the grid resistance is increased and the measurement repeated, a higher value of $-V_{g,90}$ will be noted. From the change in resistance and the change in $V_{g,90}$ the pre-striking grid current can be deduced with the help of equation (1), see above.

Fig. 7 represents the whole operation in terms of voltage and current wave-forms. It also helps to make clear exactly what is being measured and to bring out the distinction between this method and previous ones. Previously, the grid current measured was that at the moment when the valve struck; the valve was at it correct working temperature, but it was not subject to its rated peak anode voltage, and also it was impossible to fire the valve at angles greater than 90°. As pointed out earlier, the maximum peak-to-mean anode current ratio occurs when the valve is fired late in the positive anode voltage half-cycle. It was therefore impossible, using previous methods, to test a valve with both maximum anode voltage and maximum peak

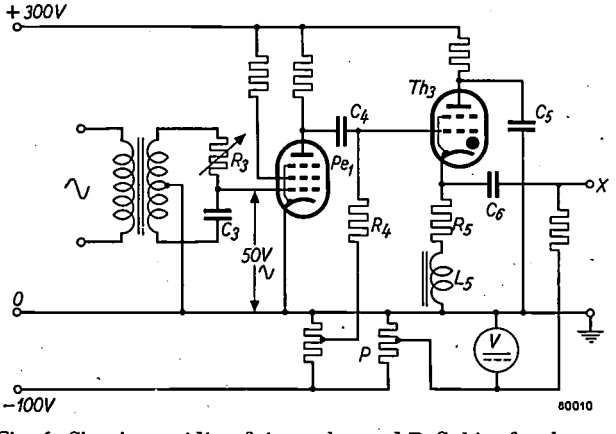


Fig. 6. Circuit providing firing pulses and D.C. bias for cheater circuit (fig. 4). A transformer connected to R_3-C_3 supplies 50 V A.C., with variable phase shift, to control grid of pentode Pe_1 (EF 91). Square-wave anode current of Pe_1 is differentiated by C_4-R_4, producing pulses which fire output stage thyratron Th_3 (Mullard 2 D 21). Th_3 being fired, capacitor C_5 discharges through Th_3, R_5 and L_5 (choke L_5 helping to extinguish Th_3). Via coupling capacitor C_6, voltage pulse across R_5-L_5 is applied to grids of test valve (at X) and of auxiliary valve (via transformer T_4, see fig. 3). Variable D.C. bias is obtained from potentiometer P.

and mean anode currents, and these are the very conditions which lead to maximum grid emission, leakage and uncontrolled cathode emission — i.e. to maximum grid current.

The new method has no such drawbacks. The grid current is necessarily measured when the anode voltage is at its peak value, whatever the peak-to-mean anode current ratio. The value obtained, is therefore a true indication of the worst possible grid current likely to flow under any given conditions.

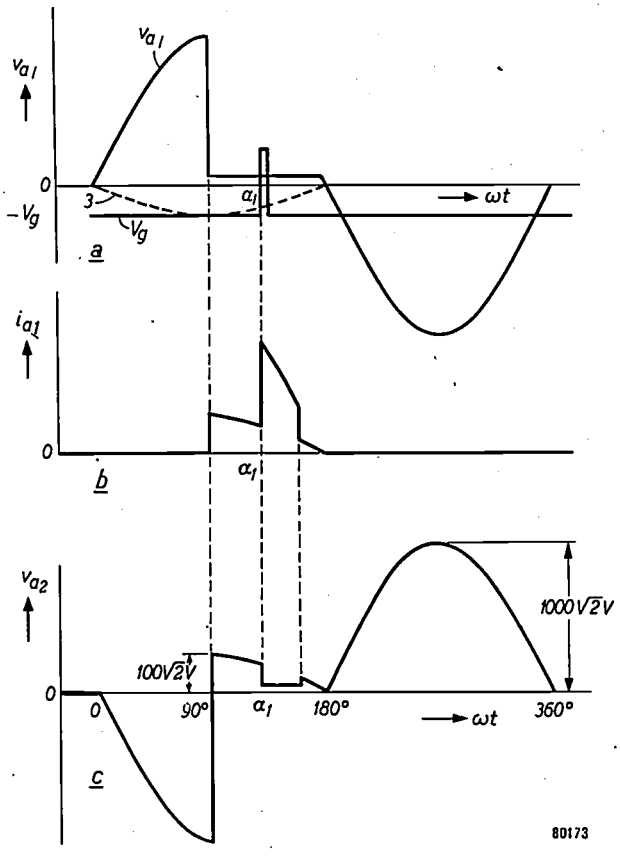


Fig. 7. Explanation of pre-striking grid current measurement with cheater circuit of fig. 4.
At (*a*), D.C. bias of test valve Th_1 is shown adjusted to top of critical grid voltage (curve *3*), causing Th_1 to fire at peak of H.T. anode voltage (see *b*). Auxilliary valve Th_2 is fired by a pulse (firing angle $\alpha_1 > 90°$), initiating main current through both valves. Voltage v_{a2} across Th_2 is shown at (*c*).

It will be appreciated that the time available for de-ionization of the auxiliary valve Th_2 is very short. In fact, if t_1 denotes the instant where Th_2 extinguishes (i.e., where the anode voltage of both valves in series has dropped to, say, 30 V), and t_2 the instant where v_{a2}, in the subsequent half-cycle, reaches the value of, say, 20 V at which ionization becomes possible again, and if the applied A.C. voltages are 100 V and 1000 V r.m.s., then

$$100\sqrt{2}\sin\omega t_1 = 30$$
$$\text{and} \quad (1000-100)\sqrt{2}\sin\omega t_2 = 20$$

(with $\omega = 2\pi \times 50$ radians per second, and the load resistors supposed to be non-inductive). If τ is the interval between t_1, and the moment where the A.C. voltages pass through zero, and τ_2 the interval between this moment and t_2, then we find from the above equations: $\tau_1 = 680$ μsec and $\tau_2 = 50$ μsec. If the high voltage is increased to 10 kV, this merely reduces τ_2 to 4 μsec, leaving τ_1 unaffected. Thus it is clearly possible to use this circuit at very high voltages, provided the low voltage does not exceed 100 V r.m.s. and provided the valve de-ionization time does not exceed 500 μsec. A circuit of this type has been successfully used for testing mercury thyratrons at voltages up to 30 kV peak.

Summing up, this method of measuring grid currents has the following advantages:

1) It measures the maximum grid current flowing under any given conditions of loading for any anode voltage up to, say, 30 kV peak.
2) It economizes in power when testing thyratrons drawing high currents at high voltages.

The power required to operate this circuit for a 6 A thyratron at 100 V, including auxiliaries, amounts to 800 W only. Comparing this with the power required for a straight test at, say, 1000 V, it is evident that the advantages of this circuit are very real, and when voltages of 20 kV or more are required, it is clear that a direct test method is prohibitive, so that this circuit is not merely useful but a necessity.

Summary. Part I of this article deals with the measurement of the pre-striking grid current of thyratrons under full operating conditions. Several conventional circuits are reviewed, and shown to be not entirely satisfactory. One of the drawbacks when large thyratrons are to be tested, is the high power consumption. A "cheater circuit" has therefore been devised, which simulates the desired working conditions with a great economy of power. Here the valve is part of a low-current H.T. circuit, until immediately after firing, when it is automatically changed over to a L.T. circuit, in which the full rated anode current is permitted to flow. This system allows the measurement of the pre-striking grid current under any conditions of loading.

CONDITIONS FOR SQUARE HYSTERESIS LOOPS IN FERRITES

by H. P. J. WIJN, E. W. GORTER, C. J. ESVELDT and P. GELDERMANS.

538.23:621.318.134

Since ferrites were introduced commercially by Philips nearly ten years ago, the applications of these soft magnetic materials have enormously increased. The further development of these materials has been influenced by the fact that it has been found possible to manufacture them with properties specially adapted to particular purposes. The present article deals with the fundamental conditions for obtaining ferrites with rectangular hysteresis loops.

Introduction

Magnetic cores with approximately rectangular hysteresis loops have a wide range of application. They are used for example in the so-called "magnetic memory" [1]) of computing machines and automatic pilots, and for magnetic switching elements.

For a memory element, the requirements are that when a square pulse of a certain height is passed through the magnetizing coil, the core shall revert to its original condition after passage of the pulse but when a pulse of *double* this height is passed through the coil, the magnetization of the core shall be *reversed* after passage of the pulse. When the material is used for switching elements, the magnetization must not be affected by a positive pulse, but must be reversed by a negative pulse of the same height.

The pulses employed in these techniques are usually of very short duration, so that, during the pulse, the variation in the current, di/dt, assumes very high values, as a result of which rapid variations dB/dt occur in the induction, and eddy currents are produced.

It is important that the magnitude of these eddy currents should be minimized. In ferromagnetic metals (nickel-iron alloys such as "Hypernik" and "Deltamax" [2])) this is to a limited extent achieved by building up the core from very thin insulated laminations of the material; in practice, however, it is difficult to construct such laminated cores to give nearly rectangular hysteresis loops.

Magnetically soft materials such as ferrites [3]), which are at the same time poor conductors, offer considerable advantages over the use of laminated ferromagnetic metals in pulse applications (given that these materials can be made with rectangular hysteresis loops [4])).

Definitions of certain quantities

One or two concepts will now be introduced which are essential to a consideration of the problem [5]).

The loop depicted in *fig. 1* (full line) is the hysteresis loop of a ferrite. It represents the magnetization J plotted against the magnetic field H for values

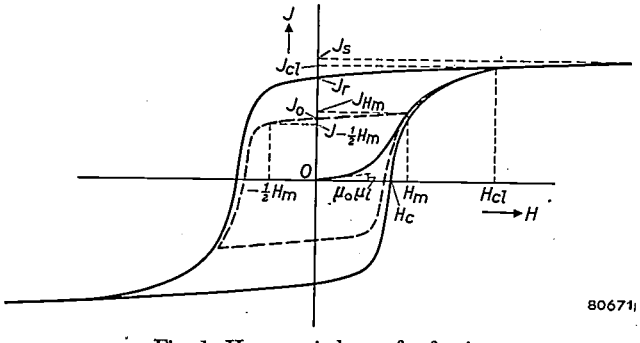


Fig. 1. Hysteresis loop of a ferrite.

of H which decrease from large positive to large negative values, and then revert to the positive value. The value of the field at which the magnetization is zero is known as the coercive force H_c of the material [6]). On decreasing the field from a large value, the point at which the magnetization

[1]) J. A. Rajchman, RCA rev., **13**, 183-201, 1953. A. Wang, J. appl. Phys. **21**, 49-54, 1950. Some specific instances are reviewed in F. van Tongerloo, T. Nederlands Radiogenootschap **18**, 265-285, 1953, No. 11.

[2]) See for example R. M. Bozorth, Ferromagnetism, Van Nostrand, New York, 1952.

[3]) J. L. Snoek, Philips tech. Rev. **8**, 353, 1946; J. J. Went and E. W. Gorter, The magnetic and electrical properties of Ferroxcube materials, Philips tech. Rev. **13**, 181-193, 1951/1952, referred to hereafter as I. As in I, the present article uses the rationalized Giorgi (M.K.S.) system, in which B is measured in Wb/m^2 and H in A/m ($\mu_0 H$ in Wb/m^2). $B = 10^{-4}$ Wb/m^2 corresponds to $B = 1$ gauss, and $\mu_0 H = 10^{-4}$ Wb/m^2 ($H = 79.5$ A/m) to $H = 1$ oersted. $\mu_0 = 4\pi \times 10^{-7}$ volt seconds/ampere metres. The formula $B = \mu_0 H + J$ takes the place of $B = H + 4\pi I$ in the electromagnetic c.g.s. system. If $J = 1$ Wb/m^2, $4\pi I = 10{,}000$ gauss and $I \approx 800$ gauss.

[4]) E. Albers-Schoenberg and D. R. Brown, Electronics **26**, April 1953, p. 146.

[5]) For a more detailed discussion of these concepts and their physical basis see J. J. Went, Philips tech. Rev. **10**, 246-254, 1948/1949.

[6]) In magnetic materials a distinction is made between a coercive force for the induction, ${}_B H_c$, i.e. the field strength at which the induction B is zero, and a coercive force for the magnetization, ${}_J H_c$, the field strength at which the magnetization J is zero. In general, ${}_B H_c$ and ${}_J H_c$ differ because $B = \mu_0 H + J$, and B and J do not become zero at the same time. In the materials discussed here the values of H are so small that the difference may be disregarded and the single symbol (H_c) employed.

begins to diverge from its previous values is denoted H_{cl}; the corresponding magnetization is denoted J_{cl}. For fields $H > H_{cl}$ the variations in the magnetization are reversible. Smaller hysteresis loops are also obtainable, as shown for example by the dotted line in fig. 1, with extreme field values of $+H_m$ and $-H_m$ ($H_m < H_{cl}$). As a measure of the effectiveness of ferrites for the cores of memory devices the concept of "squareness" is introduced, defined as:

$$R_s = \frac{B_{-\frac{1}{2}H_m}}{B_{H_m}}$$

or, what is practically equivalent,

$$R_s = \frac{J_{-\frac{1}{2}H_m}}{J_{H_m}} \quad \ldots\ldots\ldots \quad (1)$$

The demoninator and numerator of the latter represent respectively the magnetization for a field $+H_m$ and that for a field $-\frac{1}{2}H_m$. It will be clear that R_s is also a function of the maximum field H_m determining the size of the loop. When R_s is measured as a function of H_m it is found that a maximum occurs for a certain value of H_m; let this maximum be denoted by $(R_s)_{max}$. With ferrites for which $(R_s)_{max} > 0.7$, this value of H_m is roughly equal to H_c.

When ferrites are employed as switching elements the ratio J_0/J_{H_m} is important; J_{H_m} represents the above mentioned magnetization for a field H_m, and J_0 the remanent magnetization after removal of the field (see fig. 1). J_0/J_{H_m} is a function of H_m, and the maximum, $(J_0/J_{H_m})_{max}$, of this ratio is also of interest.

General considerations

Magnitude of the remanent magnetization

Consider as the starting point the demagnetized condition, that is, point O in fig. 1. It may be taken as generally known that in this condition the material is divided up into "Weiss domains" within which the material is uniformly magnetized (see [5]). The magnetization vector in each of these Weiss domains lies in a certain direction (the preferential direction) and the magnetization averaged over all the domains is zero. The preferential direction of magnetization in each domain is determined by three factors, namely the crystal anisotropy, the stress anisotropy and the shape anisotropy. These factors will be discussed presently. Very small external magnetic fields turn the magnetization vectors away from their preferential orientation, towards the direction of the applied field. The extent to which this is possible in the case of sintered ferrites is represented approximately by the quantity μ_i, the initial permeability [7]) (see fig. 1).

When the material is magnetized by a very strong field, all the magnetization vectors are parallel to each other and there is no longer any division into Weiss domains. This corresponds to the saturated state with magnetization J_s (see fig. 1). If the field be now gradually reduced to zero, the magnetization vectors turn from the direction of the field towards the nearest preferential directions as determined by the anisotropies mentioned above.

When the field H is zero a magnetization J_r remains (the remanence). For an ideal rectangular loop $J_r/J_s = 1$. If H be varied slightly, starting from the remanent state, a permeability μ_{rem} will be found which is again determined by the rotation of the magnetization vectors. The magnitude of the ratios J_r/J_s and μ_{rem}/μ_i, corresponding to the three kinds of anisotropy, will now be evaluated. The significance of the second of these ratios will appear later.

a) Crystal anisotropy. In nearly all cubic ferrites it has so far been found that the body diagonal is the preferential direction of the magnetization; there are accordingly eight such preferential directions. The magnetization energy per unit volume of a material whose magnetization vector is defined by the direction-cosines (with repect to the cubic axes) α_1, α_2 and α_3, is given to a first approximation by:

$$E = K(\alpha_1^2\alpha_2^2 + \alpha_1^2\alpha_3^2 + \alpha_2^2\alpha_3^2) \quad \ldots \quad (2)$$

In this expression, K may be both positive and negative. In materials with positive K, the value of E is at a minimum when one α is equal to unity and the others are zero, that is, when the magnetization vector is parallel to one of the sides of the cube.

In most ferrites, however, and also in some metals such as nickel, K is negative. E is then at a minimum when the factor between brackets in (2) is at a maximum, viz. when $\alpha_1 = \alpha_2 = \alpha_3$. The magnetization vector is then parallel to one of the body diagonals of the cube. If the crystal anisotropy is the only anisotropy present, when the field is reduced to zero after saturation in a strong field, the magnetization vectors of the Weiss domains in polycrystalline materials turn back from the direction of the field to the nearest cube diagonal (a (111)-direction).

---

[7]) In contrast with article I, μ_i is taken as the relative permeability. The corresponding absolute permeability is $\mu_0\mu_i$, for which the symbol μ_i was used in I. The same applies to the quantity μ_{rem} introduced later.

Since there are eight preferential directions, the vectors revert to directions which are all contained within the solid angle $\pi/2$ (*fig. 2*). According to calculations by Gans [8]), this leads to $J_r/J_s = 0.87$ and $\mu_{rem}/\mu_i = 0.36$.

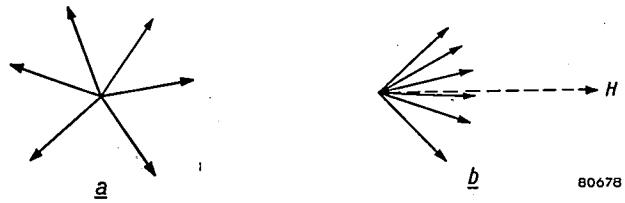


Fig. 2. *a*) Distribution of the magnetization vectors in the Weiss domains in demagnetized polycrystalline materials. *b*) As above for the remanent state when (negative) crystal anisotropy predominates.

b) *Stress anistropy*. It is well-known that the length of a rod of magnetic material changes with its magnetization; this property is known as magnetostriction. A distinction is made between positive and negative magnetostriction according to whether the magnetization is accompanied by an expansion or a contraction in the direction of magnetization.

If strains are present in the material as a result of elastic deformation, the magnetization tends to be so oriented that the accompanying variations in length oppose these strains. If the stress anisotropy predominates there will be only two preferential directions for the magnetization at every point in the material. In the case of negative magnetostriction (as usually found in ferrites), these directions correspond at each point to the two orientations at which the magnetization vectors are parallel to the greatest compressive strain or the smallest tensile strain. Here, too, the magnetization vectors turn back to the nearest preferential direction when the field is reduced to zero from the saturation value. For a random distribution of the strains in the material the vectors revert to preferential directions distributed over a solid angle of 2π (see *fig. 3b*). Under these conditions it has been calculated that $J_r/J_s = 0.5$.

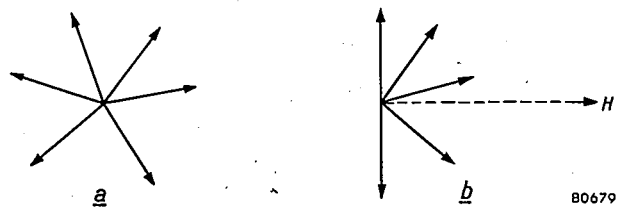


Fig. 3. *a*) As fig. 2a.
b) Distribution of the magnetization vectors for the remanent state when stress anisotropy prodominates.

[8]) R. Gans, Ann. Physik **15**, 28, 1932.

For ferrites with small crystal anisotropy but with a large magnetostriction coefficient, it is possible, when sufficiently large external forces are applied, that the latter determine the preferential direction of the magnetization. This has been demonstrated in the Eindhoven Laboratory by G. W. Rathenau and G. W. van Oosterhout in the following manner. A ring of glass was melted onto the outer cylindrical surface of a ring of Ferroxcube [9]). The coefficient of expansion of the glass was slightly higher than that of the Ferroxcube, so that after cooling to room temperature the Ferroxcube was subjected to tangential compression. The magnetization caused by the negative magnetostriction was therefore parallel to the predominant strain. Consequently the strains in the ferrite were no longer distributed in a random manner, but mainly in one direction only, so that at every point in the material only two orientations of the magnetization were possible, viz. parallel and antiparallel to the strain. The remanence after removal of a field parallel to the compression (i.e. a tangential field) can thus theoretically assume a value equal to the saturation magnetization $J_r/J_s = 1$. *Fig. 4* shows a family of hysteresis loops plotted

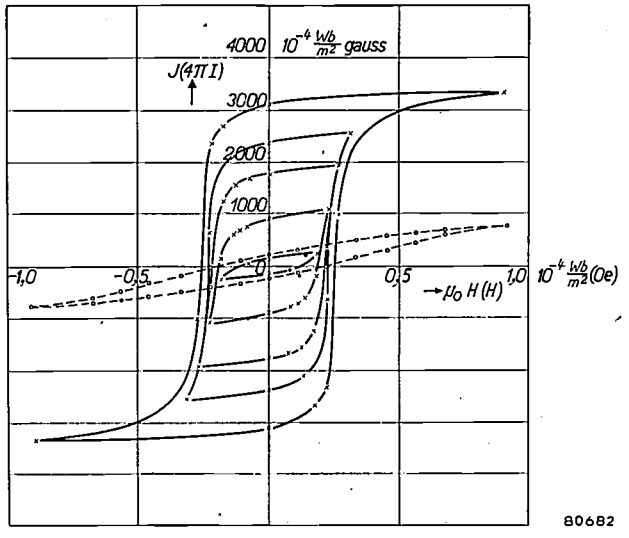


Fig. 4. Family of hysteresis loops for a ring of Ferroxcube IVA enclosed by a band of special glass. Dashed curve: hysteresis oop for Ferroxcube IVA without glass (see text).

from measurements on a ring of Ferroxcube IVA enclosed in glass. The loop drawn in broken lines is that of the ring without the glass. A great advantage of this type of material, with its rectangular hysteresis loop, is the low coercive force.

[9]) Netherlands Patent application No. 175120, dated Jan. 1953.

Similar results have been obtained by enclosing a ferrite ring in synthetic resin [10]).

If a polycrystalline material, prepared in the demagnetized state (*O*, fig. 1) by cooling from a temperature above the Curie point, contains no strain the direction of the magnetization vector in each Weiss domain is determined by the crystal anisotropy (disregarding shape anisotropy, see *c*). The magnetization vectors of the Weiss domains are then oriented in the above-mentioned preferential directions. When the material is magnetized and then brought into the remanent state, a large number of the individual magnetization vectors are turned from their original preferential direction into another. In general, this will be accompanied by a change of shape of individual crystals. For example, if the body diagonals are the preferential directions in cubic ferrites, the length of the body diagonal will depend on whether the magnetization vector is parallel to this diagonal or not. The corresponding difference in length is called the "magnetostriction in the preferential direction", λ_{111}. Even if there is no strain present in a polycrystalline material in the demagnetized state, there may nevertheless be strains present in the remanent state, unless $\lambda_{111} = 0$. In this particular case, the magnitude of the remanence is determined by the crystal anisotropy only: this is a requirement for high remanence [11]).

The results of measurements on a series of mixed crystals of nickel ferrite and ferrous ferrite shows the importance of magnetostriction in the preferential direction. *Fig. 5* shows the magnetostriction λ plotted against the magnetic field H for polycrystalline preparations of mixed crystals of Fe_3O_4 and $NiFe_2O_4$. The saturation magnetostriction (the value of λ for effective saturation) of these materials is $+ 41.6\times10^{-6}$ and $- 22\times10^{-6}$ respectively, so that it may be expected to

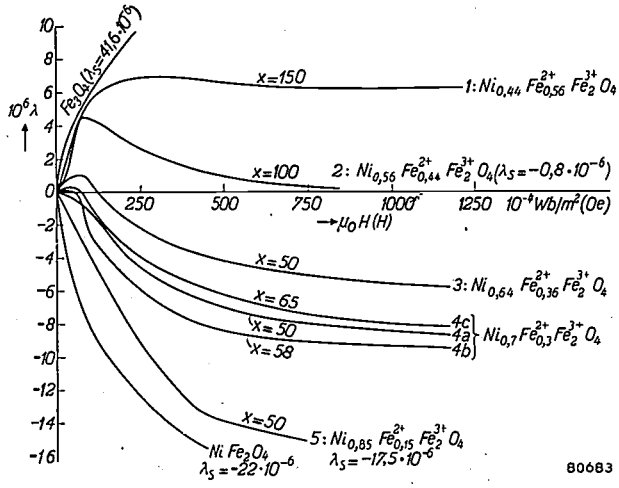


Fig. 5. Magnetostriction as a function of the magnetic field in polycrystalline samples of mixed crystals of nickel ferrite-ferrous ferrite. All samples were fired at 1350°C, but in gas currents of different composition, viz. 600 ml/min CO_2 + x ml/min of a mixture of 90% N_2 and 10% H_2. The value of x and the chemical composition of the ferrites are shown in respect of each curve.

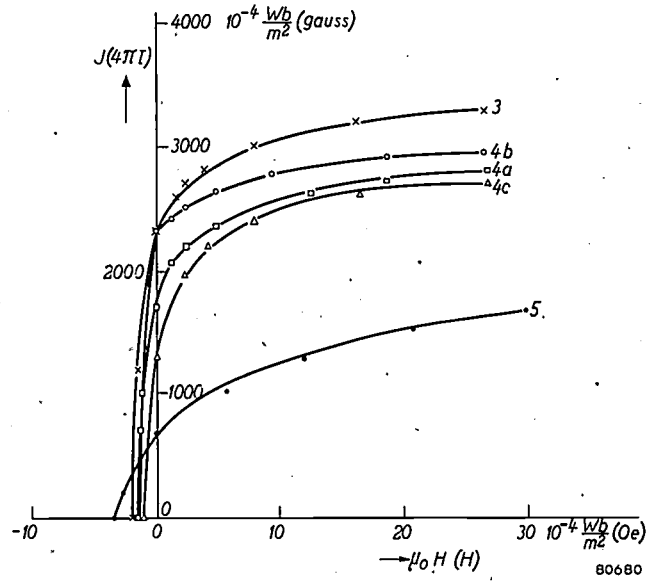


Fig. 6. Upper branch of the hysteresis loop for decreasing fields, for some of the ferrites in fig. 5.

be zero for a mixed crystal with a certain ratio of these constituents. It is, in fact, seen from fig. 5 that the saturation magnetostriction λ_s of $Ni_{0.56}Fe^{2+}_{0.44}Fe^{3+}_2O_4$ is very small ($\lambda_s = -0.8\times10^{-6}$). It was shown earlier, however, that it is the magnetostriction in the *preferred direction* that must be small in order to obtain a rectangular hysteresis loop. It can be shown quite simply that the sign of the magnetostriction of a specimen in a field of the same order of magnitude as the coercive force (i.e. $\mu_0 H_c \approx 10^{-4}$Wb/m², $H_c \approx 1$ oersted, for mixed nickel-ferrous-ferrites) is the same as the sign of the magnetostriction in the preferred crystallographic direction. Fig. 5 shows that the sign of the preferential magnetostriction is reversed in compositions occurring between those of samples *4a* and *4b* ($Ni_{0.7}Fe^{2+}_{0.3}Fe^{3+}_2O_4$ fired at 1350 °C in a current of gas consisting of 600 ml carbon dioxide and x ml nitrogen with 10% hydrogen, per minute [12])).

Part of the hysterises loops of a number of the ferrites referred to in fig. 5 are shown in *fig. 6*. It

[10]) H. J. Williams, R. C. Sherwood, Matilda Goertz and F. J. Schnetler, Commun. Electr. 9, 531, 1953.

[11]) For these considerations we are indebted to J. Smit of the Eindhoven Laboratory.

[12]) The difference in the reducing power of these gases results in a slight displacement in the ratio of Ni^{2+} to Fe^{2+} in the ferrite. The quantity of NiO or FeO that may occur as second phase can be disregarded, in comparison with the volume of the pores.

can be seen that samples *4a* and *4b* do actually yield the anticipated greater squareness compared with the other materials.

c. Shape anisotropy. It is well known that when an open magnetic circuit is magnetized an opposing field is produced, i.e. partial demagnetization occurs, due to the "free poles" at the extremities. This opposing field, which may be regarded as due to shape anisotropy, is proportional to the magnetization. Although this article is concerned only with magnetic circuits which can be considered macroscopically as closed, we are nevertheless to a large extent concerned with demagnetization due to shape anisotropy in view of the more or less porous structure of the ferrite (see fig. 10). The porosity of the sintered material varies between 1% and about 25%, and the effect of the air inclusions is such that the field H_{int} in the material is weaker than the applied field H_{ext}. In general it can be said that this difference is proportional to the magnetization of the material:

$$\mu_0(H_{ext} - H_{int}) = NJ \quad . \quad . \quad . \quad . \quad (3)$$

where N is a constant depending on the porosity.

In consequence of this, the measured hysteresis loop differs from that which would be obtained if the material were free from cavities or pores (see *fig.* 7). This effect can be regarded as a "shearing" of the

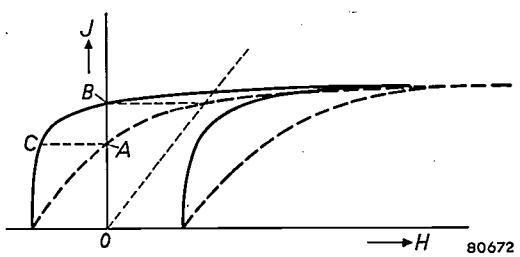


Fig. 7. Effect of "shearing" in the apparent shape of the hysteresis loop.

hysteresis loop. It will be seen from the figure that, owing to the shearing, the ratio J_r/J_s is considerably reduced, since J_r decreases whilst J_s of course remains constant. It may be noted that in the absence of shearing, the state corresponding to the point B is typified by a distribution of the magnetization vectors as in fig. 2*b* or 3*b*, according to which kind of anisotropy predominates. When shearing occurs, the point A is only apparently the remanent state; in fact the material is in a demagnetizing field $\mu_0 H = -NJ$, so that the actual state of the material is as shown at C. The distribution of the vectors is now very different from that in figs. 2*b* and 3*b*, and more resembles that of the demagnetized condition (figs. 2*a* and 3*a*). It is also to be expected that μ_{rem}/μ_i will now be more nearly equal to unity, even in the case of predominant crystal anisotropy. In ferrites with equal saturation magnetization J_s and with similar porosity (i.e. equal N), the smaller the coercive force of the ferrite, the greater the influence of the demagnetizing field on the hysteresis loop.

Of possibly greater importance than the "shearing" is the effect of the porosity on the preferential direction of the magnetization at every point in the material. It is to be expected that in porous materials the direction of magnetization at each point will be largely determined by the direction of the demagnetizing field at the point. These demagnetizing fields result in only one preferred direction at each point, so that a low value of J_r/J_s may be anticipated.

To obtain "rectangular" hysteresis loops, the ratio J_r/J_s should be as nearly as possible equal to unity With polycrystalline materials this can best be approximated when the crystal anisotropy predominates over the other anisotropies. It is especially important to ensure that shape anisotropy is absent: this means that porosity must be as low as possible. It is then found that $J_r/J_s = 0.87$. Higher values could be obtained if it were possible to ensure that the crystals are so oriented that all their body diagonals are parallel; this could even yield $J_r/J_s = 1$. This possibility cannot be pursued further here.

The coercive force

From the point of view of the application of ferrites with rectangular hysteresis loops it is important that the coercive force H_c shall be as small as possible. This is because H_c determines the number of ampere-turns necessary to reverse the direction of the remanent magnetization in the core. The coercive force is related to the field strength needed to displace the boundaries between the Weiss domains i.e. the Bloch walls. These "walls" are fixed at certain locations in the material as a result of internal strains and non-magnetic inclusions. It is to be expected that such walls in ferrites will be fixed to air pores which usually occur in a far greater number than in metals. We may therefore, for a moment, consider Néel's theory [13]), which gives an insight into the effect of porosity on the magnitude of H_c. Néel points out the fact that the internal stray magnetic fields caused by inclusions in ferromagnetic materials

[13]) L. Néel, Ann. Univ. Grenoble, **22**, 299-343, 1946, and Physica **15**, 225-234, 1949.

are limited to smaller domains when the Bloch walls pass *through* the inclusions. This is illustrated in *figs. 8a* and *b*. The magnetic "charges" of opposite sign are closer to each other in *b* than in *a*, which

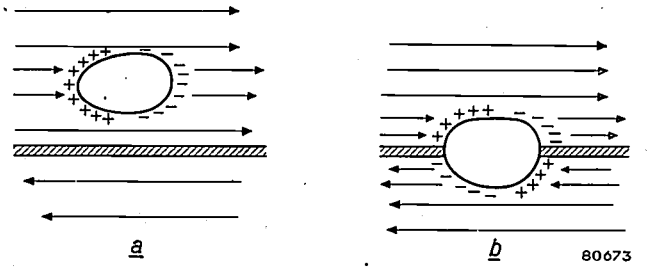


Fig. 8. Effect of inclusions or pores on the energy in a Bloch wall (Néel [13])
a) wall not intersecting the cavity.
b) wall intersecting the cavity.

means that the total magnetic energy is considerably less in *b* than in *a*. The location of the Bloch walls at which minimum free energy occurs is therefore that at which the wall intersects as many air pores as possible. The coercive force is then the field strength necessary to dissociate a wall from its air pores, and this coercive force will be related to the concentration, as well as to the size, of the air pores. It has been found that inclusions or pores of diameters comparable to the Bloch wall thickness exercise the greatest influence on H_c. According to Néel's theory, the coercive force H_c in the case of negative crystal energy (K negative), when the absolute value of K is high compared with J_s^2/μ_0 (which applies to the ferrites considered here), is given by the formula:

$$H_c = \frac{4|K|p}{3\pi J_s}\left[0.39 + \tfrac{1}{2}\ln\frac{3J_s^2}{4\mu_0|K|}\right], \quad \ldots \quad (4)$$

where p represents the porosity of the material [14]).

The relationship between p and H_c in the range of Ni-Zn ferrites of differing properties has been investigated experimentally (see *Table I)*, and the value of H_c, as computed from formula (4), is shown in the last but one column, $|K|$ being calculated from μ_i on the assumption that μ_i is determined only by rotation (i.e. $\mu_i - 1 = \frac{1}{2}J_s^2/\mu_0|K|$). It is seen that in the range of porosities considered, H_c does depend on the porosity.

If a low value of H_c is to be attained, the ferrite must be well sintered during preparation to obtain as low a porosity as possible. This can be achieved by firing at a high temperature, by employing a ferrite having a relatively low melting point, or by adopting a suitable ceramic technique. In the last two instances a high temperature is avoided, thus minimizing chemical reduction of the ferrite with its adverse consequences on the magnetic and electrical properties.

Table I. Relation between porosity and coercive force of nickel-zinc ferrites.

Ferroxcube IV Type	Firing temp. °C	Chemical composition in mole % (remainder $FeO + Fe_2O_3$) NiO	ZnO	Porosity p %	Saturation magnetization J_s 10^{-4} Wb/m²	I_s gauss	Initial permeability μ_i	Coercive force $\mu_0 H_c$: 10^{-4} Wb/m², H_c : oersted per Eq. (4)	measured
A	1250	17.5	33.2	8.9	3650	292	650	0.4	0.4
B	1250	24.9	24.9	15.4	4150	332	230	2.0	1.4
C	1250	31.7	16.5	22.5	4012	321	90	6.2	4.0
D	1250	39.0	9.4	24.3	3537	283	45	10.4	6.8
E	1250	48.2	0.7	24.8	2450	196	17	16.1	13.7
A	1450	17.5	33.2	9.5	3620	290	470	0.6	0.3
B	1450	24.9	24.9	3.2	4750	380	312	0.4	0.4
C	1450	31.7	16.5	8.0	4760	381	86	2.7	1.1
D	1450	39.0	9.4	8.9	4260	341	63	3.5	1.7
E	1450	48.2	0.7	9.9	1688	135	42	3.7	3.2

High remanence and small coercive force

Let us now briefly summarise the conditions to be fulfilled in order to obtain materials with square hysteresis loops (i.e. with a high remanence, which implies a high value of J_r/J_s), and suitable for practical purposes (i.e. having a small coercive force).

1) We have seen that if J_r/J_s is to be high, the crystal anisotropy must be predominant. A high value of $|K|$ implies a low value of μ_i, if μ_i originates only from rotational processes. The aim, then, is to produce a ferrite of fairly low initial permeability.

[14]) Thus $(1-p)$ is the ratio of the macroscopic (or apparent) density of the material to the microscopic (or true) density. The latter can be determined by X-ray diffraction methods, the former by the ordinary methods of density measurement.

At the same time its follows from formula (4) that the coecive force H_c increases with $|K|$, so that the crystal energy must not be so great that H_c assumes undesirably high values.

2) The porosity should be as small as possible; demagnetizing fields are then small and H_c will be smallest.

3) A small value of the magnetostriction in the preferred direction is favourable for squareness. Square-loop ferrites suitable for practical purposes, however, can be obtained only if the two first conditions are fulfilled. In the mixed-crystal systems referred to in figs. 5 and 6 this is not the case.

The frequency characteristics of the ferrite are also important. It is clear from the article I that a low value of μ_i will be accompanied by a high ferromagnetic resonance frequency: this is a primary requirement if the hysteresis loop is to be rectangular with pulses of about 1 μ sec. It is also seen from I that it is just with those ferrites which have the smallest initial permeability that the irreversible Bloch-wall displacements are able to follow the current variations up to the highest frequencies. If difficulties due to eddy currents are to be avoided, moreover, the resistivity of the material must be sufficiently high. It appears to be not difficult to attain values higher than 10^2 or even 10^4 ohm metres (M.K.S. system), which means a factor of 10^9 to 10^{11} higher than that of metallic soft magnetic materials.

Examples of ferrites with rectangular hysteresis loops

Measurements of the hysteresis loop have been carried out with a ballistic galvanometer. The dependence of the results on the frequency will not be discussed here. Before considering the square loop materials, a brief review of some well-known materials will be given.

The porosity of Ferroxcube IIIB is about 10%. The coercive force and the crystal anisotropy are both low, and it can therefore be expected that the porosity will result in a considerably reduced value of J_r/J_*[15]); in fact, an average value of 0.27 is obtained. Accordingly $\mu_{rem}/\mu_i = 0.96$. The squareness ratio is of course very small: $(R_s)_{max} \approx 0$.

Ferroxcube IVE has a higher coercive force ($\mu_0 H_c = 14 \times 10^{-4}$ Wb/m^2) and a larger crystal anisotropy (μ_i is only 17), both factors being favourable for a rectangular hysteresis loop. If this material is fired at the normal temperature, however, it is very porous ($p = 25\%$), and the demagnetizing fields again become significant. It is found that $J_r/J_* < 0.6$; $\mu_{rem}/\mu_i = 0.63$, and $(R_s)_{max} = -0.15$. These values show an improvement when the material is fired at a higher temperature; the porosity is then only 10%. In this case, notwithstanding a decrease in coercive force (down to $\mu_0 H_c = 3 \times \times 10^{-4}$ Wb/m^2), the loop is more rectangular, viz. $J_r/J_* = 0.6$, $\mu_{rem}/\mu_i = 0.55$, and $(R_s)_{max} = 0.70$.

The chemical compositions and properties of a number of ferrites giving rectangular hysteresis loops, together with those of some other ferrites, are given in *Table II*. The relationship in sintered ferrites between μ_{rem}/μ_i and J_r/J_* can be seen from the table and from *fig. 9*. For low values of J_r/J_*, μ_{rem}/μ_i approximates to 1. For the maximum value that can be anticipated, viz. $J_r/J_* = 0.87$, the ratio μ_{rem}/μ_i should approach a value of 0.36, and this is roughly the case as shown in fig. 9. Striking results were obtained from the following ferrites.

1) $Co_{0.02}Mn_{0.48}Fe_2O_4$. The hysteresis loop of this ferrite is not particularly square, but the material has that advantage of a small coercive force: $\mu_0 H_c = \times 10^{-4}$ Wb/m^2.

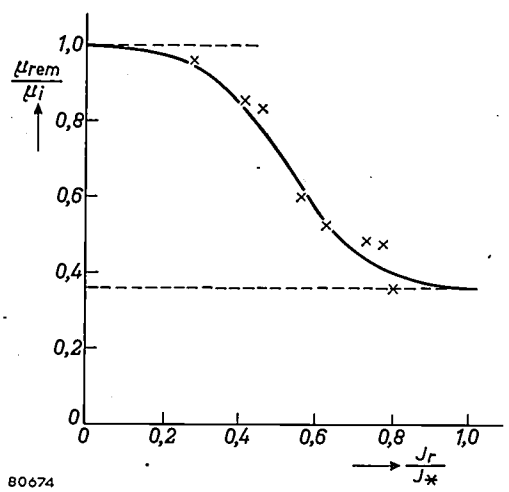


Fig. 9. Relationship between the quantities μ_{rem}/μ_i and J_r/J_*.

2) $CuO_{0.1}(MnO_{1+\delta})_{1.1}Fe_2O_3$ is remarkable for its low coercive force: $\mu_0 H_c = 0.67 \times 10^{-4}$ Wb/m^2.

3) $MgO_{0.5}(MnO_{1+\delta})_{0.875}Fe_2O_3$ is a very close-grained ferrite. Some idea of the porosity can be obtained from the photomicrograph of the polished surface shown in *fig. 10a*. For comparison fig. 10*b* shows a similar photograph of the porous Ferroxcube IIIA (p=9%). From the table it is seen that $(R_s)_{max} = 0.81$, whilst $(J_0/J_{H_m})_{max} = 0.96$.

4) $Mn_{0.1}Ni_{0.5}Mg_{0.4}Fe_2O_4$. Porosity 5%. This ferrite has remarkably good characteristics, viz.

[15]) By J_* is meant the magnetization of a ring measured at $\mu_0 H = 0.01$ Wb/m^2 (H = oersted). Since H_c is very much smaller, an adequate approximation to J_s is obtained by replacing it by the slightly smaller value J_*.

Table II. Properties of ferrites with rectangular hysteresis loops.

No.	Chemical composition	Porosity p %	μ_i	μ_{rem}/μ_i	J_r/J_*	$(J_0/J_{Hm})_{max}$	$(R_s)_{max}$	$\mu_0 H_c$: 10^{-4} Wb/m² H_c: oersted	$\mu_0 H_m$ for $(R_s)_{max}$ 10^{-4} Wb/m²
	Ferroxcube IIIB	10	1230	0.96	0.27	0.32	~0	0.8	
	Ferroxcube IVE,								
	fired at 1250 °C	25	17	0.63	<0.6	0.70	−0.15	14	
	„ „ 1450 °C	10	42	0.55	0.6		0.70	3	
1	$Co_{0.02}Mn_{0.48}Fe_2O_4$	6	83	0.86	0.41	0.83	0.59	0.43	0.48
2	$(CuO)_{0.1}.(MnO_{1+\delta})_{1.1}.Fe_2O_3$	3	86	0.56	0.60	0.93	0.76	0.67	0.85
3	$(MgO)_{0.5}.(MnO_{1+\delta})_{0.875}.Fe_2O_3$		55	0.49	0.73	0.96	0.81		1.35
4	$Mn_{0.1}Ni_{0.5}Mg_{0.4}Fe_2O_4$	5	138	0.48	0.76	0.95	0.83		1.7
5	$Mg_{0.4}Ni_{0.6}Fe_2O_4$		28	0.53	0.62	0.94	0.84		2.6
6	$Li_{0.46}Ni_{0.08}Fe_{2.40}O_4$	5	40	0.36	0.8		0.78		4.3
7	$Li_{0.25}Cu_{0.5}Fe_{2.25}O_4$	4					0.75		

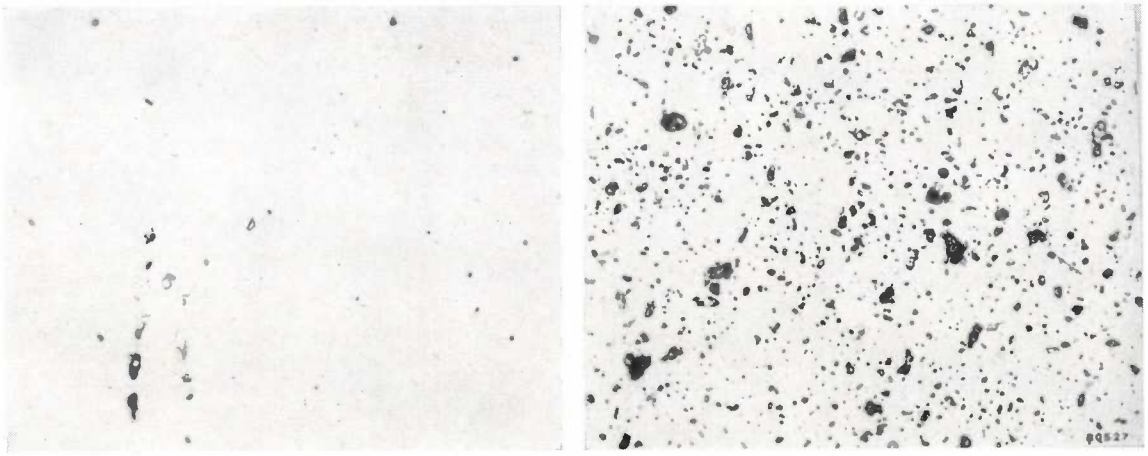
$(J_0/J_{Hm})_{max} = 0.95$ and $(R_s)_{max} = 0.83$ with a field $\mu_0 H_m = 1.7 \times 10^{-4}$ Wb/m².

5) $Mg_{0.4}Ni_{0.6}Fe_2O_4$. The oustanding feature of this ferrite is that the squareness ratio is only very slightly dependent on the temperature. We shall return to this point later. It is found among the mixed crystals of Mg ferrite with Ni ferrite, that firing at 1450 °C does not always ensure low porosity; now and then a rectangular loop is obtained with $p = 20$ %. Microscopic examination of the polished surface gives the explanation: the high porosity is in this case due to much larger pores than in other ferrites. The initial permeability is low ($\mu_i = 28$).

6) $Li_{0.46}Ni_{0.08}Fe_{2.40}O_4$. It appears that "lithium ferrite" ($Li_{0.5}Fe_{2.5}O_4$) fired at 1000 °C in oxygen exhibits slightly negative magnetostriction. If it is chemically reduced by firing at a somewhat higher temperature (1150°C) so that a mixed crystal of lithium ferrite and ferrous ferrite is produced, a positive magnetostriction in the preferred direction is obtained. Intermediate firing temperatures, give the greatest squareness but it is still insufficient for practical purposes because the relatively low temperature results in too much porosity. Firing at higher temperatures gives a lower porosity, but an increased ferrous ferrite content, and hence not a low magnetostriction. An improvement was obtained by starting with a mixed crystal of lithium and nickel ferrite. The apparent density of this material is 4.60, and the actual density of the material itself is 4.85. This, combined with a high value of $|K|$ (since $\mu_i = 40$) promotes squareness of the loop (see Table II).

7) $Li_{0.25}Cu_{0.5}Fe_{2.25}O_4$. Porosity 4%, $(R_s)_{max} = 0.75$.

As explained above, the squareness ratio R_s of a hysteresis loop is a function of the maximum field strength H_m at which the loop is measured. *Fig. 11* shows R_s as a function of H_m for the ferrites listed in Table II. It is seen that the lower the field strength



a *b*

Fig. 10. Photo-micrographs of polished surfaces (magn. 400 ×). *a*) Ferrite No. 3 (Table II). *b*) Ferroxcube IVA. $p = 0.09$.

at which R_s reaches a maximum, the more R_s depends on H_m. The figure also shows the ratio J_0/J_{H_m} plotted against H_m. It may be noted that the optimum field for R_s is practically the same as for J_0/J_{H_m}.

Special properties

In many applications of square hysteresis loop ferrites a low temperature coefficient and high stability of the constants $(R_s)_{max}$ and $(J_0/J_{H_m})_{max}$ are also required. These factors will now be briefly reviewed.

Effect of temperature on $(R_s)_{max}$

Different meanings can be attached to the dependence of $(R_s)_{max}$ upon the temperature. We choose a temperature range of 20 to 60°C. in which R_s is determined in respect of values of H_m such that R_s at 40°C is equal to $(R_s)_{max}$. The values obtained for a number of ferrites are plotted in *fig. 12* against the temperature. At any temperature (except 40°C) a loop can be obtained which is more rectangular, corresponding to a different value of H_m, but it is found that the spread in $(R_s)_{max}$ within a certain range of temperatures is smallest for those loops which refer to the optimum field H_m for the average temperature (40°C) in the working range.

Stability of the hysteresis loop

Fig. 13 shows an ideal rectangular hysteresis loop. It will be clear that if this loop be followed round a number of times to the point where $H = H_m$, this will be the point I in the diagram. When H is made zero, point II will be reached. A current pulse which causes H to drop to $-\frac{1}{2}H_m$ brings us to point III and, if H then again becomes zero, the material will return to the state represented by point II. In the applications for which this material is employed the cycle II-III-II may

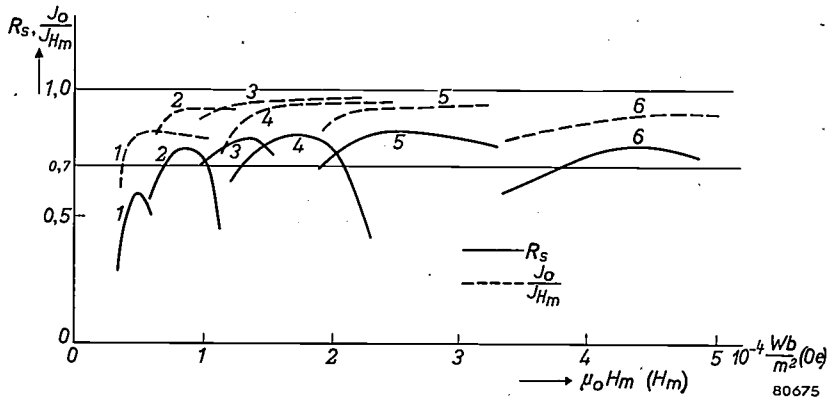


Fig. 11. R_s and J_0/J_{H_m} plotted against H_m for a number of the ferrites in Table II.

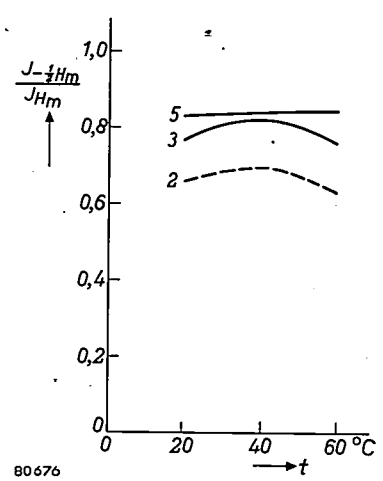


Fig. 12. R_s as a function of the temperature, for the three ferrites (2), (3) and (5) in Table II.

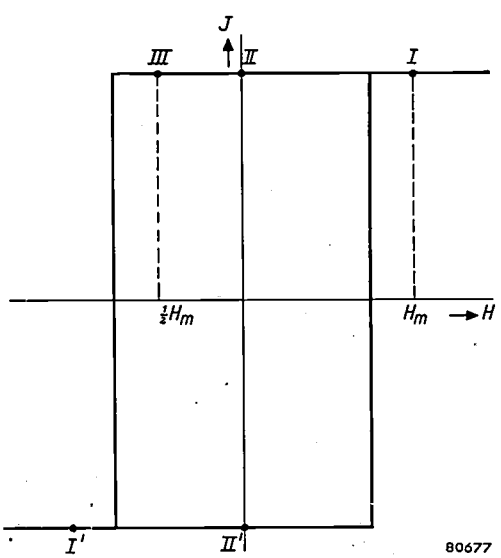


Fig. 13. Ideal rectangular hysteresis loop.

occur many times before a pulse of magnitude $-H_m$ arrives to bring the material into condition I' and then II', i.e. to reverse the sign of the magnetization.

How far the non-ideal available materials approximate to this behaviour may be gathered from *fig. 14*. This shows one half of the hysteresis loop corresponding to the optimum squareness ratio for one quality of ferrite. The point I was measured after the field H_m had gone through a number of cycles. When H_m is removed the point II is reached. An opposing field $-\frac{1}{2}H_m$ gives point III and, when this in turn is removed, a point IV is reached which is lower than II. The variation of points I - IV when the cycle I-II-III-IV is completed a number of times has also been investigated. Ferrite No. 2 in Table II was used for this purpose. It was found that after a large number of cycles of the subsidiary loop I-II-III-IV, the induction at point I had dropped by less than 1%. Point II remained constant within experimental error; point III rose: the corresponding induction value may increase by more than 5%. Point IV also rose considerably. The final situation is that the subsidiary loop has moved to the position shown by the broken line in fig. 14. Clearly, the squareness ratio R_s does not diminish but even increases, in this case from 0.76 to 0.81.

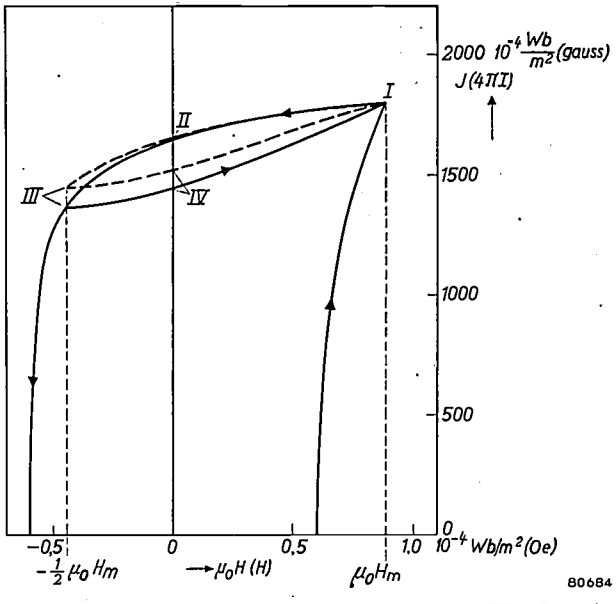


Fig. 14. Hysteresis loop of ferrite No.2 (see Table II) when the field is varied a number of times from $H = H_m$ to $H = -\frac{1}{2}H_m$.

Summary. For certain purposes (computing machines, switching elements) cores of magnetically soft material (i.e. with small coercive force) are required, having almost rectangular hysteresis loops. Ferrites fulfil these requirements and also have the advantage that eddy currents and other losses are only small when the field is varied rapidly. The shape of the hysteresis loop of ferrites is determined by the nature of the anisotropy governing the direction of the magnetization vector (crystal, stress or shape anisotropy). Pronounced crystal anisotropy is an advantage (and with it the accompanying low initial permeability μ_i), but it should not be so high that the coercive force becomes too great. In order to minimize the other kinds of anisotropy, internal strain and porosity should be avoided. A number of suitable ferrites especially developed for the purposes mentioned above are described and their properties enumerated.

MIRROR CAMERAS FOR GENERAL X-RAY DIAGNOSTICS

by W. HONDIUS BOLDINGH.

778.33:771.31:616-073.75

The use of fluorography is becoming more and more common and is now also employed in general X-ray diagnostics. Attempts to minimize the dosage to which the patient is exposed during this type of examination have developed along two quite separate lines, namely, the improvement of optical efficiency in photographic systems, and the use of electronic aids (e.g. the X-ray image intensifier) to increase the image luminance. It is difficult at the present stage to predict the ultimate relationship between the two methods; the former, however, has now attained a considerable measure of perfection. The present article describes some of the latest designs of the fluorographic cameras used.

Fluorography, that is, the photographing of fluorescent X-ray images with the aid of a camera instead of by direct contact with a film, was originally developed for mass chest survey. The merits of the method as applied to this particular branch of diagnostics have been discussed fully in earlier issues of this Review [1] [2]). All that we need recall here is that documentation is thus achieved without undue expense of film and filing space, and

[1]) A. Bouwers and G. C. E. Burger, X-ray photography with the camera, Philips tech. Rev. **5**, 258-263, 1940.

[2]) H. J. di Giovanni, W. Kes and K. Lowitzsch, A transportable X-ray apparatus for mass chest survey, Philips tech. Rev. **10**, 105-113, 1948/1949.

occur many times before a pulse of magnitude $-H_m$ arrives to bring the material into condition I' and then II', i.e. to reverse the sign of the magnetization.

How far the non-ideal available materials approximate to this behaviour may be gathered from *fig. 14*. This shows one half of the hysteresis loop corresponding to the optimum squareness ratio for one quality of ferrite. The point I was measured after the field H_m had gone through a number of cycles. When H_m is removed the point II is reached. An opposing field $-\frac{1}{2}H_m$ gives point III and, when this in turn is removed, a point IV is reached which is lower than II. The variation of points I - IV when the cycle I-II-III-IV is completed a number of times has also been investigated. Ferrite No. 2 in Table II was used for this purpose. It was found that after a large number of cycles of the subsidiary loop I-II-III-IV, the induction at point I had dropped by less than 1%. Point II remained constant within experimental error; point III rose: the corresponding induction value may increase by more than 5%. Point IV also rose considerably. The final situation is that the subsidiary loop has moved to the position shown by the broken line in fig. 14. Clearly, the squareness ratio R_s does not diminish but even increases, in this case from 0.76 to 0.81.

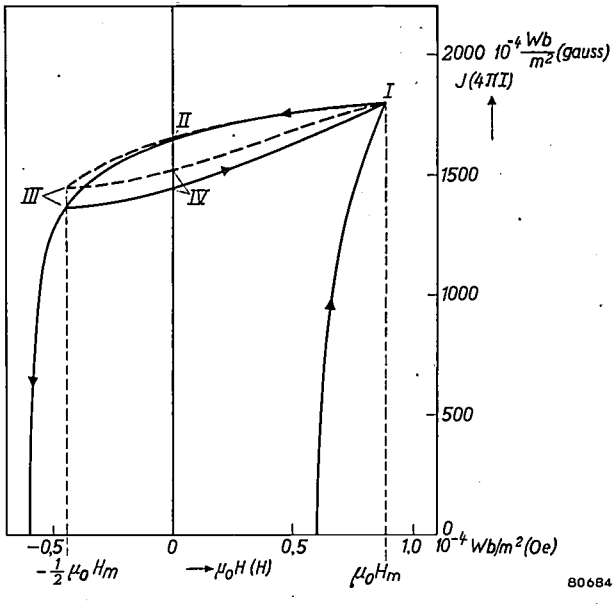


Fig. 14. Hysteresis loop of ferrite No.2 (see Table II) when the field is varied a number of times from $H = H_m$ to $H = -\frac{1}{2}H_m$.

Summary. For certain purposes (computing machines, switching elements) cores of magnetically soft material (i.e. with small coercive force) are required, having almost rectangular hysteresis loops. Ferrites fulfil these requirements and also have the advantage that eddy currents and other losses are only small when the field is varied rapidly. The shape of the hysteresis loop of ferrites is determined by the nature of the anisotropy governing the direction of the magnetization vector (crystal, stress or shape anisotropy). Pronounced crystal anisotropy is an advantage (and with it the accompanying low initial permeability μ_i), but it should not be so high that the coercive force becomes too great. In order to minimize the other kinds of anisotropy, internal strain and porosity should be avoided. A number of suitable ferrites especially developed for the purposes mentioned above are described and their properties enumerated.

MIRROR CAMERAS FOR GENERAL X-RAY DIAGNOSTICS

by W. HONDIUS BOLDINGH.

778.33:771.31:616-073.75

The use of fluorography is becoming more and more common and is now also employed in general X-ray diagnostics. Attempts to minimize the dosage to which the patient is exposed during this type of examination have developed along two quite separate lines, namely, the improvement of optical efficiency in photographic systems, and the use of electronic aids (e.g. the X-ray image intensifier) to increase the image luminance. It is difficult at the present stage to predict the ultimate relationship between the two methods; the former, however, has now attained a considerable measure of perfection. The present article describes some of the latest designs of the fluorographic cameras used.

Fluorography, that is, the photographing of fluorescent X-ray images with the aid of a camera instead of by direct contact with a film, was originally developed for mass chest survey. The merits of the method as applied to this particular branch of diagnostics have been discussed fully in earlier issues of this Review [1] [2]). All that we need recall here is that documentation is thus achieved without undue expense of film and filing space, and

[1]) A. Bouwers and G. C. E. Burger, X-ray photography with the camera, Philips tech. Rev. **5**, 258-263, 1940.

[2]) H. J. di Giovanni, W. Kes and K. Lowitzsch, A transportable X-ray apparatus for mass chest survey, Philips tech. Rev. **10**, 105-113, 1948/1949.

that a well-organised routine has been evolved during the examination of entire population groups.

The principal problem associated with the introduction of fluorography was the speed of the camera required to photograph the faint image on the fluorescent screen with a very short exposure. Lens cameras were initially used for fluorography but a considerable improvement was effected by introducing mirror cameras. Such a camera, based on the Schmidt optical system, and designed for 45 mm film, has been described earlier in this Review [3] [4]. This camera, subsequently modified to some extent, can make almost distortion-free photographs of a flat fluorescent screen of (effective) area 42 × 42 cm, reduced in size by a factor $r \approx 10.5$. It contains an optical system with a mirror of 166 mm diameter and acorrecting plate [5] of diameter (D) 125 mm. The focal length f of the system is 104 mm. The effective aperture ratio (see the article referred to in [4])) of the camera at the above reduction factor is $1{:}N_{\text{eff}} = 1 : 1.03$. To compute this quantity (which is a true measure of the light-gathering power, see article referred to in note [4])), use is made of the formula:

$$1 : N_{\text{eff}} = \frac{D}{f} \frac{r}{r-1} \sqrt{S}$$

where S is the transmission of the optical system (otherwise termed the shadow factor): $1 - S$ indicates what fraction of the light proceeding towards the mirror is intercepted by the film holder.

The above-mentioned camera is provided with a wide range of attachments for automatic and foolproof operation in mass chest surveys.

Two more cameras of a similar type have recently been developed for use with 35 mm and 70 mm films, primarily because these sizes have been either standardized or recommended in several countries. In principle, the mirror optical system is the same in all three cameras (the 70 mm model is an enlarged version of the 45 mm camera scaled-up approximately proportionally to the ratio of the film sizes). Moreover, the film transport mechanism and the accessories of the two new cameras are not fundamentally different from the earlier camera; no further description is therefore necessary in this article.

[3]) P. M. van Alphen and H. Rinia, Projection-television receiver, I. The optical system for the projection, Philips tech. Rev. 10, 69-78, 1948/1949.

[4]) W. Hondius Boldingh, Fluorography with the aid of a mirror system, Philips tech. Rev. 13, 269-281, 1951/1952.

[5]) Special consideration has been given to the diameter of the correcting plate. In practice the particular diameter adopted produces the optimum combination of light-gathering power and picture definition.

However, developments of another kind were taking place during the course of the work on the new cameras for mass chest survey, viz. the development of cameras for general diagnostics. The continued increase in the use of X-rays for general diagnostics has led to the desire to use fluorography also in this field. In some large hospitals the number of such examinations may mean over 1000 radiographs per day, the usual size, using contact radiography, being 30 × 40 cm; hence the use of fluorography for even a portion of the daily examinations can mean an appreciable saving in the use of film. The application of fluorography to this field has become practicable as a direct result of the introduction of the mirror camera, which permits of a shortening of the exposure and consequently gives photographs of improved quality, suitable for many diagnostic purposes. This possibility was anticipated in the previous article [4]).

Certain requirements for mass chest survey, e.g. simplicity of operation and the positive identification of photographs, apply perhaps less stringently to a camera for general diagnostics; greater emphasis however, must now be placed on the picture-quality. Moreover, the equipment must be adapted to suit the methods of general diagnostic examination. With this in view, three new cameras have been designed, one for single exposures, one for a series of up to 30 photographs, and one for a similar series at high speed; these will now be described.

Picture-size of the new cameras

Each of the mirror cameras for general diagnostics is designed to take 70 mm film, this being so economical as compared with the full-size contact picture that there is virtually no incentive to adopt a smaller size. The special merit of 70 mm film is that in many cases of general diagnostics the relevant details can be seen direct from the film without enlargement, and that a critical examination of the photograph can be accomplished quite well with a simple optical aid such as a magnifying glass. The relatively greater weight and volume as compared with cameras for smaller film-sizes is not inconvenient in this application, which does not involve transportation. It is in fact generally recognized that 70 mm is the most appropriate film-size for this purpose.

The actual picture is of course narrower than the film, which is masked on either side by the film gate against which it is pressed during the exposure. The strips of film thus obscured are wider than in cameras used for ordinary photography, since with

the Schmidt mirror optical system it is necessary to give a spherical curvature to the film: a not too narrow margin is required to give adequate purchase on the film during the process of spherical deformation. Hence the picture-width of the 70 mm film for camera chest examination was limited to 58 mm (reduction factor 7.2). This figure has also been adopted in the new cameras, now to be described; the optical systems of all these 70 mm cameras are therefore identical.

Single exposure camera

As will be seen from *fig. 1*, the design of the single exposure camera is relatively simple. To position the film (flat film, cut to size 70 × 70 mm) between the concave mirror and the correcting plate a sliding cassette is used; in this the film is moulded to the required spherical shape by spherical pressure plate (*fig. 2*). The axial tolerance of the position of the film, or, more precisely, of the position of the centre of curvature of the spherical film-surface, is extremely critical: owing to the unusually high aperture-ratio of this camera, the depth of focus is so minute that a film displacement of only a few tens of microns is enough to produce a perceptible decrease in definition. The problem of attaining so high a standard of precision in the positioning of the film cassette could of course be solved by employing an extremely accurate finishing process for the cassette slide, but in view of the inevitable wear on the sliding faces, and to ensure reproducibility of position when changing the cassette, another method was adopted. Each cassette is provided with three lugs (fig. 2) whose surfaces facing the correcting plate form a continuation of the convex film surface or, more accurately, of the contact frame around the film gate. When the cassette has been inserted, it can be moved towards the correcting plate by turning a lever, until the three cassette lugs rest against the points of three set screws rigidly fixed with respect to the optical components (*fig. 3*). Since a spherical surface of a given radius (and direction) of curvature is uniquely located by three fixed points, precise positioning of the film is ensured.

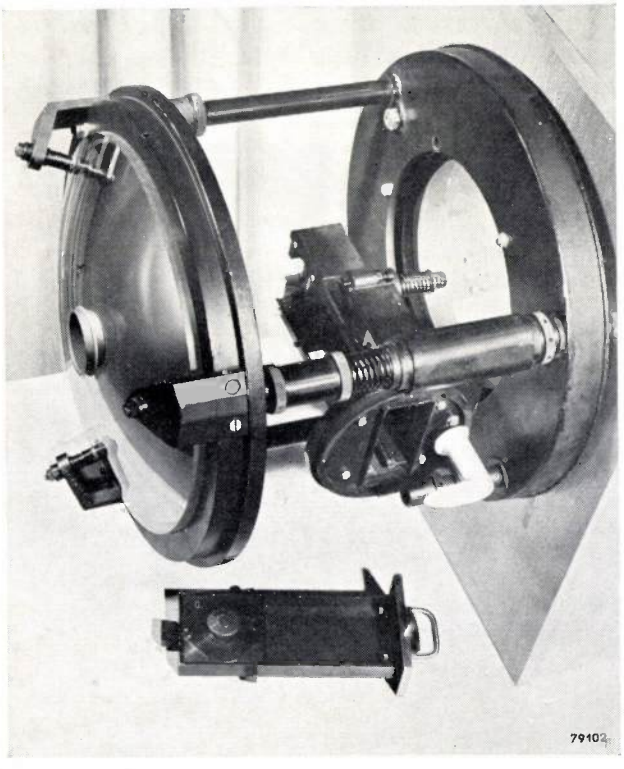


Fig. 1. Mirror camera for single exposures, with cover removed. (For the purposes of the photograph the camera is screwed to a wooden mount.) The concave mirror is on the left, and the correcting plate on the right of the photograph: between them is the cassette holder with slot for inserting the cassette, and a crank for pressing the cassette against the centering screws. A cassette is shown beneath the camera.

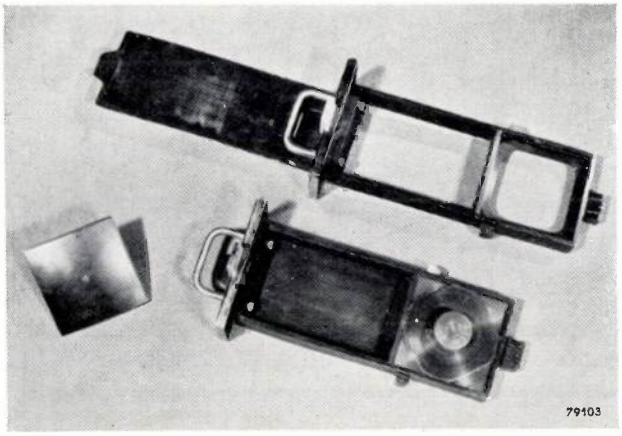


Fig. 2. A cassette (foreground) with the spherical pressure plate inserted; note the three lugs provided to ensure accurate centering. Above, the cassette with the spherical pressure plate (left) removed and the cassette cover withdrawn.

The loss of light in this camera is smaller than in the other mirror cameras mentioned above, owing to the fact that cut film is used: no light-tight film-guide is therefore required between the gate and the outside of the camera, and when the cassette cover is withdrawn to expose the sensitized film surface, the optical system is substantially free of obstruction save for the film itself in the centre. Thus the transmission in this camera is very high, viz. $S = 0.79$, as compared with 0.49 in the 70 mm camera for mass chest survey, 0.55 in the 70 mm serial camera and 0.53 in the 45 and 35 mm chest survey cameras. The present instrument thus has by far the highest light-gathering power, its effective aperture ratio being 1 : 0.80 as compared with 0.96, 1.0, 1.03 and 1.05 respectively for the other types referred to. The exposures required for this camera

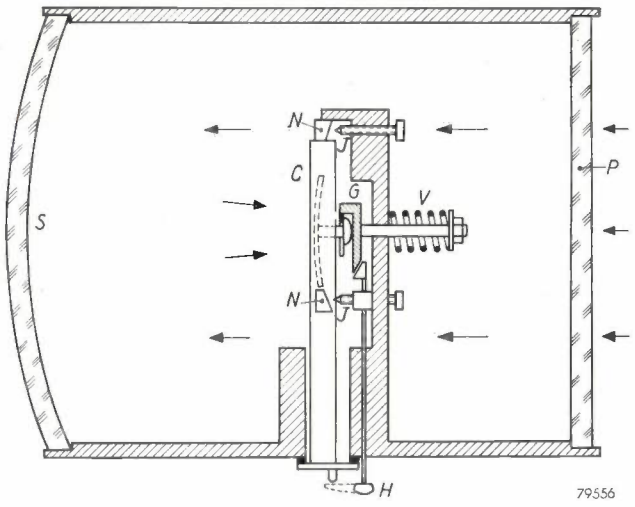


Fig. 3. Diagram illustrating the centering of the film. When the cassette *C* is inserted, the knob of the spherical pressure plate (fig. 2) engages with a bracket *G*. When the crank *H* is rotated a cam bearing on the bracket *G* allows the latter and hence also the cassette to be pulled to the right by spring *V* until the lugs *N* of the cassette rest upon the points of the set screws *J*. *S* is the concave mirror; *P* is the correcting plate.

are only about three times longer than those necessary for contact photographs, assuming the same voltage and current on the X-ray tube; they can of course be shortened considerably by increasing the voltage on the X-ray tube. The consequent loss of contrast may be largely off-set by the use of a film with a higher gamma.

Camera for serial exposures

The serial camera (*fig. 4*) is designed for use in cases where frequent X-ray examinations are to be made so that it is not convenient to change the cassette before each exposure. The film is transported from a dispenser cassette capable of accommodating 30 m of film (enough for 400 photographs) to the film gate, and from there to a receiver-cassette capable of storing up to 30 photographs. When this total is reached, or sooner if desired (if necessary immediately after each individual exposure), the film can be cut off and the receiver-cassette removed from the camera to develop the exposed strip of film.

This camera is so designed that the cutting-off of individual photographs can be done without wasting relatively long strips of film; this has been achieved by a special modification of the usual film transport. Normally the film is *drawn* through the camera and over the spherical pressure plate,

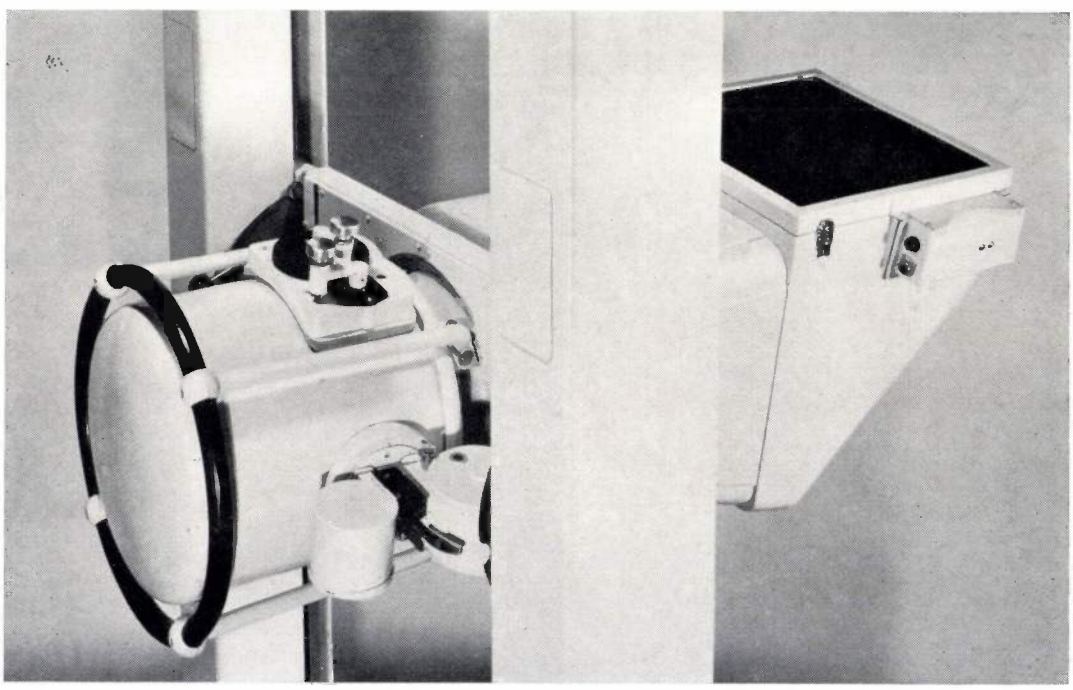


Fig. 4. Mirror camera for a series of up to 30 photographs. The dispenser cassette for 400 photographs (30 m of film) and the receiver-cassette are seen at the side of the camera; the film transport lever and the exposure counter, and another lever for film cutting are at the top of the camera. The camera is secured to an angular hood which slides up and down on vertical rails: the other end carries the fluorescent screen and a plane mirror set at an angle of 45° to the screen. The hood and camera can be rotated through approximately 270° about the axis of the camera so that, for example, photographs of a patient lying on an ordinary examination table can be taken vertically upwards or downwards.

by a traction wheel outside the optical system. The film cannot, of course, be severed between the traction wheel and the film gate, since the wheel could then exercise no pull on the film. To develop a photograph immediately after exposure (that is, without waiting until a number of other photographs has been taken), the portion of film concerned must therefore be advanced beyond the traction mechanism, leaving several picture-lengths unexposed.

In the serial camera, this is avoided by *pushing* instead of pulling the film strip past the gate. This permits the strip of film exposed and fed forward to be cut off very close to the film gate (*fig. 5*); only $2 \times {}^1/_2 = 1$ picture-length per cut need then be spoiled (this being necessary to ensure the complete exclusion of light from the film on either side of the cut); hence the loss per strip of film cut off is equivalent to only one photograph. When the film is wound forward after the next exposure it passes through a funnel-shaped guide into a new receiver-cassette, in the place of the one removed.

To operate the film transport of the camera, a crank is rotated one full turn by hand. This releases the film in the film gate, transports the film by one frame and re-applies the spherical deformation to the film.

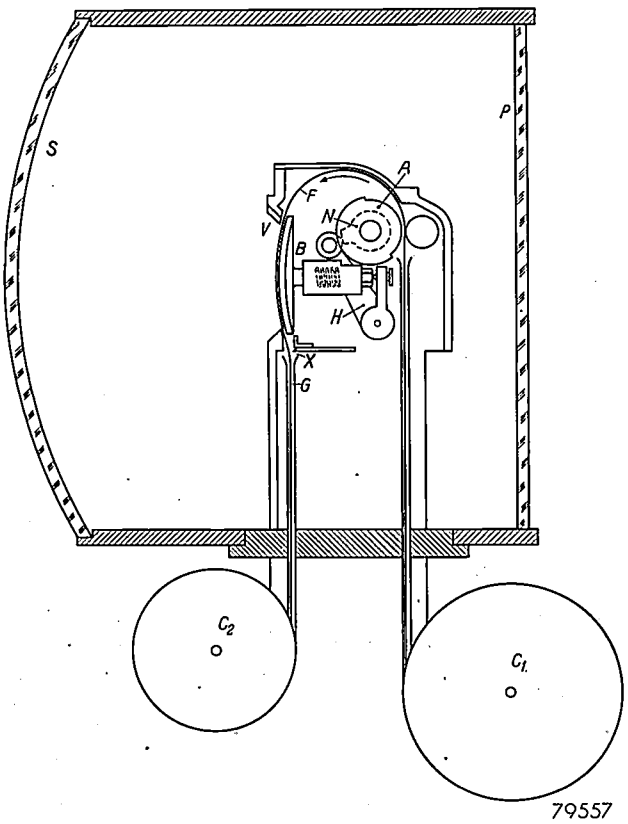


Fig. 5. Section (greatly simplified) of the camera for serial exposures. Note the friction roller A which actuates the spherical pressure plate B by means of a cam N and a lever H and aided by a pressure roller, pushes the film strip F past the film gate V. Also shown are the cutter at X and the funnel-shaped guide G to the receiver-cassette C_2, which receives the film as fed forward. C_1 is the dispenser-cassette.

Although the optical system is the same as that of the single exposure camera described above and the mass chest survey camera, the three cameras differ so appreciably in other particulars of design that no attempt has been made to furnish them with identical or interchangeable components. Uniformity of this kind would merely make each type individually more complex and more expensive.

Serial camera with rapid film transport

"Functional" X-ray examination, whereby the motions of functioning organs are demonstrated, usually with the aid of a contrast agent introduced into the body, is an important branch of radiology. Such examinations often involve taking upwards of ten photographs in quick succession. The development of this valuable diagnostic technique using contact radiography has been hitherto impaired by the high cost of film and the difficulty of attaining the desired rapid succession of exposures.

With the introduction of fluorography conditions have become much more favourable for the development of this technique. It has been found possible to design a film transport mechanism for the serial camera operating at a speed that will produce 5 pictures per second. This rapid film movement is accomplished with the aid of an electro-mechanical drive operating in the manner demonstrated in *fig. 6*. An electric motor mounted on the camera,

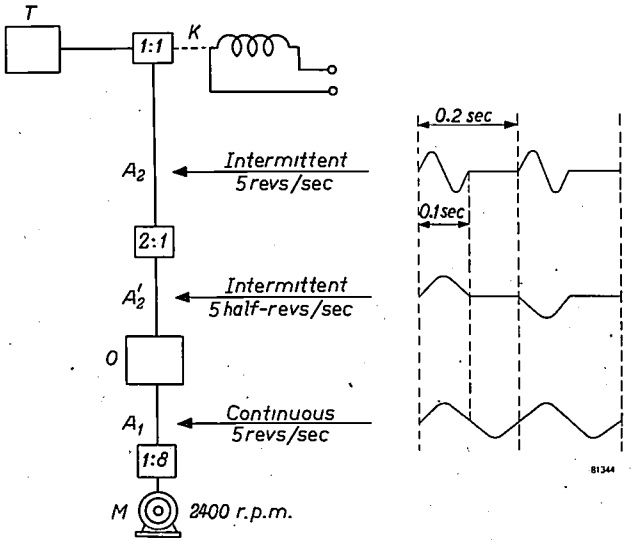


Fig. 6. Schematic diagram of the film transport mechanism of the rapid sequence serial camera. The revolutions of the continuously rotating spindle A_1 and of the intermittently rotating spindles A_2' and A_2 are shown schematically on the right. M motor, O feed-mechanism, K electromagnetic clutch, T film transport and actuating mechanism for pressure plate. As long as K is engaged, the film is advanced 5 times per second and a photograph (exposure not exceeding 0.1 sec) is taken after each advance.

drives a spindle at a speed of 5 revolutions per second (A_1), which actuates a feed-mechanism (O). This in turn drives another spindle (A_2) intermittently, so that it rotates one full turn in $^1/_{10}$ second, remains stationary for a similar period, then executes another full turn, and so on. This intermittently rotating spindle can be coupled by means of an electromagnetic clutch to the film-transport drive, which is otherwise identical with that of the serial camera already described. When this clutch is engaged (during a stationary period) the next revolution of A_2 causes a complete cycle of the film transport, that is, the retraction of the spherical pressure plate, the feeding forward of the film and the spherical deformation of the next frame of film. During the subsequent stationary period of A_2, which lasts $^1/_{10}$ of a second, a fluorogram can be recorded; for this purpose a contact in the camera transmits an electric signal to the time switch of the X-ray apparatus. At the end of the stationary period, the spindle again rotates, the film is fed forward, and so on.

It will be seen that as long as the magnetic clutch remains engaged, 5 photographs per second will be recorded, each with an exposure not exceeding $^1/_{10}$ second (the time switch of the X-ray apparatus is of course pre-set to give the desired exposure). The electromagnetic coupling, however, between the intermittently rotating spindle A_2 and the film transport can be interrupted at will for one or more periods after the forward movement of the film, so that fewer photographs are taken per second (if necessary with exposures exceeding $^1/_{10}$ second). This is accomplished with the aid of a specially designed accessory known as the programme selector (*fig.* 7), containing a series of 50 contact knobs

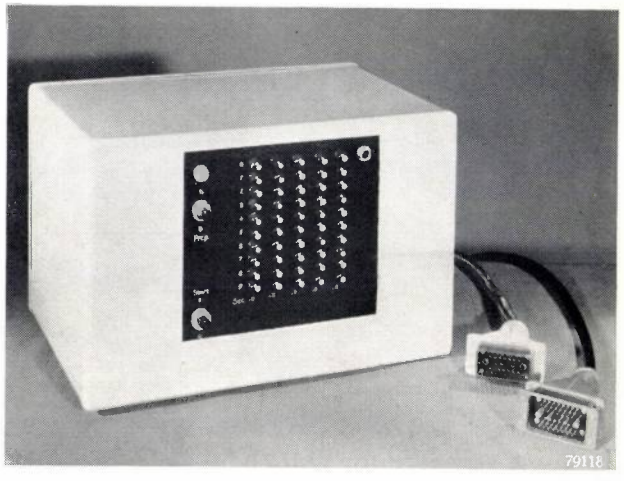


Fig. 7. Programme selector for camera with high-speed film transport. The 50 press-buttons correspond to the 50 periods of $^1/_5$ second occurring within 10 seconds. Whether or not the magnetic clutch is to be engaged, i.e. whether or not a photograph is to be taken in any particular period is determined in advance by means of the appropriate button.

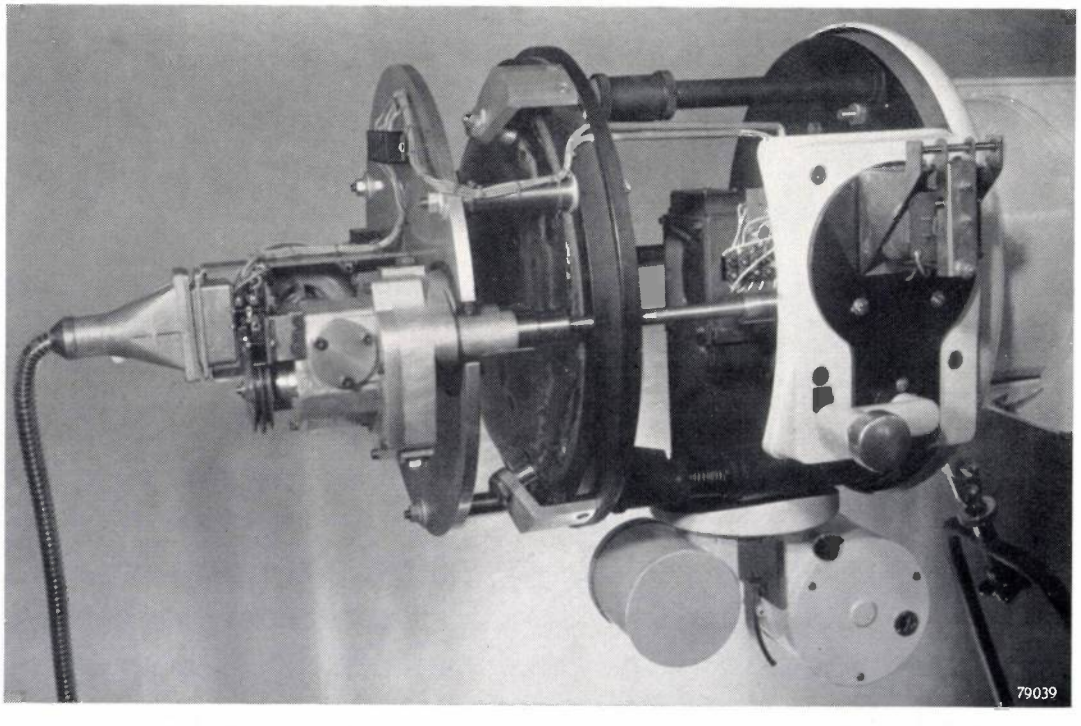


Fig. 8. Interior of mirror camera with high speed film transport. The feed mechanism is seen on the left, and behind it the electric driving motor; on the right is the electromagnetic clutch which couples the intermittent spindle (A_2 in fig. 6) to the actual film transport at the predetermined intervals

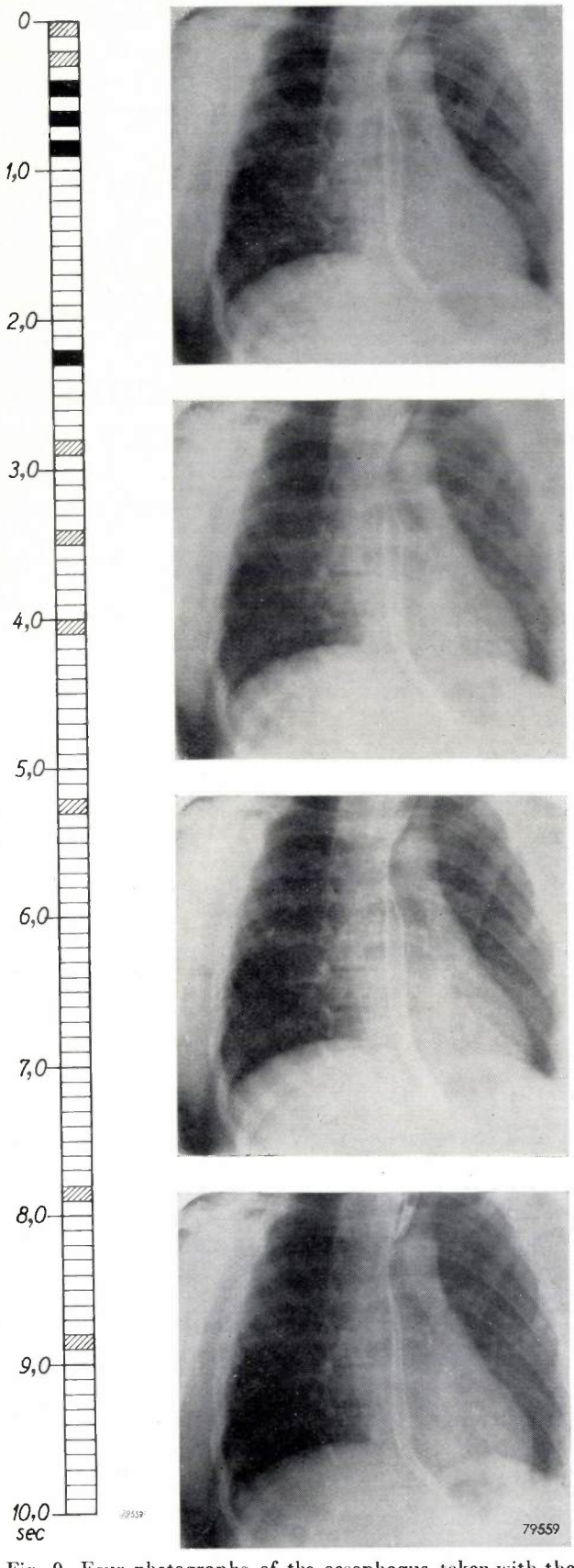


Fig. 9. Four photographs of the oesophagus taken with the rapid sequence mirror camera. The scheme of the whole series is seen on the left; hatched areas indicate periods in which photographs were taken; the four blackened areas correspond to those reproduced here. The contrast agent administered enables the act of swallowing to be followed.

corresponding to the 50 revolutions (cycles) performed by the intermittently rotating spindle in 10 seconds. Before every series of exposures, the magnetic coupling for each period is pre-set through a system of relays by means of the appropriate knob, that is, it is possible to determine beforehand whether or not the film shall be advanced and the X-ray apparatus switched on during a particular period. In this way it is possible to map out the whole scheme of photographs before exposure.

The feed-mechanism (driven by the continuous spindle A_1), which governs the intermittent rotation of spindle A_2, is a counterpart of the well-known Maltese-cross mechanism used in film projectors [6]). The Maltese cross mechanism of the type most commonly used, rotates $^1/_4$ turn in a quarter period of the driving shaft and then remains stationary for the remaining three-quarters of the period. In our system, which is based on a picture-period of $^1/_5$ second, the above time-ratio would leave only $^1/_{40}$ second for the actual film transport (another $^1/_{40}$ second being required for the spherical deformation of the film). This involves a film acceleration so high as to involve a serious risk of damage to the film. For this reason a new mechanism with a feed period equal to the stationary period was designed to supply the intermittent movement required. The intermittent spindle A_2 of this mechanism performs one half revolution in the first $^1/_2$ cycle of the continuous spindle A_1, and remains stationary during the remainder of that cycle. A 2:1 gear coupling A'_2 to A_2 produces in the latter (which carries the friction wheel of the film transport) a rotation through one full turn. This gear-up is necessary because if the friction wheel were carried on the shaft A'_2 ($^1/_2$ turn per cycle of operations) the diameter necessary to transport the film with the picture-size adopted would be inconveniently large.

A photograph of the high-speed camera with covers removed is shown in *fig. 8*.

As an example of the results obtainable with the high-speed camera, *fig. 9* shows a series of photographs taken during a functional diagnostic examination of the oesophagus. The rate at which the contrast agent (or an air-bubble contained in it) descends through the oesophagus during the act of swallowing can be ascertained by comparing these photographs. A good deal of the clarity is

[6]) A description is given in an article by J. J. Kotte on a professional 16 mm film projector, to be published shortly in this Review.

necessarily lost in the half-tone reproduction but the quality of the original photographs is in every way sufficient for such an examination.

Summary. Fluorography, originally developed for mass chest surveys, is now also becoming important in general X-ray diagnostics. Three Schmidt type mirror cameras designed for this purpose, each using 70 mm film and having a reduction factor of 7.2, are described. The first, of effective aperture ratio 1 : 0.80, is for single exposures. The method of film-centering in this camera is described in detail. The second camera, which is fitted with dispensing and receiving cassettes, can take a series of up to 30 photographs, but the film can be cut off and developed after only one or more exposures. The film wastage is minimized by pushing, instead of pulling the film through the film gate. The last of the three, largely identical with the serial camera just mentioned, has a high-speed film transport which permits 5 photographs to be taken per second, a valuable feature for functional radiography.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Administration of the Research Laboratory, Eindhoven, Netherlands.

2078: W. K. Westmijze: Gap-length formula in magnetic recording (Acustica **2**, 292, 1952).

A formula is derived for the dependence of the frequency characteristic of a magnetic reproducing head on the gap-length.

2079: H. G. van Bueren: On the attraction between a perfectly conducting plate and a thin perfectly conducting cylinder (Proc. Kon. Akad. Wet. Amsterdam **B 55**, 493-499, 1952, (No. 5).

By analogy with the formula of Casimir, $E = -(\pi^2/720)(h/2\pi)cL^2/R^3$, for the interaction between two parallel conducting plates of area L^2, due to zeropoint energy, it is found that the interaction energy between a thin wire (radius r_0, length L) and a square plate of area L^2 is

$$E = -(3/16)\ (h/2\pi)\ cr_0^2L/R^4.$$

If $r_0 = 2\mu$ and $R = 5\mu$ the force of attraction is equal to that due to a potential difference of about 0.3 mV.

2080: H. Bruining: Quelques points de vue nouveaux concernant la construction et l'utilisation de l'image iconoscope (Le Vide **71**, 1248-1255, 1952, No. 42). (Some new points of view regarding the construction and utilization of the image iconoscope: in French.)

In this article on the construction of an image-iconoscope, special attention is giving to means of suppressing ion burn. This is achieved by placing a fine-mesh grid close to the photo-cathode, thus avoiding concentration of the ion beam. The advantage of using a L-cathode (diffusion cathode) in the electron gun is stressed. In addition, an electron lens system is described allowing a continuous change of the focal length of the camera. A special device is described for ensuring equality of brightness, especially at the boundary of the image field. See also Philips tech. Rev. **14**, 327-335, 1952-53.

2081: B. D. H. Tellegen: Synthesis of four-poles (Proc. Symp. Modern Network Synthesis, New York N.Y., 1952, pp. 40-49, publ. by Polyt. Inst. Brooklyn N.Y.).

General considerations on the synthesis of four-poles with preconceived properties, by means of inductances, capacitances, resistances and ideal transformers. A fifth possible type of network element is the gyrator. See these abstracts, No. **R73**.

2082: J. L. Meijering: Calculs thermodynamiques concernant la nature des zones Guinier-Preston dans les alliages aluminium-cuivre (Rev. Métall. **49**, 906-910, 1952). (Thermodynamical calculations concerning the Guinier-Preston zones in aluminium-copper alloys; in French).

According to calorimetric measurements on solid Al-Cu alloys by Oelsen and Middel the enthalpy of mixing is negative over the entire range of concentrations. This appears to be in contradiction to the current picture of the initial stages of precipitation hardening in aluminium with 5% Cu, which demands (when rather forced explanations are to be avoided) a segregation tendency in the face-centered cubic phase. Such a tendency is commonly due to a *positive* mixing-enthalpy curve. In this paper it is shown, by combining the calorimetric data with the solubility curve of Al_2Cu in Al, that the mixing-enthalpy curve is partly concave, partly convex, this making the contradictions disappear. Similar strongly asymmetric mixing-enthalpy curves must also appear in the systems Al-Ag and Pt-Ag.

2083: J. I. de Jong, J. de Jonge and H. A. K. Eden: The formation of trimethylol urea (Rec. Trav. Chim. Pays-Bas **72**, 88-90, 1953).

In concentrated aqueous solution and in the presence of an excess of formaldehyde, more than two

methylol groups may be attached to one molecule of urea. The equilibrium constant of the formation of trimethylol urea is evaluated.

2084: K. H. Klaassens and C. J. Schoot: Derivatives of p-diethoxybenzene, I. 1,4-diethoxy-2-chlorobenzene-5-diazonium-borofluride (Rec. Trav. chim. Pays-Bas **72**, 91-93, 1953).

Description of the preparation of the above-named compound, and confirmation of its structure.

2085: W. J. Oosterkamp: The radiography of the human body with radioactive isotopes (Brit. J. Radiology **26**, 111, 1953).

The activity of a number of radioactive isotopes, per mm^2 of surface of a layer of a thickness equal to one half-value layer (with a maximum of 10 mm), is compared to the emission, per mm^2 focus, of X-ray tubes (stationary and rotating anode). It is shown that the use of radioactive isotopes, even if short-lived and carrier-free, for medical radiography is only attractive for those applications where the use of X-ray tubes is not practicable.

2086*: J. D. Fast and E. M. H. Lips: Metallurgical research in the Netherlands (Metal Progress **63**, 109-111, 1953).

Some practical results in the field of metallurgical research are enumerated, e.g. dies for deep drawing, blanking operations, drawability and brittleness of sheet metal, permanent magnets ("Ticonal" and "Ferroxdure"), gases and metals, hardening by internal oxidation, influence of admixtures on scaling rate, embrittlement of iron by oxygen, ageing, and welding with contact electrodes.

2087: J. M. Stevels: Note on the ultraviolet transmission of glasses (Proc. 11th Int. Congress pure & appl. Chem. **5**, 519-522, 1953).

Considerations on the influence of bridge-oxygen ions and non-bridging oxygen ions on the ultraviolet absorption limit of glasses. Bridge oxygen shifts the absorption edge towards smaller wavelengths, non-bridging oxygen towards larger wavelengths.

2088: J. I. de Jong and J. de Jonge: Kinetics of the formation of methylene linkages in solutions of urea and formaldehyde (Rec. Trav. chim. Pays-Bas **72**, 139-156, 1953).

A kinetic investigation is made of the reactions between monomethylol urea, dimethylol urea and urea in acid aqueous solution. Strong indications are obtained that these reactions are all of one type, viz. bimolecular hydrogenion-catalyzed reactions between an amidomethylol group and an amide group, leading to the formation of methylene linkages between urea fragments:

$$-NH.\,CH_2OH + -NH_2 \rightarrow -NH.\,CH_2.\,NH- + H_2O.$$

The rate constant of this reaction appeared to depend on the type of amide group or amidomethylol group that is reacting. The activation energy is about 15 kcal/mole. The reaction between two molecules of dimethylol urea was found to be very slow if occurring at all. The possibility of cyclic structures arising from trimerisation of monomeric methylene urea and of the formation of dimethylene ether linkages between urea fragments may be excluded.

R 206: J. D. Fast and J. L. Meijering: Anelastic effects in iron containing vanadium and nitrogen (Philips Res. Rep. **8**, 1-20, 1953, No. 1).

The maximum quantity of nitrogen taken up by an iron wire containing 0·5 atomic % vanadium during heating at 950 °C in N_2 with 1% H_2, corresponds to one atom N per atom, plus a further amount of the same order as the quantity taken up by pure Fe under identical conditions. The first amount combines chemically with the vanadium and causes no internal friction. The extra quantity gives rise not only to a damping peak corresponding to that in pure iron (with an oscillation period of 1.3 seconds at 21.5 °C), but also to a peak at higher temperatures. This peak cannot be described with a single relaxation time only. It is due to the presence of (submicroscopic) VN particles in the metal. These do not directly cause damping, but cause the free N-atoms to be bound much tighter in the surrounding interstitial sites than in the normal interstices. These abnormal interstices, therefore, will capture free N-atoms rapidly, whereupon the latter give rise to the abnormal damping. The binding energy in the abnormal interstices is not the same for all sites, and with coarsening of the VN precipitate the distribution of these energies is displaced in the direction of stronger binding. This is deduced from a shift of the summit of the second peak towards higher temperatures (from 80 °C to 88 °C in the authors' experiments) as the heating time at 950 °C is prolonged. From the intermediate state where they cause the abnormal damping, the N-atoms pass over rapidly into the fully precipitated state (iron nitride), where they cause no damping. Consequently, the VN precipitate exerts a strongly accelerating influence on the precipitation of dissolved nitrogen.

R 207: C. G. J. Jansen and R. Loosjes: The velocity distribution of electrons of thermionic emitters under pulsed operation, Part I, Apparatus and measuring technique (Philips Res. Rep. **8**, 21-34, 1953, No. 1).

This paper describes the construction of a tube for detecting the electron-velocity distribution of emitting surfaces, especially of oxide coatings at high current densities. With the apparatus used in conjunction with this tube it is possible to measure the electron velocities with an accuracy of about 1%. The i-V characteristic of the total emitting surface (8 mm²) and of the areas (0·03 mm²) whose velocity spectra are observed, can be determined with rectangular pulses or D.C. Complete i-V characteristics of the total emitting surface were also determined with pulses with a linearly sloping flank having a time interval of about 5 microseconds. Typical velocity spectra obtained from (BaSr)O, BaO and SrO coatings, and from the L-cathode (diffusion cathode) are shown. With tubes of similar construction equipped with an L-cathode, peak voltages can be determined with an accuracy of about 1 volt, independently of repetition frequency or pulse width.

R 208: H. C. Hamaker: The efficiency of sequential sampling for attributes, Part I. Theory (Philips Res. Rep. **8**, 35-46, 1953, No. 1).

In principle, Wald's probability-ratio sequential plans require three parameters for their specification. It is shown that for practical purposes we may with advantage use the two-parametric set of plans with decision lines symmetric with respect to the origin. There is no specific advantage in using an asymmetric position of the decision lines, while for the symmetric position the equations for sequential sampling can be greatly simplified.

R 209: J. W. A. Scholte and W. Ch. van Geel: Impedances of the electrolytic rectifier (Philips Res. Rep. **8**, 47-72, 1953, No. 1).

The system aluminium/aluminium oxide/electrolyte behaves as a rectifier. The conductivity of the layers out of which the oxide is composed changes with the externally applied potential difference. This article describes a method of deriving an electrical equivalent circuit of the oxide layer (consisting of a number of capacitors with resistors in parallel) from the frequency dependence of the impedance. The dependence of the state of the oxide layer on the D.C. potential across it is shown in a number of diagrams, in which the specific resistance is shown as a function of the position in the oxide layer. All diagrams show a layer with a high resistance and a layer in which the resistance decreases sharply, followed by a layer of low resistance. With electric fields in the conducting direction, no permanent change of the oxide appears and only the resistance of the high-resistance layer varies with the applied voltage. In stronger electric fields which do cause a permanent change of the oxide layer, the highly conductive layer grows at the expanse of the less conducting layer or vice versa. The electrical equivalent circuit is a starting point for discussion of the structure of the oxide layer. The conclusion is that both at the aluminium boundary and at the electrolyte boundary, the composition of the oxide shows a deviation from the simple stoichiometric ratio and is a semiconductor. It is assumed that rectification occurs by means of the contact between p-type and n-type semiconducting layers.

R 210: C. G. J. Jansen and R. Loosjes: Graphs for rapid calculation of the work function of thermionic emission (Philips Res. Rep. **8**, 81-90, 1953, No. 2).

These graphs are based on Richardson's formula, either in the form $j_s = A_0 T^2 \exp(-e\varphi/kT)$, or in the form $j_s = AT^2 \exp(-e\varphi_0/kT)$, where j_s is the saturation current density, φ the work function at temperature T, φ_0 the same at $T = 0$, e the electronic charge, k Boltzmann's constant, and $A_0 =$ 120 A/cm². The latter form takes into account the dependence of φ on T, making the constant A differ from A_0. The authors prefer the first form, however, from which φ can be derived at a given temperature. j_s is measured by a rapid oscillographic method. By computing φ for a number of temperatures it is possible to determine φ_0 and A, if desired. A number of graphs are added, making calculations unnecessary.

R 211: C. M. van der Burgt: Dynamical physical parameters of the magnetostrictive excitation of extensional and torsional vibrations in ferrites (Philips Res. Rep. **8**, 91-132, 1953, No. 2).

Tensile and torsional vibrations can be easily excited by magnetostriction in non-conducting ferromagnetics like ferrites. A simple experimental method permits rapid determination of the dynamic elastic and magnetoelastic constants, the complex nature of which is discussed after a comprehensive survey of the four sets of simultaneous magnetostriction equations of interest, under adiabatic conditions and under arbitrary depolarization. In reasonable agreement with the theory, the stress-

sensitivity constant and the magnetomechanical coupling coefficient of several Ni-Zn ferrites proved to be of the same order of magnitude as those of metallic magnetostrictive materials in common use. Moreover these ferrites (Ferroxcube IV materials) show much lower elastic dissipation at ultrasonic frequencies. Mechanical Q-factors up to 15 000 have been obtained at 50 kc/s. The variation of elastic and magnetic lag with frequency and biasing polarization is discussed. A considerable part of the total elastic losses at optimum bias consists in macro-magnetoelastic losses accompanying the macro-polarization induced magnetostrictively. The experimental correlation between the conductivity and the elastic and magnetic losses is explained in terms of elastically and magnetically excited micro strains that give rise to an electronic diffusion process. The order of magnitude of the molecular field is derived from the influence of the magneto-caloric effect on the magnetoelastic constants near saturation.

R 212: S. Duinker: An approximate graphical analysis of the steady-state response of non-linear networks (Philips Res. Rep. **8**, 133-147, 1953, No. 2).

The steady-state response of essentially non-linear networks containing iron-cored inductors with simultaneous a.c. and d.c. magnetization is analysed by means of an approximate graphical procedure. It is shown, that under various conditions, jump phenomena (so-called ferro-resonance effects) may occur when circuit parameters (e.g., applied voltage, polarizing voltage, etc.) are varied gradually. Instabilities are found to exist in the series circuit consisting of a non-linear inductor, a linear capacitor and a small resistor when driven by a voltage generator, and also in the parallel circuit of the same elements but with a high value of the resistor and driven by a current generator. Three different kinds of jump phenomena can be distinguished depending on whether or not a jump corresponds to a transition from a capacitive to another capacitive state or from a capacitive to an inductive state or vice versa and, further, on the number of stable possibilities linked with the effect. It is pointed out that the analysis is not restricted to the special configuration of non-linear inductors considered but that it also applies to circuits containing capacitors and resistors with centro-symmetrical characteristics. The results arrived at may be of value in connection with the investigation of the influence of resistive and reactive loads of magnetic and dielectric amplifiers and the application of jump phenomena in ferro-resonant flip-flops.

R 213: W. K. Westmijze: Studies on magnetic recording, Part I (Philips Res. Rep. **8**, 148-157, 1953, No. 2).

In this series of papers some problems are treated concerning the physics and mathematics of magnetic recording. In particular those problems are dealt with that arise in the recording of sound, where a strictly linear relationship between original and reproduced signals is required. This introduction gives a brief survey of the principle, some technical details and the history of magnetic recording. The mutual relation of the problems to be treated is explained.

R 214: W. K. Westmijze: Studies on magnetic recording, Part II (Philips Res. Rep. **8**, 161-183, 1953, No. 3).

The magnetic field in front of the gap of some simple types of recording head is calculated as the solution of a two-dimensional potential problem. Applying the reciprocity principle the magnetic flux through the coil of a reproducing head, originating from a sinusoidally magnetized tape in front of the gap, is deduced. It is shown that the well-known gap-loss formula $(\sin 2\pi l/\lambda)/(2\pi l/\lambda)$, where l is the gap length and λ the wavelength on the tape) holds only in a theoretical case. A more general formula is given.

R 215: E. S. Rittner: A theoretical study of the chemistry of the oxide cathode (Philips Res. Rep. **8**, 184-238, 1953, No. 3).

A comprehensive theoretical analysis of the chemistry of the oxide cathode, based upon thermochemistry and diffusion theory, is presented. The treatment is based on the conventional supposition that excess barium is required to activate the coating, and accordingly a search is made for materials of sufficient reducing power to serve as activators. Metals which fall into this category include: Th, Mg, Be, Hf, Sc, Y, Sm, Nd, Pr, La, Zr, U, Al, Si, C, and possibly Ti and Ce. A detailed analysis of the factors limiting the generation of free Ba reveals that the most favourable reaction mechanism is that in which the reaction speed is limited by the rate of diffusion of activator in the core metal. The free Ba subsequently finds its way into the individual oxide via the processes of Knudsen flow of the vapour and volume diffusion. An important requirement for the latter process is that the coating be constituted of a porous mass of fine particles. The paper concludes with a discussion of the evaporation loss during life of the excess Ba.

VOL. 16 No. 3, pp. 69-104 SEPTEMBER 1954

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

FUNDAMENTALS OF THE GAS REFRIGERATING MACHINE

by J. W. L. KÖHLER and C. O. JONKERS.

621.573

A hot-air engine suddenly deprived of heat from its burner, but kept in motion by an electric motor, functions as a refrigerator ("gas refrigerating machine"). In this sense it is entirely analogous to the steam engine which, although perhaps not widely known, functions as a (compression) refrigerator if the process of heating is discontinued and motive power is supplied from outside to keep the engine turning: the water in the boiler then cools rapidly and finally reaches a temperature below that of the surrounding atmosphere (the pressure in the condenser becomes higher than that in the boiler). It will be noted, however, that whereas the steam engine and the compression refrigerator have developed along entirely different lines, a gas refrigerating machine can be designed on principles very closely related to those of the hot-gas engine. This is substantiated by the theoretical arguments contained in this article, and will be further elucidated in a subsequent article dealing with the gas refrigerating machine recently developed at Philips.

Although even in primitive times man was able to heat objects (the knowledge of fire dates back to the very beginnings of the stone age), thousands of years elapsed before he learned how to effect cooling to temperatures below that of the surrounding atmosphere. The Egyptians are known to have kept their beverages cool by storing them in porous vessels, but ice could not be obtained even in Roman times by any means other than by fetching it from some natural source, often a distant mountain. Machines for generating "cold" were described for the first time during the nineteenth century.

The late development of methods of effecting cooling may be related to the fact that it is very much less essential to human life than heating. However this may be, it is certainly also largely attributable to the appreciably greater complexity of methods of producing cold. Methods of generating heat are legion and require no enumeration; on the other hand, only a few — usually complex — methods of generating cold are known even at the present time, and only a few of these are applied in practice. Some of these methods are outlined below.

Probably the oldest, and still the most widely used method of effecting a decrease in temperature is cooling by evaporation. A liquid is allowed to evaporate, and the vapour is removed, so that no state of equilibrium can be reached; the evaporation thus continues without interruption. Heat is absorbed by the evaporation process and the liquid therefore cools. The simplest example of this process is the sensation of cooling perceived when a moistened finger is held in the wind; another is the use of porous vessels which, as we have already seen, effect cooling by the gradual evaporation of liquid through the porous walls. In machines that operate continuously on this principle, the vapour (the refrigerant) is restored to the liquid phase usually by pumping and compressing the vapour; machines which do this are compression refrigerators. The absorption refrigerator is another type based on cooling by evaporation.

The lowest temperature economically attainable with these refrigerators depends primarily upon the decrease in the vapour pressure of the particular refrigerant with the temperature; the great volume of vapour to be pumped in the case of a low vapour pressure would necessitate the use of a very large machine. For this reason, temperatures below about —60 °C to —80 °C cannot be attained with such refrigerators, despite the use of modern refrigerants. These temperatures are low enough,

e.g. for the production of solid carbon dioxide ("dry ice", sublimation point —78 °C at 1 atm.), a process nowadays widely used for storing and conveying cold to places where it is needed (refrigerative containers) — in the same way as with ordinary ice. To obtain temperatures below —80 °C with an evaporation refrigerator, two or more cyclic processes, each involving a different refrigerant must be linked together; this "cascade method" was used at one time but is now almost completely obsolete, owing to its complexity.

There is another, very much lower range of temperatures which is now frequently and relatively easily attained by means of machines operating on quite different principles. This is the range associated with liquid air, the boiling point of which is —194 °C at 1 atm. (*fig. 1*). The liquefaction of air on a commercial scale was first accomplished by Carl von Linde. He employed the Joule-Kelvin effect that is, the decrease in temperature that takes place in all gases below a certain temperature when passed through an orifice from a high to a low pressure. Another method of producing a decrease in temperature was used shortly afterwards by Claude, this being the adiabatic expansion of compressed air with the performance of mechanical work. Modern equipment for the production of liquid air embodies a combination of these two methods.

The range of temperatures between about —80 °C and —180 °C, is at present virtually unexploited; there are few simple machines working within this range.

Naturally, this brief survey is far from complete. In particular, it fails to reflect the great increase in the use of refrigeration ever since the first continuously operated machines were designed. The development of new applications continues and the introduction of new refrigerating techniques can be expected to lead to still further applications. The development described below should be regarded in this light. In proceeding to this description, we shall refer to a series of articles on hot-air engines published in this Review a few years ago [1]). In the first of these it was stated that the hot-air cycle described could also be used for refrigeration, and that very low temperatures could, in fact, be attained in this way. Investigations carried out during recent years have shown that refrigerators based upon this "gas refrigeration cycle" can exhibit particularly favourable properties precisely within the unexploited range between —80 °C and —180 °C. In fact, the gas refrigerating machine reaches this range in one stage from room temperature, thus considerably simplifying the equipment.

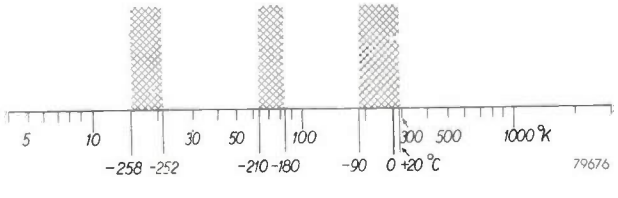


Fig. 1. Temperature ranges covered by conventional refrigerators and liquefiers. Evaporation type refrigerators normally go down to about —80 °C. The equipment used for the liquefaction of air operates in the range —180 °C to —210 °C. Hydrogen liquefiers operate between —252 °C and —258 °C.

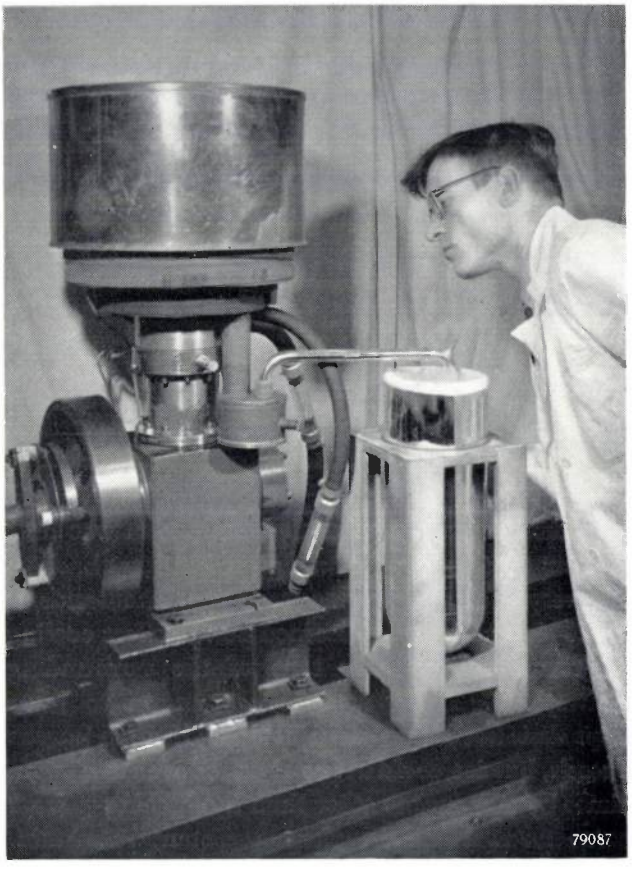


Fig. 2. Gas refrigerating machine used for the liquefaction of air in the Philips Laboratories at Eindhoven.

Machines of this type, capable of generating cold very efficiently at still lower temperatures, e.g. —200 °C, can also be designed. This is demonstrated by the refrigerating machine illustrated in *fig. 2*, constructed in the Philips Laboratories at Eindhoven, which is capable of liquefying air at

[1]) H. Rinia and F. K. du Pré, Air engines, Philips tech. Rev. **8**, 129-160, 1946.
H. de Brey, H. Rinia and F. L. van Weenen, Fundamentals for the development of the Philips air engine, Philips tech. Rev. **9**, 97-124, 1947.
F. L. van Weenen, The construction of the Philips air engine, Philips tech. Rev. **9**, 125-160, 1947.
The hot-air engine is more appropriately named *hot-gas engine*, since the working fluid is not restricted to air.

atmospheric pressure. In this process the air to be liquefied need not be compressed. This fact permits some surprising simplifications, which have led to an extremely compact apparatus requiring very little attention.

In this article the fundamentals of the gas refrigeration cycle will be discussed. The practical limitations of this cycle and the design of the actual machine with be dealt with in a second article.

The gas refrigeration cycle

In gas refrigeration, a quantity of gas is compressed at room temperature and the heat of compression thus generated is dissipated, e.g. by cooling-water. Next, the gas is cooled to the desired low temperature (it is assumed that the steady state has been reached, i.e. that a region of the machine is already cooled to this particular temperature), and it is permitted to expand with the performance of mechanical work. The cold of expansion [2]) thus liberated is the essence of the entire process. Finally, the expanded gas is re-heated to room temperature and another cycle is commenced.

In principle, then, the process may be divided into four phases:

I. Compression of the refrigerant at the ambient temperature.
II. Cooling to the working temperature.
III. Expansion at the working temperature, with a consequent generation of cold (absorption of heat).
IV. Re-heating to the ambient temperature.

The refrigerant, unlike that in evaporation refrigerators, remains gaseous throughout this process. Since a *closed system* is used, the choice of refrigerant is not restricted to air; any other refrigerant may be used, provided that its behaviour approximates sufficiently well to that of a perfect gas.

It will be seen that the gas refrigeration process is somewhat similar to the Claude process mentioned in the introduction. In both processes gas is compressed at room temperature and allowed to expand at the working temperature. The object of the Claude process, however, like that of the Linde process, is to extract the cooled and liquefied gas from the system, whereas in gas refrigeration the gas merely acts as the working fluid (analogous to the liquid in the evaporation refrigerator). Moreover, the practical forms of these two processes differ so considerably that there is little to be gained by comparing them.

[2]) Reference to the generation of "cold" in the qualitative sense has already been made in the earlier part of this article. Following the accepted usage of refrigeration terminology, we shall now apply the same term quantitatively; thus the production of a quantity of cold by a working fluid simply means the absorption of a quantity of heat by the fluid from its surroundings.

The process outlined here was at one time carried out by means of two separate machines, one for compression and one for expansion (in this form it is known in the literature as the air refrigeration process). However, a single machine containing two compartments in communication with each other but at different temperatures can also be used; the refrigerator then contains no valves and is therefore extremely simple. This type of machine is similar to the well-known Stirling air engine. The operation of the Philips gas refrigeration cycle will first be discussed in terms of a schematic cycle in which the four distinct phases of the cycle can be identified. This schematic process is not suitable for practical application, since it involves a discontinuous movement of pistons, but it will be shown in due course that an equivalent cycle can be obtained with a *continuous* piston movement.

Schematic cycle

Fig. 3a shows a cylinder containing two pistons; the one on the right moves in a region of the cylinder which is at high temperature T_C, i.e. at room temperature (about 300 °K), and the one on the left in another region of the cylinder at a low temperature (T_E). Between the two portions is a regenerator R (heavy dotted line), the purpose of which is described in the next section. In phase I of the cycle (the transition $1 \rightarrow 2$ in the diagram) the gas is compressed at temperature T_C (C = compression) from volume V_1 to volume V_2. It is assumed that the compression is isothermal, i.e. that the heat of compression is removed from this

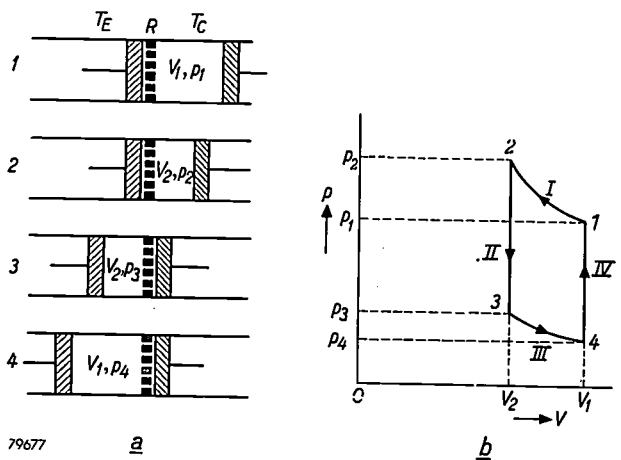


Fig. 3. The schematic cycle of the gas refrigeration process.
a) Two pistons move in a cylinder, one region of which is at room temperature (T_C), another region (the freezing space) being at a low temperature (T_E). A regenerator (R) is situated between these two regions. The process comprises four phases *I-IV*, the initial states (*1-4*) of which are shown in the diagram.
b) Variation in the pressure and volume of the working fluid. Phases *I* and *III* of the cycle are isothermal, phases *II* and *IV* are isochoric (constant volume).

part of the cylinder. In phase *II* of the cycle (*2*→*3*) the gas is transferred without change in volume to the low-temperature (T_E) region of the cyclinder (left); the decrease in temperature of the gas during this phase causes a drop in pressure. In phase *III* (*3*→*4*) the gas expands at temperature T_E (E = expansion) from volume V_2 to the original volume V_1, whereupon the pressure diminishes still further, and heat is absorbed from the surroundings of this region of the cylinder. In the final phase *IV* (*4*→*1*) the gas is transferred isochorically (at constant volume) back to the right hand section of the cylinder which is at room temperature (T_C). During this process the pressure increases and condition *I* is restored. The whole process thus consists of passing the same quantity of gas to and fro between the two temperature regions at T_C and T_E, compressing it in the former and expanding it in the latter. In the p-V diagram shown in fig. 3*b*, the process is defined by the two isotherms *I* and *III* and the two isochores *II* and *IV*. The actual absorption of heat, i.e. production of cold, takes place during phase *III*.

The regenerator

The regenerator mentioned in the preceding section is essential to the process, since if the gas compressed during phase *I* at temperature T_C were to pass direct into the space at temperature T_E, the effective output of cold on expansion would be reduced by the amount necessary to cool the gas from T_C to T_E; in fact this loss would be so great that there would be no real output at all.

In the regenerator, then, heat is extracted from the gas passing from the high-temperature to the low-temperature space, and the gas is thus cooled. The quantity of heat extracted is stored temporarily (that is, it is not discharged into the cold space and does not, therefore, reduce the output of cold). On returning during phase *IV* from the low-temperature to the high-temperature zone, the gas re-absorbs the heat stored in the regenerator; hence it enters, and leaves the high-temperature region of the cylinder at the same temperature.

In view of the essential function of the regenerator it is important to recognize that, theoretically, 100% regeneration is possible. This may be shown as follows. The heat energy transferred from the gas to the regenerator in phase *II*, as well as the heat energy transferred to the gas from the regenerator in phase *IV*, are given by

Energy = mass × temp. difference × specific heat.

Since we are considering a closed system, the mass of the working fluid is the same in both phases. The temperature difference is also the same in each case. Furthermore, the specific heat remains constant although the average pressure is greater during phase *II* than during phase *IV*, for we are here considering perfect gases whose specific heats are independent of pressure. Hence the heat transferred from the gas to the regenerator in phase *II* is equal to that transferred to the gas from the regenerator in phase *IV*.

In practice, of course, the regenerator does not achieve 100 % regeneration, but it nevertheless drastically reduces the loss of cold in phase *II*. For the present we shall proceed on the assumption that the refrigerator considered is provided with an ideal regenerator, i.e. that this loss may be ignored.

Energy balance in the schematic process

It can be shown (see below, small type) that the quantity of cold produced per cycle in the expansion cylinder is given by

$$Q_E = mRT_E \ln \frac{V_1}{V_2}, \quad \ldots\ldots \quad (1)$$

and that the mechanical work which has to be performed per cycle is

$$W = mRT_C \ln \frac{V_1}{V_2} + mRT_E \ln \frac{V_2}{V_1} = \\ = mR(T_C - T_E) \ln \frac{V_1}{V_2}. \quad \ldots \quad (2)$$

In these formulae, R is the gas constant and m the mass (in gram-molecules) of the working fluid.

It will be seen that both Q_E and W are proportional to the logarithm of the volume ratio V_1/V_2, and to the quantity of gas present in the system; W is positive, i.e. there is a net surplus of work to be performed (also shown by the *sense* of the cycle indicated in the p-V diagram, fig. 3*b*). It will also be seen that Q_E is proportional to T_E; hence the output of cold decreases as the temperature of the freezing-space decreases. But the required mechanical work is proportional to the temperature difference $T_C - T_E$; it is thus zero when $T_C = T_E$, and increases steeply according as T_E decreases.

The refrigeration "efficiency", or better, the "coefficient of performance", is the ratio between the cold produced and the mechanical work expended. Applying formulae (1) and (2), we find that this ratio is:

$$\eta = \frac{Q_E}{W} = \frac{T_E}{T_C - T_E}.$$

This corresponds to the thermal efficiency of the Carnot cycle which, according to thermodynamics, is the highest efficiency attainable in a cycle operating between temperatures T_C and T_E. To attain this efficiency the cycle must be reversible; provided

that the assumptions of isothermal compression and expansion and perfect regeneration are valid, the present process satisfies this condition. It is seen that the efficiency decreases sharply as the

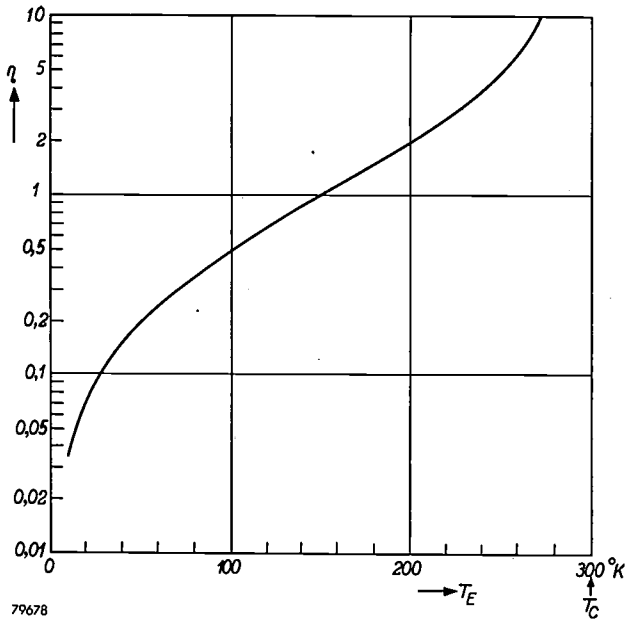


Fig. 4. Theoretical coefficient of performance η of the gas refrigeration cycle (i.e. thermal efficiency of Carnot cycle, $\eta = \frac{T_E}{T_C - T_E}$) plotted against the freezing-temperature T_E, for an ambient temperature $T_C = 300$ °K.

temperature of the cold region of the machine is lowered (*fig. 4*). On the other hand it is seen — and the same applies in principle to all refrigerators — that the efficiency exceeds unity when the temperature difference $T_C - T_E$ is small, and even becomes infinite in the extreme case where $T_C \rightarrow T_E$.

For this reason the term "coefficient of performance" is to be preferred to "efficiency" for the ratio Q_E/W. The latter suggests that a certain percentage of the required mechanical work is converted into the desired product ("cold"), whereas in reality the cold is produced by *transferring* ("pumping") heat from a low-temperature, to a high-temperature reservoir, the mechanical work merely being used to drive the "pump".

Formulae (1) and (2) may be derived briefly as follows. The work done by a piston in compressing a certain quantity of gas in a cylinder at a pressure p is

$$W = -\int p \, dV,$$

where dV is the change in volume. For the mass of gas considered here, we have, from the gas laws:

$$W = -mRT \int \frac{dV}{V}.$$

Thus the work performed by the piston of the compression cylinder in phase *I* of the schematic cycle corresponds to:

$$W_I = -mRT_C \int_{V_1}^{V_2} \frac{dV}{V} = mRT_C \ln \frac{V_1}{V_2},$$

which is converted into heat of compression. Similarly, we have for the expansion cylinder in phase *III*:

$$W_{III} = -mRT_E \ln \frac{V_1}{V_2} \quad \text{and} \quad Q_E = RT_E \ln \frac{V_1}{V_2}.$$

Formulae (1) and (2) follow direct from the above.

It will be noted that the pressure of the working fluid changes also during phases *II* and *IV*; the pressure (in both cylinders) decreases during phase *II* and increases during phase *IV*, with the result that the fluid has to absorb heat from the cylinder walls in the former, and give up heat to the walls in the latter phase. However, since the pressure ratio in phase *II* is equal the pressure ratio in phase *IV*, and since the quantity of heat transferred depends only on the pressure ratio, the heat absorbed by the fluid in each of the two cylinders during phase *II* is equal to the heat given up during phase *IV*; hence the two terms cancel each other out in the energy balance.

Process based on continuous piston movement

As already mentioned, the schematic cycle described above is unsuitable for practical application by reason of the discontinuous piston movement involved.

As can be seen from *fig. 5*, however, an approximation to this cycle can be achieved by a *harmonic* movement of the pistons. It is then necessary to introduce a phase displacement between the movements of the two pistons, such that *the change in volume in the expansion cylinder is given a phase*

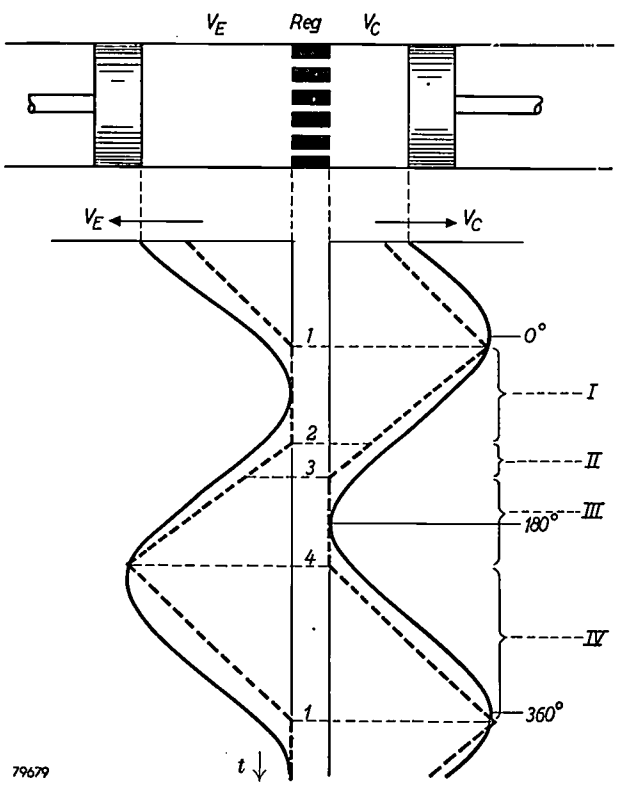


Fig. 5. The gas refrigeration cycle operating with a harmonic piston movement. The variation of V_C (volume of room-temperature region) and V_E (volume of low-temperature region) are plotted against time t. The dotted line represents the corresponding variation in the schematic process (non-harmonic piston movement). This diagram also shows that the regenerator (*Reg*) has a finite volume of its own (see fig. 6) which, together with other spaces not swept out by the pistons, represents a "dead space". This reduces the pressure variation in the refrigerator and thus also the refrigerating capacity, but does not affect the efficiency of the process.

lead with respect to the corresponding volume change in the compression cylinder. Fig. 6 shows that a simple drive can be used to impart the required movement to the two pistons (another method, which has certain advantages over this, will be described in the second article). In the case of a harmonic piston movement, the p-V diagram becomes a smooth continuous curve (*fig. 7*), which implies a merging of the four phases *I-IV*. Compression still takes place mainly — but not entirely — at temperature T_C; similarly, expansion is not confined exclusively to temperature T_E. Moreover, some change of volume now occurs during phases *II* and *IV*, i.e. the isochoric character of the gas transfer is lost. The more important properties of the cycle thus obtained will be discussed in the remainder of this article. A more thorough analysis (which is beyond the scope of this article) shows that the usefulness of the cycle is not lost when a harmonic piston movement is employed.

The fact that the isochoric character of phases *II* and *IV* is not essential may be seen as follows. The cycle may equally well be schematised by a discontinuous piston movement involving an *isobaric* (constant pressure) instead of an isochoric transfer of the gas during phases *II* and *IV*.

The conclusions already reached remain valid also in this case. In many ways, the practical form of the cycle based on a harmonic piston movement may be considered as an intermediate form between that with the isochoric and that with the isobaric phases.

Now that the basic theory of the gas refrigeration cycle has been outlined, a few historical details may be given. Like the hot-air engine, this method of refrigeration originated a long time ago. The possibility of generating cold by the Stirling process was conceived by John Herschel in 1834. A patent for a machine operating on this principle was granted to Alexander Kirk [3]) in 1862; this machine was in use for more than a decade. More recently Lungaard in the U.S.A. worked on a refrigerator based on the Stirling process. It appears that very low temperatures have never been attained with these machines.

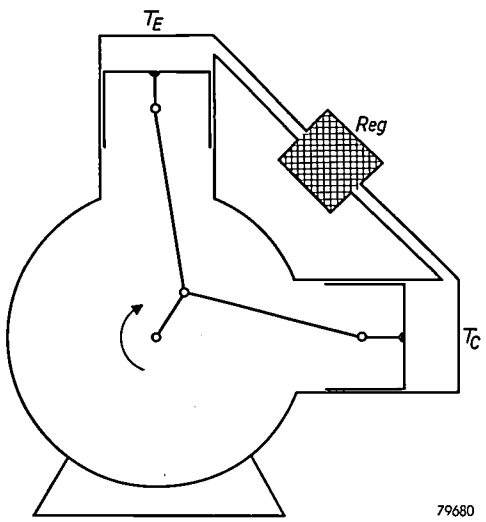


Fig. 6. Crank drive which can be used to produce a harmonic piston movement with the required phase displacement. In the case shown, the phase angle between the volume variations of the compression and expansion spaces is 90°.

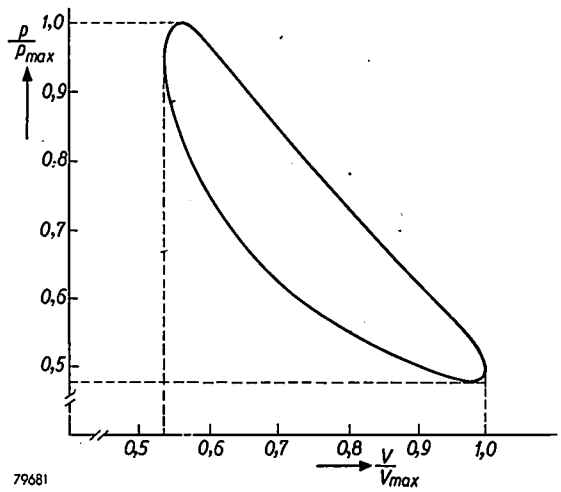


Fig. 7. p-V diagram for the gas refrigeration cycle operating with a harmonic piston movement. The phases *I-IV* of the cycle merge to some extent; hence a smooth envelope is obtained instead of the discontinuous diagram shown in fig. 3*b*.

Relationship between the gas refrigerating machine and the hot gas engine

Comparing the gas refrigerating machine described above, with the hot-gas engine as described in the first of the articles referred to in [1]), their close resemblance in principle will at once be noted. In view of their quite different purposes, however, they should be regarded from different points of view.

The gas refrigeration cycle is frequently referred to as the "hot-air cycle in reverse", but because this phrase is so easily misinterpreted, we shall now proceed to a closer examination of the relationship between the two machines.

Consider first the operation of the hot-gas engine. Its cycle is entirely analogous to that demonstrated in fig. 5, but in the engine the temperature T_E of the expansion cylinder is kept high, e.g. 600 °C; the other cylinder is again at the ambient temperature T_C. In these circumstances, i.e. when heat is supplied to the expansion cylinder at temperature T_E and the other cylinder at T_C is cooled, the engine delivers mechanical power to the shaft. If the heating of the expansion cylinder is discontinued, but the machine is maintened in rotation at the

[3]) A. G. Kirk, Min. Proc. Inst. Civ. Eng. **37**, 244, 1873/74.

same speed and in the same direction (by applying power from outside), the change will not be detected by the interior of the machine, since the working fluid will continue to go through the same cycle as before; hence the fluid will continue to absorb heat from the expansion cylinder. Deprived of its source of heat, this cylinder will therefore cool rapidly until its temperature finally drops below the ambient temperature; the hot-air engine is thus converted into a refrigerator. The decrease in temperature continues until the residual influx of heat (e.g. due to imperfect insulation) to the expansion cylinder, cooled below the ambient temperature, is balanced by the rate of heat absorption (cold output) of the working fluid.

Thus in both the hot-gas engine and the gas-refrigerating machine, heat is absorbed from outside by the working fluid in the expansion cylinder. The principal difference between the two machines lies in the temperature T_E at which this absorption takes place: in the engine T_E is higher than the ambient temperature: in the refrigerator T_E is lower than the ambient temperature. Both machines, however, rotate in the same direction and with a phase difference of the same sign between the two piston movements [4]).

Performance calculations on the cycle based on harmonic piston movement

The calculation of cold output, mechanical power (shaft power) and coefficient of performance, outlined above for the schematic cycle of the Stirling process, may now be made for the cycle involving a harmonic piston movement. From what has already been said it will be clear that the calculation will be substantially the same as that previously described for the hot-air engine, provided that the quantities employed are suitably defined.

The following notation will be used:

V_E, T_E = volume and absolute temperature of the expansion space (referred to in the articles of [1]) as the hot space of the hot-air engine, with suffix h).

V_C, T_C = volume and absolute temperature of the compression space (referred to in the articles of [1]) as the cold space of the hot-air engine, with suffix c).

V_0 = maximum value of V_E.

wV_0 = maximum value of V_C.

τ = T_C/T_E, the temperature ratio of the machine; for the engine $\tau<1$ and for the refrigerator $\tau>1$.

V_s = volume of the dead space, that is the total volume of the circulation system other than that of the cylinders proper.

T_s = absolute temperature of the dead space (average for this space),

s = $\frac{V_s}{V_0}\cdot\frac{T_C}{T_s}$, the relative reduced dead space.

Assuming that the variation of V_E and V_C is harmonic, we write:

$$\left.\begin{aligned} V_E &= \tfrac{1}{2}V_0(1+\cos\alpha) \\ \text{and}\quad V_C &= \tfrac{1}{2}wV_0[1+\cos(\alpha-\varphi)] \end{aligned}\right\} \quad \ldots \quad (3)$$

where

α = the crank angle of the machine, measured from the position associated with maximum expansion space. If the time t is likewise measured from the moment when α is zero, then $\alpha=\omega t$, ω being the angular velocity of the shaft;

φ = the phase difference between the variations in volume of the expansion and compression spaces (φ is positive when V_E leads in phase).

To carry out the calculation it is necessary to establish the variation of the pressure p in the machine as a function of the crank angle α.

Variation of the pressure

The expression for the pressure variations derived previously [1]) simply by applying the gas laws is:

$$p = p_{max}\frac{1-\delta}{1+\delta\cos(\alpha-\Theta)} \quad \ldots \quad (4)$$

where p_{max} is the maximum pressure,

$$\delta = \frac{\sqrt{\tau^2+w^2+2\tau w\cos\varphi}}{\tau+w+2s},$$

Θ is the phase angle of the pressure with respect to the volume of the expansion cylinder (p is at its minimum when $\alpha=\Theta$); hence

$$\tan\Theta = \frac{w\sin\varphi}{\tau+w\cos\varphi}. \quad \ldots \quad (5)$$

Formula (4) applies to both the gas refrigerating machine and the hot-gas engine. The value of the quantity δ, which is only slightly dependent on the temperature ratio, is about the same for machines of both types, i.e. 0.3-0.4. *Accordingly, the pressure variations in both machines are virtually the same; there is, however a consider-*

[4]) We draw attention to this fact, because in the article quoted, the transition from engine to refrigerator is effected by reversing the direction of rotation. This is not quite so simple (especially as regards the calculation given below), since the sign of the heat flux changes with the direction, and the two cylinders thus exchange functions.

able difference between the two as regards the phase Θ of the pressure with respect to the piston movement. This is substantiated by expression (5) for tan Θ; the denominator of this formula contains the ratio τ, which in the engine is less than unity and in the refrigerator greater than unity. Thus Θ is greater for the engine than for the refrigerator. This is also demonstrated by *fig. 8*, which is a polar diagram showing p as a function of the crank angle, in accordance with equation (4); the resultant curve is an ellipse, one of whose foci coincides with the origin. Fig. 8*a* applies to a refrigerator, 8*b* to an engine; it will be seen that there is a difference in the position of the major axis, which corresponds to the difference in phase angle Θ.

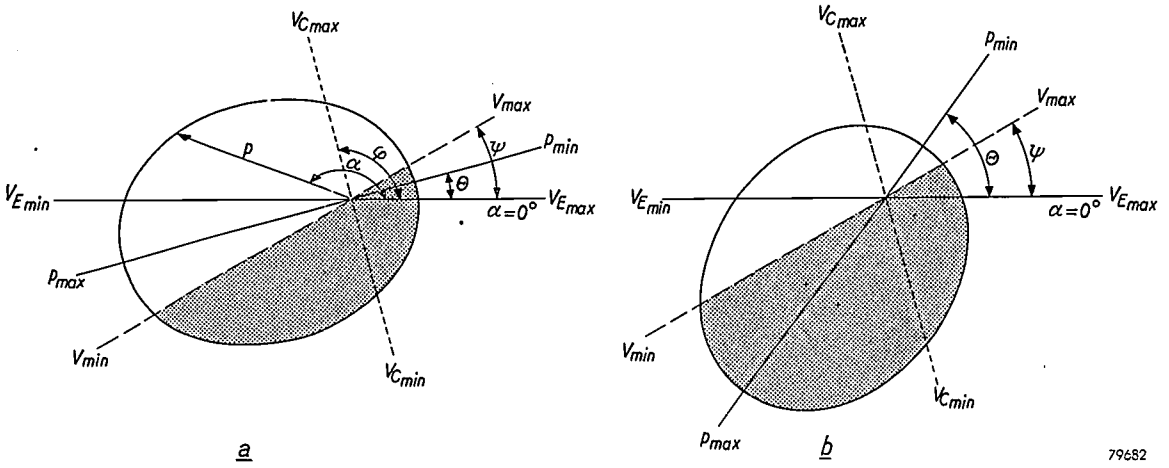


Fig. 8. Polar diagram showing the variation of pressure with the crank angle α,
a) for the gas refrigerating machine ($\tau > 1$),
b) for the hot-gas engine ($\tau < 1$).
Angle α is measured from the position of the crank corresponding to maximum expansion space ($V_E = V_0$). For clarity in the drawing, the pressure ratio p_{max}/p_{min} is given a value appreciably higher than the values employed in practice (2 to 2.5). Usually, angle φ is chosen between 60° and 120°. In case (*a*) the phase angle Θ of the pressure variation is smaller than the phase angle ψ of the variation in the total volume $V(= V_E + V_C)$, whereas in case (*b*) Θ exceeds ψ. As a result, the average pressure in the *refrigerator* is lower during the expansion stroke (from V_{min} to V_{max}, shaded part of ellipse) than during the compression stroke (from V_{max} to V_{min}); hence mechanical power has to be supplied to the refrigerator. The opposite is the case for an *engine* (*b*): this therefore *delivers* mechanical power.

To further clarify the phase relationship, the pressure p defined by equation (4), the volumes V_E and V_C, and the total volume $V_E + V_C = V$ ("working space") of the refrigerator, are shown in *fig. 9* as functions of α on cartesian co-ordinates. It is seen that the pressure variation is almost sinusoidal.

The minimum and maximum pressures occur at $\alpha - \Theta = 0$ and $\alpha - \Theta = \pi$ respectively. Substitution in (4) gives us for the pressure ratio:

$$\frac{p_{max}}{p_{min}} = \frac{1+\delta}{1-\delta}. \quad \ldots\ldots \quad (6a)$$

By integrating the pressure with respect to the crank angle α, we find an expression for the mean pressure:

$$\bar{p} = p_{max}\sqrt{\frac{1-\delta}{1+\delta}}, \quad \ldots\ldots \quad (6b)$$

which will be used in the calculation that follows.

Refrigerating capacity, shaft power and efficiency

Consider now, quite generally, a working fluid in an enclosed space. The quantity of heat absorbed by this fluid when volume (V) and pressure (p) both vary *sinusoidally* can be determined very simply.

This quantity is, per cycle,

$$Q = \oint p\,dV. \quad \ldots\ldots\ldots \quad (7)$$

Now, suppose that

$$V = \tfrac{1}{2}V_0(1+\cos\alpha)$$

and

$$p = \bar{p}\,[1 - \Delta\cos(\alpha - \vartheta)], \quad \ldots \quad (8)$$

where ϑ is the phase lag of the pressure relative to the volume. Substituting the above in (7) we have:

$$Q = \tfrac{1}{2}\bar{p}V_0 \oint [1-\Delta\cos(\alpha-\vartheta)]\sin\alpha\,d\alpha,$$

i.e.

$$Q = \pi\bar{p}V_0\frac{\Delta}{2}\cdot\sin\vartheta \quad \ldots\ldots \quad (9)$$

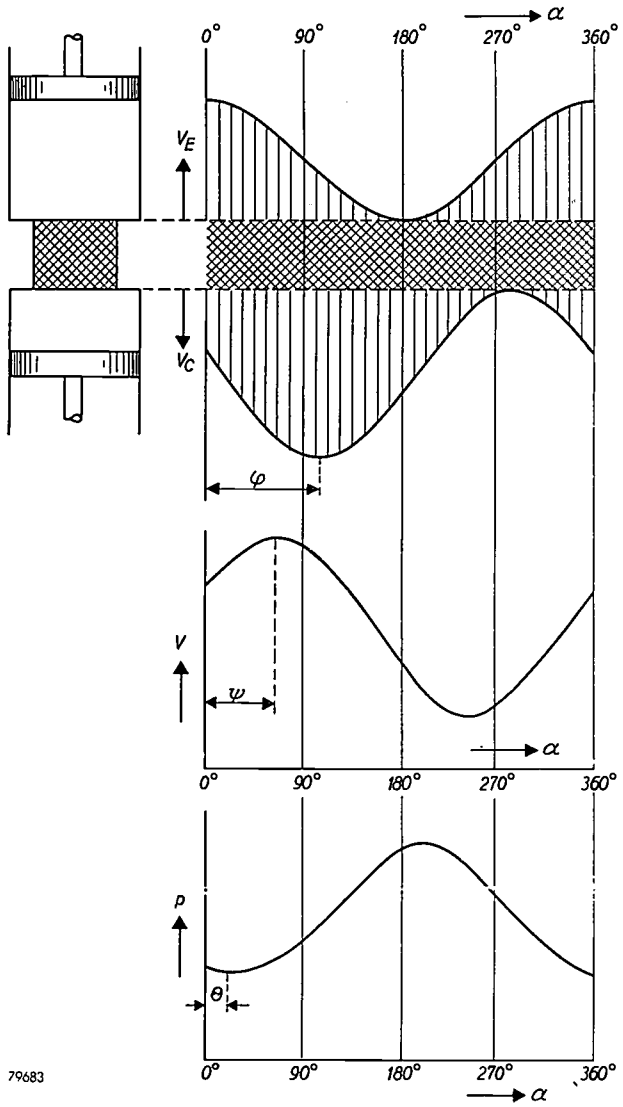


Fig. 9. The variation of the volumes V_E and V_C, of the total volume V ("working space"), and of the pressure p of the refrigerating machine as functions of the crank angle α. The phase relationships are clearly seen; the pressure variation is almost sinusoidal.

It will be seen that Q is positive when angle ϑ is positive, and negative when this angle is negative. This is actually the case in the machine: the pressure lags behind the volume of the expansion space (where the working fluid absorbs heat), whereas it leads the volume of the compression space (where the working fluid liberates heat).

To apply (7) to the cylinders of the gas refrigerating machine, it is necessary to substitute the sinusoidally varying volumes V_E and V_C according to (3), and the non-sinusoidally varying pressure according to (4) [5]. As is well-known, only the fundamental component of (4) will contribute to the integral of the product of (3) and (4). Hence we again obtain (9) in which, as the Fourier series shows, $2\delta/(1+\sqrt{1-\delta^2})$ must be substituted for Δ, and $\vartheta = \Theta$ must be employed for the expansion space and $\vartheta = (\Theta - \varphi)$ for the compression space. We then find the heat absorption per cycle in the expansion space to be:

$$Q_E = \pi \bar{p} V_0 \frac{\delta}{1+\sqrt{1-\delta^2}} \sin \Theta \quad . \quad (10)$$

and that in the compression space:

$$Q_C = \pi \bar{p} w V_0 \frac{\delta}{1+\sqrt{1-\delta^2}} \sin (\Theta - \varphi) . \quad . \quad (11)$$

Using (5), which relates Θ, φ and τ, we deduce from (10) and (11) that:

$$Q_C = -\tau Q_E . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

The output of cold per second (q_E) is given by multiplying Q_E by the number of revolutions per second. Measuring $\bar{p}$ in kg/cm^2, V_0 in cm^3 and the speed of rotation n in r.p.m., we find [6]) that:

$$q_E = 5.136\, \bar{p} V_0 \frac{\delta}{1+\sqrt{1-\delta^2}} \sin \Theta \cdot \frac{n}{1000} \text{ watt.} \quad (13)$$

By analogy with (12), we have $q_C = -\tau q_E$. The power P to be applied to the shaft follows from this and from the equation $P = -q_E - q_C$:

$$P = (\tau - 1)\, q_E . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

Hence the coefficient of performance of the refrigerator is

$$\eta = \frac{q_E}{P} = \frac{1}{\tau - 1} = \frac{T_E}{T_C - T_E} . \quad . \quad (15)$$

This is again the thermal efficiency of the Carnot cycle as already derived in the case of the schematic Stirling process and which was to be expected, since the change to a harmonic piston movement does not affect the reversibility of the cycle.

The refrigerating capacity q_E and the shaft power P given by equation (13) and (14) are shown in *fig. 10* as functions of the temperature T_E of the expansion space, when $T_C = 300$ °K (ambient temperature). This diagram demonstrates once again the close relationship between the hot-gas engine and the gas refrigerating machine. Formulae

[5]) That the expression $Q = \oint p\, dV$, which applies to an enclosed space, is also valid in this case can be demonstrated by means of a thermodynamic argument which has been omitted from this article as being too involved.

[6]) On the continent of Europe, refrigerating capacity is generally expressed in kcal/hour: this necessitates changing the constant 5.136 in formula (13) to 4.418. In England and America the units generally used are Btu/hour; in this case the constant in (13) should be 17.7. The watt, however, is a very convenient unit for the calculation of efficiencies, since the power of electric motors is usually expressed in watts.

(12) and (14) apply to both these machines, the difference being that for the hot-gas engine $\tau = T_C/T_E < 1$ and for the refrigerator $\tau > 1$ (hence the phase angles Θ differ in the two cases; see above).

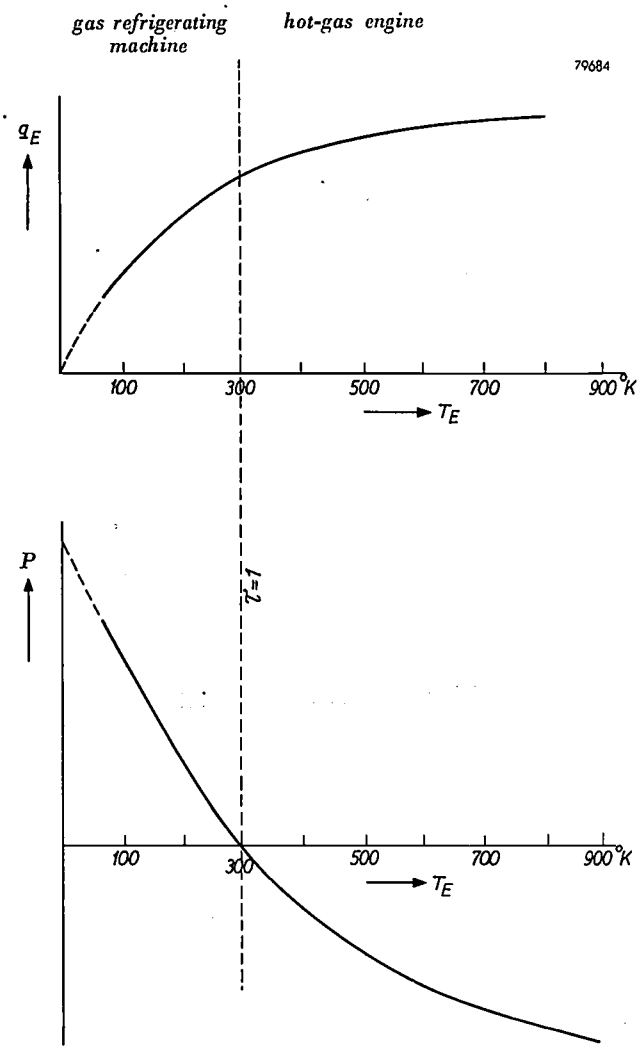


Fig. 10. Refrigerating capacity q_E and applied shaft power P as functions of the temperature T_E of the expansion space (for $T_C = 300$ °K). One continuous curve is obtained, which applies to both the hot-gas engine ($\tau<1$, to the right of T_C) and the gas refrigerating machine ($\tau>1$, to the left of T_C).

It is seen that a decrease in T_E is accompanied by a continuous decrease in q_E, and that the passing of the ambient temperature takes place without any discontinuity. Similarly, the variation of the applied shaft power P is continuous, but changes sign when the expansion cylinder drops below ambient temperature; when this happens the machine changes from a hot-gas engine into a refrigerator.

Fig. 10 illustrates the fact, deduced earlier from considerations of the schematic process, that the efficiency decreases rapidly as the freezing-temperature T_E decreases. This happens because the output of cold diminishes and the required shaft power increases as T_E becomes lower (naturally, a high ambient temperature T_C also has an unfavourable effect on the efficiency). *For economical refrigeration, then, it is essential to ensure that the freezing-temperature at which the machine is to operate* is *not lower than strictly necessary.* Incidentally, this conclusion is valid for every refrigerating process. For example, it is most uneconomical to store cold in the form of liquid air when the objects need be cooled only to —100 °C. This demonstrates the importance of a refrigerator capable of operating economically within the virtually unexploited temperature range between —80 °C and —180 °C. The Philips gas refrigerating machine is particularly well adapted to this role.

This article treats only the *ideal* gas refrigeration process. Naturally, the practical application of this process will involve certain losses. The factors governing these losses will be studied in detail in the second article. It will be shown that although the losses are not to be underestimated, it is none the less possible to design a refrigerating machine of this type which combines high refrigerating capacity with small dimensions and high efficiency. This claim will be substantiated by a detailed description of the gas refrigerating machine illustrated in fig. 2 in this article. A number of machines of this type are now in regular use in the Philips factories for the production of liquid air.

Summary. Although the possibility of generating cold by "reversing" the hot-air engine has been known for more than a century, previous attempts to usefully apply the gas refrigeration cycle have been unsuccessful. The application of principles analogous to those of the Philips hot-gas engine has made possible the construction of a highly satisfactory gas refrigerating machine. A number of machines of this type which cool down to about —200 °C, are at present in regular use in the Philips factory at Eindhoven, where they are used for the liquefaction of air: this low temperature is reached in one stage, which makes for a refrigerator of small dimensions and high efficiency. The gas refrigerating machine is, however, also very suitable for operation throughout the entire range of temperatures between —80 °C and —200 °C, which is just the range not covered by refrigerators of conventional design. In this article, the gas refrigeration cycle is described and analyzed with reference to a schematic cycle involving a discontinuous movement of two pistons in two cylinders. It is demonstrated that this schematic cycle can be replaced by a practical cycle based on a harmonic piston movement. The close relationship between this cycle and that of the hot-gas engine is explained in detail. Finally, the pressure variation for an ideal gas refrigeration cycle (no losses) involving a harmonic piston movement, is deduced by applying the previously derived theory of the hot-gas engine. The refrigerating capacity, required shaft power and coefficient of performance (efficiency) are computed from the pressure variation.

In a subsequent article the losses involved in the practical application of this process will be analyzed, and the construction of a complete refrigerating machine described.

A NEW AUTOMATIC HYSTERESIS CURVE RECORDER

by F. G. BROCKMAN *) and W. G. STENECK *).

621.317.44.087.4

In the development of new magnetic materials, the taking of hysteresis curves is one of the recurring measurements. For this purpose an apparatus has been developed which traces a complete hysteresis loop on paper in about a minute. The scales along both axes of this graph can be set at convenient values. According to requirements either the induction or the magnetization may be plotted against the field strength.

Introduction

Measurement of magnetic quantities by integration

The various curves which relate the magnetic induction B (or alternatively the magnetization J), to the field strength H, in ferromagnetic materials (hysteresis loops, initial magnetization curves, minor loops) contain data of practical as well as of theoretical value. Determinations of these data are therefore of considerable importance. More often than not, the determination of at least one, and frequently of both quantities B and H is accomplished by measuring the electromotive force produced in a coil of wire by a change in the magnetic flux. This electromotive force, e, expressed in volts, obeys the relationship:

$$e = \frac{\mathrm{d}N\Phi}{\mathrm{d}t}$$

where Φ is the flux in webers (Vsec) linked by each of the N turns of the coil. In order that the electrical effect can furnish a measure of the flux change producing it, an integration of the voltage developed must be performed:

$$\int_0^t e\,\mathrm{d}t = \Delta N\Phi \ldots\ldots\ (1)$$

Let us suppose, in order to consider a definite case, that the hysteresis curve of a permanent magnet material is required. The sample, which is in the form of a cylinder, is clamped between the pole pieces of an electromagnet (*fig. 1*). A number of windings are laid closely around the sample, constituting a coil for the measurement of B. A small coil, having many turns, is placed adjacent to and with its axis parallel to the sample. This serves to measure H, the value of which, at the site of this small coil, may be taken as equal to the value inside the sample. (Alternatively, and more exactly, a Rogowski or Chattock potentiometer coil may be used to measure H).

*) Philips Laboratories, Irvington, N. Y., U. S. A.

Each measuring coil is connected to an integrating device. Such a device may take a number of forms, the most widely known of which is probably the ballistic galvanometer. The ballistic galvanometer is a single deflection device (returning to zero after each reading of a change in flux). Consequently, point by point readings of flux changes must be made, which make the procedure very time-consuming. An advantage of the method is its high accuracy.

Other methods of carrying out the integration make use of the fact that the current through an inductor is proportional to the time integral of the voltage drop across it, or that the voltage developed across a capacitor is proportional to the time integral of the current flowing into it.

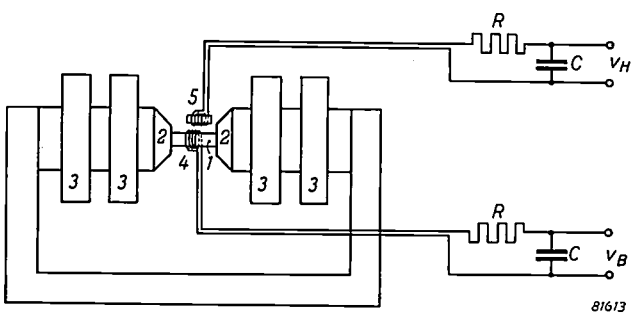


Fig. 1. Principle of hysteresis curve plotting by the yoke method The sample *1* is clamped between the polepieces *2* of the soft iron yoke with magnetizing coils *3*. A coil *4* of a few turns, laid closely round the sample, serves for the measurement of B, and a small coil *5* of many turns for the measurement of H. Both measuring coils are connected to an RC integrating circuit. The voltages v_B and v_H across the capacitors measure changes of the induction and of the field strength in the sample.

Integration with an RC-circuit

The method actually used here is based on the last mentioned principle. Let us for a moment suppose that each of the measuring coils in fig. 1 is connected to a capacitor C via a resistor. If the total resistance in such a circuit is R, the current i and the electromotive force generated in the coil by a changing flux e, the voltage across C being denoted by v, there exists the relation:

$$e = iR + v.$$

By combining this with (1) and supposing that we start with an uncharged capacitor, one easily deduces:

$$\Delta N\Phi = RCv + \int_0^t v\,dt. \quad . \quad . \quad . \quad . \quad (2)$$

It can be arranged that the last term is small compared with the error allowed in the determination of the flux change. In that case we may neglect this term and consider the voltage across the capacitor as a measure for the flux change that has occurred since the start of the experiment.

As the cross-section A_B of the permanent magnet sample and the number of turns N_B around it are known, it is now possible to compute the change in the value of B from the voltage v_B across the corresponding capacitor with the formula:

$$\Delta B = \frac{RC}{N_B A_B} v_B. \quad . \quad . \quad . \quad . \quad . \quad (3)$$

By a similar procedure we may calculate the value of the change in field strength ΔH from the voltage across the capacitor associated with the field-coil. The effective turns-area ($N_H A_H$) of this coil must first be determined by calibrating it in a known magnetic field. When the magnetizing current of the electromagnet is varied, each pair of instantaneous values v_B and v_H provides a point of the required hysteresis curve.

The error on integration

A fuller analysis, not reproduced here, shows that if the magnetic flux is any given function of time, the value of the neglected error term in (2) decreases relative to the main term RCv as the time constant $\tau = RC$ of the integrating network increases. It further appears that in practical cases this error term may be neglected when τ is long compared with the time for the integration. Obviously, however, when the value of τ is increased by increasing R, C, or both, the value of the voltage v itself is decreased, which implies a loss in *sensitivity*. On account of this, and also because it is difficult to obtain large-valued resistors and capacitors that are sufficiently stable, the simple RC integrating circuit turns out to be practicable only for relatively short integration times (e.g. 1/50 second).

RC-integrating circuits, such as outlined above, have been applied to the recording of magnetization curves under alternating current conditions. In the case of the permanent magnet sample we are considering here, rather intense magnetizing fields are required. It is not practicable to produce these fields at A.C. mains frequencies because of the great inductive reactance of the electromagnet and the large eddy currents that would be set up in its unlaminated core. Moreover the eddy currents present in a massive metallic sample would impair the accuracy of the measurements by their shielding effect in the sample. Carrying a laboratory electromagnet of modest size through a cycle from zero field to a large positive field, through zero again to an equally large negative field and back to zero,

Fig. 2. Front of the instrument for the automatic recording of hysteresis curves, showing two integrator panels (one for B and one for H) and the output control panel (the uppermost one). The instrument is used together with an X-Y recorder of a commercial type.

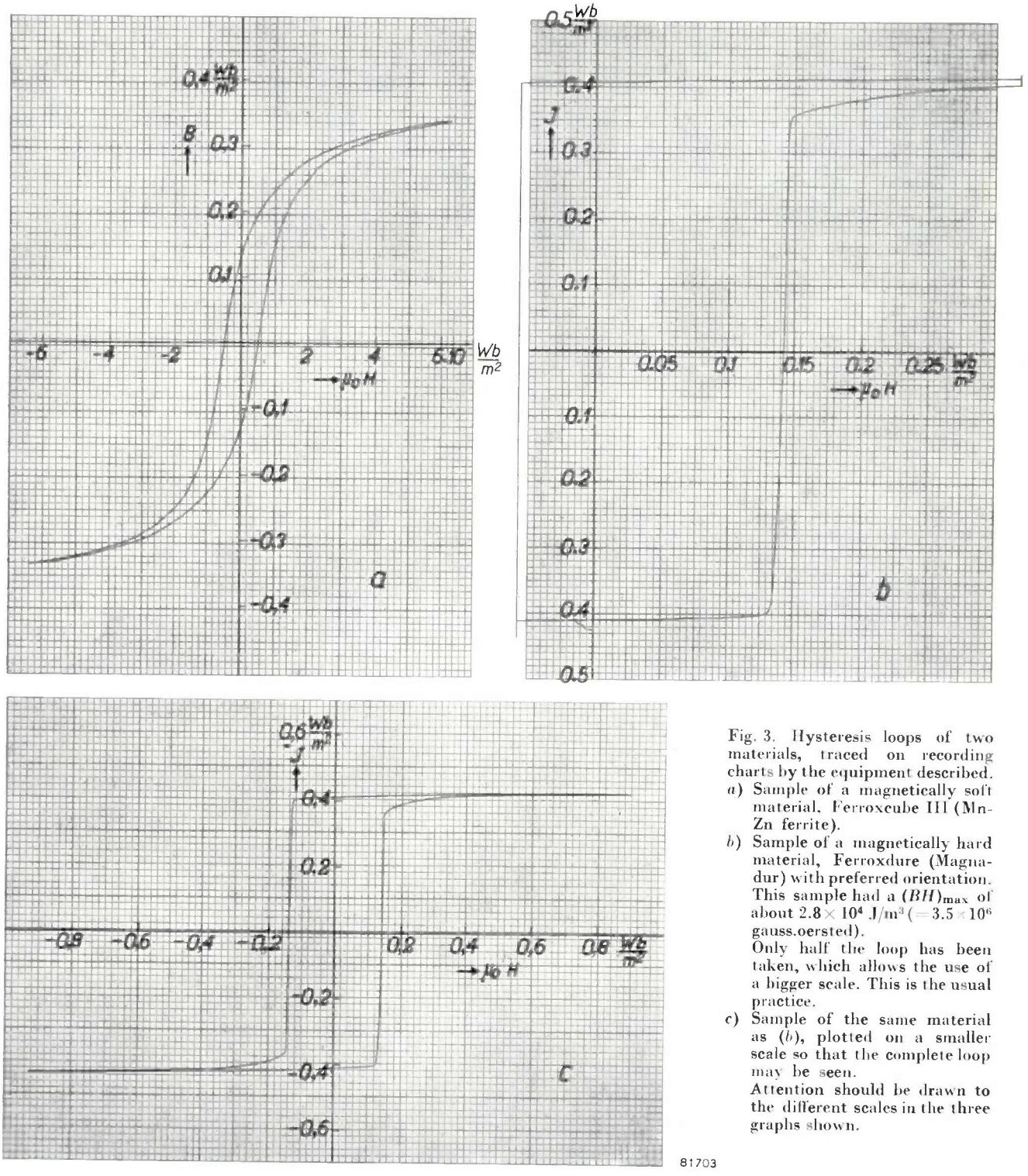


Fig. 3. Hysteresis loops of two materials, traced on recording charts by the equipment described.
a) Sample of a magnetically soft material, Ferroxcube III (Mn-Zn ferrite).
b) Sample of a magnetically hard material, Ferroxdure (Magnadur) with preferred orientation. This sample had a $(BH)_{max}$ of about 2.8×10^4 J/m³ ($= 3.5 \times 10^6$ gauss.oersted).
Only half the loop has been taken, which allows the use of a bigger scale. This is the usual practice.
c) Sample of the same material as (*b*), plotted on a smaller scale so that the complete loop may be seen.
Attention should be drawn to the different scales in the three graphs shown.

would require perhaps one or two minutes. Obviously the *RC*-integrating circuits would fail under these circumstances.

The new equipment for the recording of hysteresis curves

The instrument which is the subject of this article, and which was developed at the Philips Laboratories at Irvington, N.Y., U.S.A., comprises two identical integrating units and an output control panel (*fig. 2*). The integrating units are refinements of the *RC*-circuits discussed above. They make possible the use of the relatively long integration times necessary with permanent magnet materials. The integrators are calibrated as fluxmeters to an accuracy of about 0.1 %. The unit measuring B and that measuring H are used to move the writing pen and the paper respectively of an X-Y recorder. When

the sample is cycled by varying the magnetization current of the electromagnet in a suitable way, the hysteresis curve of the sample is then automatically recorded. In this way a complete hysteresis-loop can be obtained in one or two minutes (*fig. 3*).

If the Rowland ring method of taking the magnetization data is employed, where the current in the primary winding is used as a measure of the field strength, the voltage drop across a series resistor in the primary can serve directly to operate the H coordinate of the recorder. Only one of the integrating units is necessary in this case.

Basic circuit of the integrators

The basic circuit used in the fluxmeters (integrators) is shown in *fig. 4*. In this circuit, where A represents a D.C. amplifier of very high gain, the principle of negative feedback has been used to increase the time available for the integration [1]).

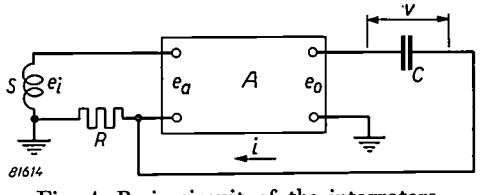


Fig. 4. Basic circuit of the integrators.

For simplicity, let us suppose that the gain G of the amplifier is infinite. Then the input voltage e_a of the amplifier must always be zero and consequently the same applies to the current through the search coil S. The voltage e_i generated in the search coil, will therefore be equal and opposite to the voltage drop across R developed by the output current i:

$$e_i = iR, \quad \ldots\ldots\ldots \quad (4)$$

or, by integration, assuming that we start with an uncharged capacitor:

$$\Delta N\Phi = \int_0^t e_i \, dt = R \int_0^t i \, dt = RCv. \quad . \quad (5)$$

It is seen that the voltage v across the capacitor is proportional to the change $\Delta N\Phi$ in the flux linked by the search coil.

As explained in the appendix, with a *finite* value of G, the circuit acts with the same accuracy as would a resistance-capacitance integrator in which the time constant RC is multiplied by a factor approximately $G/3$. The loss in sensitivity that would occur in an ordinary RC circuit with a large time constant, however, is now completely avoided.

To raise the time constant of a resistance-capacitance combination of convenient value (say $RC = 10^{-3}$ seconds) to a usuable value (say 10^3 seconds), the amplification must accordingly be several millions.

A physical picture of this increase in time constant can be gained in the following manner. Let a charge reside in C, and suppose that there is no e.m.f. in the search coil. Without amplification, the charge flows out of C through R with the usual decay rate associated with the discharge of a capacitor. In the circuit of fig. 4, however, the voltage drop across R gives rise to a voltage across the input of the amplifier, the output of which is fed into the RC network again. Assuming that this output voltage has the correct polarity, it tends to maintain the charge in C. If the gain, G of the amplifier were infinite, the charge in C would remain an infinitely long time.

The amplifier output voltage as a measure of the flux change

Equations (2) and (5) are both valid on the assumption that the charge on the capacitor is equal to the time integral of the current flowing into it. This is true, however, only when the insulation between the capacitor plates is perfect, a condition which is never completely fulfilled. Accordingly equation (2) does not hold exactly and neither does equation (5), not even when G is infinite. It is obvious that the error caused by a non-perfect insulation is much more serious in the case where the time for the integration, and consequently the time for the leakage current to flow, is so much longer. Consequently special care has been given to the insulation in the capacitor. Polystyrene is used as the dielectric and the insulation resistance is in excess of 10^{12} ohms. It is difficult to measure in a simple way the voltage across the capacitor without at the same time seriously reducing this very high insulation resistance. This is why, instead of the voltage v across the capacitor proper, the output voltage e_o of the amplifier is used as a measure of the flux change.

In the case of infinite G there exists between e_0 and v the relation (cf. (4) and (5))

$$e_o = v + iR = \frac{1}{RC} \int_0^t e_i \, dt + e_i. \quad . \quad . \quad (6)$$

In practical cases, the error made by equating e_o to v is negligible (see appendix), and equation (5) may be replaced by:

$$\Delta N\Phi = \int_0^t e_i \, dt = RCe_o, \quad . \quad . \quad . \quad (7)$$

[1]) This circuit using capacitive negative feedback is known as the *Miller integrator* and is commonly used for integration in analog computers, but usually the time constants involved are small compared with the values here. Cf. J. M. L. Janssen and L. Ensing, Philips tech. Rev. **12**, 319-335, 1950/51. R. H. Dicke, Rev. sci. Instr. **19**, 533-534, 1948, has described a capacitative feedback integrator with a long time constant. His method of obtaining D.C. amplification differs from the method used here.

and (3) becomes:

$$\Delta B = \frac{RC}{N_B A_B} e_0 . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

The amplifier

Of the various systems of D.C. amplification it appears that the galvanometer-photoelectric amplifier is the type which has been developed to the state were useful gains of the required order can be realized. This type of amplifier is used in the integrators.

The principle of the photoelectric amplifier may be understood from *fig. 5*. The terminals of the galvanometer *Ga* constitute the input terminals of the amplifier. *F* is a twin phototube whose two photoelectric elements *I* and *II* are

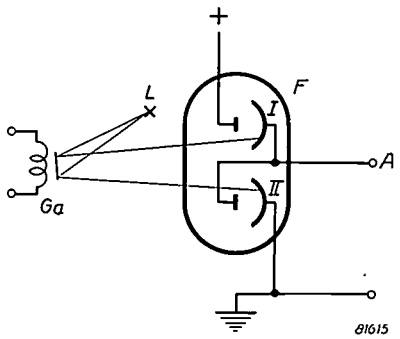


Fig. 5. Illustration of the principle of the photoelectric amplifier

connected in series. Suppose that, when no current is flowing through the galvanometer coil, the light spot from the galvanometer illuminates small but equal parts of each of the photosensitive cathodes. The two elements (supposed identical) then offer the samc resistance to the photo-current and consequently the potential of *A* is half the potential drop across the whole phototube. When a small voltage is established across the galvanometer, a tiny current flows through its coil and the latter is deflected, in such a way, say, that the cathode of element *I* receives more and the cathode of element *II* less light. The internal resistance of *I* then decreases but that of *II* increases. As a consequence the potential of *A* rises sharply. The device thus acts as a voltage amplifier.

By using a galvanometer of medium-high sensitivity and a working distance of 1 metre, it is possible to obtain an amplifier with a voltage gain of about 50×10^6. This amplification is obtained without the use of additional electron tube amplification or positive feedback, and the system thus recommends itself by its obvious simplicity.

Fig. 6 is a simplified circuit diagram of an integrating unit, drawn in such a way as to emphasize its similarity to the basic circuit of fig. 4. A twin phototube *F* of a gas-filled type is used. This gas-filling makes a current limiting resistor necessary in series with the tube. The tube acts as a voltage source with very high internal resistance (> 100 megohms). In order that the output can provide the charging current for the integrating condenser and also operate an indicating instrument, it is necessary to convert this high resistance into one much lower. This is accomplished by a cathode follower vacuum tube circuit.

The meter, shown by $N\Phi$ in the diagram, corresponds to the meters visible in fig. 2. It measures the output voltage of the amplifier and is calibrated directly in units of flux, thus giving a direct reading of changes in the flux linked by the search coil.

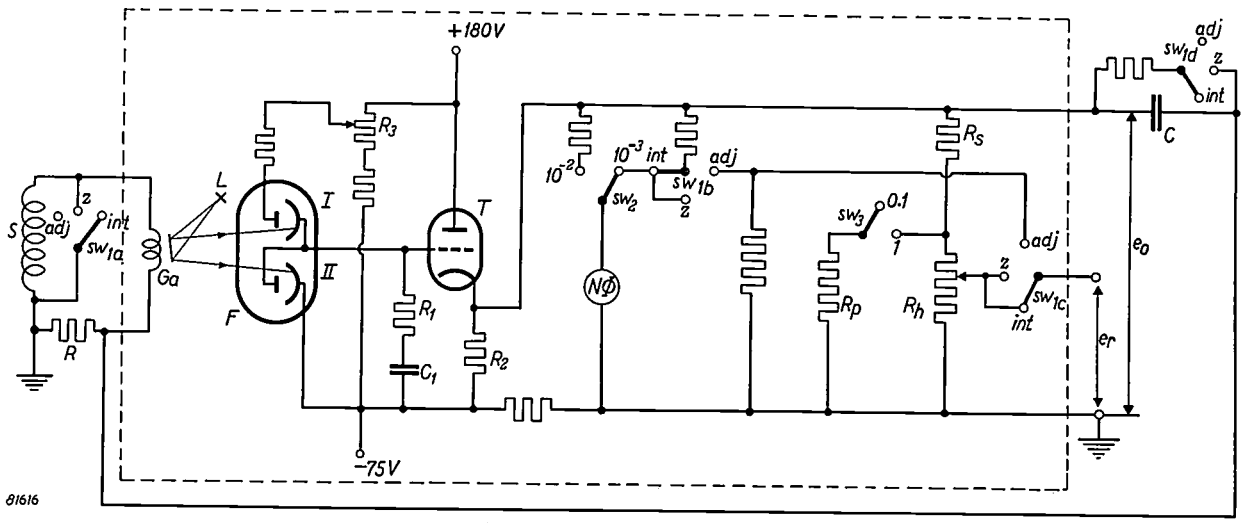


Fig. 6. Simplified diagram of an integrating unit. *S* search coil; *R* integrating resistor; *Ga* gal vanometer; *L* light source for galvanometer; *F* gas-filled twin phototube, containing the two photoelectric elements *I* and *II*; *T* cathode follower tube; $N\Phi$ meter giving a direct reading of the change of flux linked by the search coil. By means of the switch sw_2 this meter can be set to two ranges. R_2 ten turn helical potentiometer, serving as an "output scaler"; R_s and R_p resistors in series and parallel to R_h, serving for the adjustment of the output scaler; sw_3 switch to select the range of the output scaler; *C* integrating capacitors; sw_1 (*a*, *b*, *c* and *d*) multiple switch, which can be set to the "integrating position" (int), to zero (z) and to the "adjusting position" (adj.).

With equal light distribution on the two photocells, it is required that the $N\Phi$ meter indicates zero, which implies that the voltage drop across R_2 has to be 75 volts. In the initial setting-up of the equipment, this requirement is met by adjusting R_3 for zero output voltage when the light beam is equally divided on the photocells. The cathode follower circuit is designed so that under these initial conditions tube T operates linearly.

The resistor R_1 and the capacitor C_1, which connect the grid of T to the -75 V line, serve to provide adequate damping of the system.

The measuring range

With the chosen values of R and C ($R = 1000$ ohms, $C = 10^{-6}$ farad) it follows from equation (7) that

$$N\Phi = 10^{-3}\, e_o.$$

With the small receiving type tubes used in the cathode-follower, a maximum value of 10 volts for e_o is easily attainable. This means that flux changes up to 10^{-2} weber can be measured. By means of the switch sw_2 the measuring range of the indicating meter can be set to 10^{-2} weber or 10^{-3} weber.

This is one of the advantages in the use of the capacitance type integrator as compared with the inductance type. Edgar [2]) described a fluxmeter using a mutual inductance. Although he analyzed his circuit in a different manner, it is readily described as a feedback amplifier with a basic circuit as given in *fig. 7*. When the polarity of e_m is in opposition to that of e_i, and the amplification is large, then approximately $e_i = e_m$.

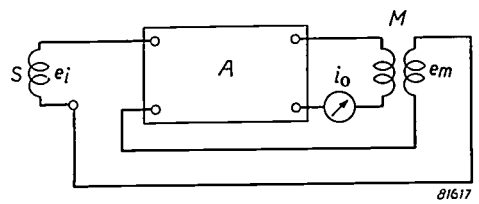


Fig. 7. Basic circuit of the mutual inductance type integrator described by Edgar [2]).

If the input voltage e_i is due to flux changes in a pick-up coil of N turns, then:

$$N\frac{d\Phi}{dt} = M\frac{di_o}{dt},$$

where M is the mutual inductance in henries and i_o is the output current flowing in the primary of the mutual inductor. This gives:

$$N\Phi = Mi_o.$$

If the largest practical value of an air-cored mutual inductor (the use of an iron or ferrite cored inductor would impair the accuracy) be 0.1 henry, then to obtain a maximum range of 10^{-2} weber, a change of output current of 0.1 ampère would be required. In vacuum tube circuits, current changes of this magnitude can be obtained only by the use of power tubes. Cioffi [3]) has described an adaptation of the Edgar fluxmeter for use with a pen X-Y recorder.

The output-"scaler"

It is highly desirable that in the recorded curves the scale factors for both coordinates are round figures. Each integrator is therefore provided with an "output scaler", permitting, as explained below, a free choice of these scale factors.

Such an output scaler is in fact a rather simple voltage divider connected across the output terminals of each of the integrators. It consists of a ten turn helical potentiometer R_h in series with a fixed resistance R_s (fig. 6). The recorder is of the self-balancing type (i.e. the input voltage is continuously balanced by an opposing voltage established automatically inside the recorder [4])): the current in its input circuit is therefore zero and its input voltage e_r is exactly proportional to the resistance tapped off. The series resistance R_s is included because in practical cases the output voltages of the integrator are much too large to directly operate the recorder. R_s is so chosen that the voltage e_r at the recorder terminals is variable from 0 to 0.1 of the integrator output voltage e_o. This range can be reduced by means of the switch sw_3 whereby a resistance R_p is connected in parallel with the potentiometer. The value of R_p is such that e_r is now variable from 0 to 0.01 e_o.

The helical potentiometer has effectively 1000 divisions so that, when S is the dial reading, there exists between e_r and e_o the relation:

$$e_r = \frac{S}{1000} \cdot m \cdot e_o, \quad \ldots\ldots \quad (9)$$

where m is a factor, either 0.1 or 0.01 chosen by the switch.

The X-Y recorder used gives full scale deflections (100 divisions) on both axes at input voltages of 10^{-2} volts (i.e. 10^{-4} volts per division). A deflection of y divisions will accordingly correspond to:

$$e_r = 10^{-4}\, y \text{ volts}. \quad \ldots\ldots \quad (10)$$

From (8) we find, using (9) and (10);

$$\Delta B = \frac{1}{S}\,\frac{1000\,RC}{m}\,\frac{1}{N_B A_B}\,10^{-4}\, y.$$

[2]) R. F. Edgar, A new photo-electric hysteresigraph, Trans. Amer. Inst. El. Engrs. **56**, 805-809, 1937.

[3]) P. P. Cioffi, A recording fluxmeter of high accuracy and sensitivity, Rev. sci. Instr. **21**, 624-668, 1950.

[4]) A description of such a system may be found in H. J. Roosdorp, An automatic potentiometer for industrial use, Philips tech. Rev. **15**, 189-198, 1953/54 (No. 7).

The scale factor for the movement connected to the B-integrator is thus

$$1 \text{ division} = \frac{1}{S}\frac{1000\,RC}{m}\frac{1}{N_B A_B}10^{-4} \text{ webers/m}^2. \quad (11)$$

R and C are known, and so is the quantity $N_B A_B$, though this value may differ for different experiments. It is seen from (11) that the dial reading S of the potentiometers can be adjusted to give the scale factor along the B-axis a convenient value. The setting of the scale along the H-axis is accomplished in an analogous way.

If we want full scale deflection (100 divisions) to correspond to a value ΔB_1 say, of the change in the induction, 1 division must correspond to $\Delta B_1/100$ webers/m^2. From (11) it then follows that the dial of the potentiometer should be set to:

$$S = \frac{1}{N_B A_B \Delta B_1}\frac{10\,RC}{m} \quad \ldots\ldots\ldots \quad (12)$$

If R and C had exactly their nominal values of 10^3 ohms and 10^{-6} farads respectively, $10\,RC/m$ would be either 10^{-1} or 1, depending on the position of the switch sw_3. Now R is made equal to its nominal value to an accuracy of 0.1%, but C only to 0.5%. This precision is adequate for the indicating meter $N\Phi$. A closer accuracy, however, is desirable at the recorder. A measurement of C is therefore made to an accuracy of 0.1% (and as a check the same is done for R). In order not to complicate the calculation of the required potentiometer dial setting, R_s and R_p are readjusted to make m deviate from 0.1 or 0.01, as the case may be, to correct for the deviation of the RC product. In this manner $10\,RC/m$ is made equal to 10^{-1} or 1 respectively to an accuracy of about 0.1%. As the helical potentiometer also is accurate to 0.1%, the same order of accuracy is obtained in the voltages applied to the recorder. A commercially available recorder accurate to 0.1% is used and the cumulative errors (see also appendix) then allow a normal accuracy in the recorded curves of about 0.5% of the maximum reading.

Influence of thermo-electric voltages

The input circuit of an integrator unit involves the galvanometer, the search coil and the resistor R. Much care has been taken to reduce the possibility of the generation of thermal e.m.fs in this loop. The precautions include the use of all copper conductors, massive copper terminals and the inclusion of the resistor R (the only non-copper conductor in the circuit) in a thick walled aluminium box. This last measure serves to keep both ends of R at the same temperature. Even with these precautions a means of balancing the last traces of thermal e.m.fs is required; this balancing is done by a network (not shown in fig. 6) which makes it possible to insert a small controllable and reversible current through R. The voltage drop in R can be set to oppose any thermal forces in the amplifier input circuit.

An impression of the necessity for such a control can be gained by the following: when the fluxmeter is set to the highest sensitivity and is used with a recorder giving full scale deflection for 10 mV, a constant input voltage at the integrator of 10^{-7} V will cause the recorder to deflect 10% of full scale in 100 seconds.

Thermal e.m.fs are a real source of trouble in any flux measuring system. In addition to true thermo-electric forces, there is another effect which might be called an "effective" thermal e.m.f. With the precautions taken to eliminate true e.m.fs this "effective" e.m.f. can be observed. It is the shift of the mechanical zero of the galvanometer with temperature. A shift of the mechanical zero has the same effect as a current flowing in the galvanometer coil and therefore as an e.m.f. in the input circuit. Shielding of the galvanometers from temperature fluctuations aids in keeping this effect at a minimum.

The multiple switch sw_1 (a, b, c, d) serves three purposes. In the centre position, the integrating capacitor C is shorted through a high resistance. This resets the fluxmeter to zero output voltage. By switching to the "integrate " position the unit is ready to act as a fluxmeter. The third switch position is used to adjust the thermal e.m.f. balancing current. In this position the unit is also in "integrate" condition, but the sensitivity at the recorder terminals is twice as great as in the most sensitive operating condition. Drift compensation is made under this enhanced sensitivity.

Recording the magnetization

Since the magnetization is given by $J = B - \mu_0 H$, and the two voltage outputs to the recorder are directly proportional to B and $\mu_0 H$, these may be subtracted electrically to give J. The one requirement is that the output scalers of the two units must be set so that the factors relating their output voltages to B and $\mu_0 H$ respectively, are the same.

This requirement is a disadvantage, because it prevents us from choosing the most convenient full scale deflections for both movements of the recorder. The magnetization can, however, be recorded by a somewhat different method, whereby this disadvantage is avoided. This method comes down to the performance of the necessary subtraction on the input rather than on the output side of the integrators. If the effective turns-area products of the B and the H coil were the same ($N_B A_B = N_H A_H$), the desired effect could be obtained by simply subtracting the voltage generated in the H coil from that generated in the B coil, and feeding the resultant voltage into the B (now changed to J) integrator. In practice, the voltage generated in the H coil must first be reduced to the effective turns-area product $N_B A_B$ of the B coil, which means that only a fraction $N_B A_B / N_H A_H$ of this voltage must be subtracted. The required fraction is obtained by the insertion of a voltage dividing potentiometer across the

terminals of the H coil. (*fig. 8*) It is one of the advantages of the large amplification of the feedback amplifier that its effective input resistance is so large that use can be made of a potentiometer of a relativvely low resistance, accurate potentiometers of very high resistance not being available. In this case a

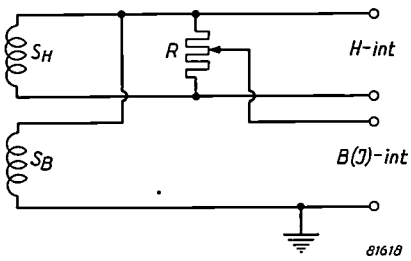


Fig. 8. Input-system for the integrators, suitable for the recording of the magnetization J as a function of the field strength. Correction for the wire size of the B coil can also be made with this circuit. S_H search coil for the measurement of H; S_B search coil for the measurement of B or J; R potentiometer.

1000 ohm potentiometer is used. If the electrical resistance of the H coil exceeds 1 ohm, this must be taken into account in the adjustment of the potentiometer, if an accuracy of 0.1% in the voltage division is to be maintained.

It is necessary that the B coil is so wound that its effective turns-area product is less than that of the H coil. Otherwise a fraction greater than unity of the H coil voltage would have to be tapped off the potentiometer, which is of course impossible.

The potentiometer used is a ten-turn helical potentiometer. These components use dissimilar metals for the resistance wire and for the contacts. As has already been pointed out, thermal e.m.fs. in the input circuit are very undesirable and extra care must therefore be exercised to prevent their occurrence. For this reason the potentiometer is contained within a double thermal enclosure. The inner enclosure is of thick-walled copper and the potentiometer is immersed in oil. In this manner temperature differences in the non-copper portions of the circuit are kept to a minimum.

Correcting for the wire size of the B coil

It follows from theory that the total area enclosed by a turn of wire is that enclosed by its centre line. Thus, even when the B coil is laid in intimate contact with the sample, the flux through the turns is made up of two parts, 1) the flux in the sample and 2) the flux passing between the sample and the centreline of the wire of the coil. The second contribution causes an error in the observed induction B. This can be automatically corrected when use is made of the potentiometer across the H coil mentioned above. If the active area of the flux in the annulus between the sample and the centre line of the wire, be s, then the effective turns-area product is $N_B{\cdot}s$ (N_B is the number of turns of the B coil). Now in the H coil, an effective turns-area $N_H A_H$ corresponds to the total voltage generated in this coil, so that an effective area of $N_B{\cdot}s$ corresponds to the fraction $N_B{\cdot}s/N_H A_H$ of this voltage. If by means of the potentiometer we subtract this fraction of the H coil voltage from the voltage generated in the B coil, the correction for the wire-size of the B coil is automatically made.

Of course the same correction can be made when the magnetization J is plotted. We then have an effective turns-area of $N_B(A_B + s)$ for the H-flux through the B coil and consequently have to subtract a fraction $N_B(A_B + s)/N_H A_H$ of the voltage generated in the H coil.

Appendix: The influence of finite amplifier gain

Equation (5) in the text was developed under the assumption that the gain G of the amplifier was infinite. The circuit conditions when G is finite are given in this appendix. The input resistance of the amplifier and the resistance of the search coil are also included in this analysis. Fig. 4 is redrawn as *fig. 9*. The circuit equations are:

$$\begin{aligned} e_a &= r_i i_1 \\ e_o &= G\, e_a \\ v &= e_o - R i_1 - R i_2 \\ e_i &= (r_s + r_i + R) i_1 + R i_2 \\ \int_0^t i_2 dt &= C\, v \end{aligned}$$

Of the six variables (e_o, e_a, i_1, i_2, e_i and v) in these five equations, the first four can be eliminated, leaving a relation

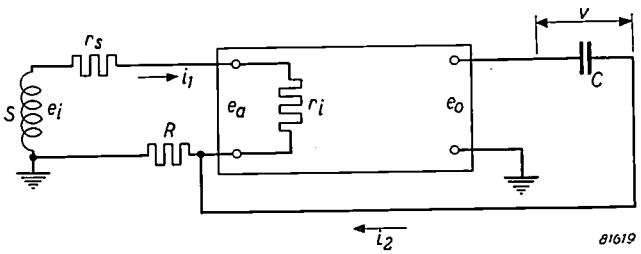


Fig. 9. Basic circuit of integrating unit for the deduction of the circuit equations.

between e_i and v. This relation may conveniently be written in the form

$$\int_0^t e_i dt = RCv + \frac{r_s + r_i + R}{G r_i - R}\left[RCV + \int_0^t v dt\right]. \qquad (13)$$

This result may be compared to the equivalent relation in the case of the simple RC-integrating circuit. There we had:

$$\int_0^t e_i dt = RCv + \int_0^t v dt \;.\;.\;.\;.\;.\;.\;. \qquad (14)$$

In our case $R = 1000$ ohms and $r_i = 500$ ohms, r_s may be

a few ohms and thus its influence is only small. If $G = 50 \times 10^6$, then:

$$\frac{r_s + r_i + R}{Gr_i - R} \approx \frac{1}{G/3} = 6 \times 10^{-8}.$$

From this it follows at once that the first correction term in (13) is negligible as compared with the main term. The second correction term is identical to that in (14), except that in (13) it is preceded by a factor of the order 10^{-7}. It follows that the circuit acts as an RC-integrator in which the error on integration is reduced by a factor of about $G/3$ as was stated in the text.

The time constant

If equation (13) is solved for v under the condition that $e_i = 0$, then the decay of the voltage v across the capacitor, when left to itself, can be obtained. This solution shows that v will decay exponentially according to

$$\frac{v(t)}{v(0)} = \exp - \left\{ \frac{1}{RC} \frac{r_s + r_i + R}{(G+1)r_i + r_s} t \right\}.$$

If the time constant τ is defined as the time required for v to fall to $1/e$ of its initial value,

$$\tau = RC \frac{(G+1)r_i + r_r}{r_s + r_i + R}. \quad \ldots\ldots \quad (15)$$

With the values mentioned for r_i, r_s, R and G, we find $\tau = \frac{1}{3} G \cdot RC = 17 \times 10^6\, RC$. Furthermore, since the RC product used is 10^{-3}, the time-constant is about 1.7×10^4 sec $\approx$ 4.7 hours.

The error on integration

In practice not the voltage v across the capacitor, but the output voltage e_o of the amplifier is used as a measure of the time integral of the input voltage e_i. Instead of the relation between v and e_i we can also deduce from the circuit equations a relation between e_o and e_i. Using the expression (15), this relation turns out to be:

$$e_o = \frac{1}{RC} \int_0^t e_i dt + e_i - \frac{1}{\tau} \int_0^t e_o dt. \quad \ldots\ldots \quad (16)$$

(If G tends to infinity τ does the same and (15) reduces to (6)). Equation (16) shows that the output voltage is not exactly proportional to the time integral of the input voltage, but that two additional terms exist in the solution of the circuit equations. An evaluation of the effect of these terms upon the accuracy of the basic assumption is therefore appropriate. It will not be surprising that certain restrictions exist as regards the time over which the integrations (and hence flux measurements) are made. Although it is not immediately self-evident from (16), a closer comparison of the different terms in this equation shows that the second term on the right ($+e_i$) is associated with a short time limit, and the third term with a long time limit to the integration operation. These two limits are considered separately. In either case it is necessary to define the manner in which the flux in the measuring coil changes in time before an evaluation of these time limits can be made. Both cases have been considered for

a) a flux increasing linearly with time by a fixed amount.
b) a flux increasing according to the square of time by a fixed amount.
c) a flux changing sinusoidally for one cycle, with a fixed maximum value.

The long time limit. — For the analysis of this case, the short time factor in (16) may be ignored and this equation becomes:

$$e_o = \frac{1}{RC} \int_0^t e_i dt - \frac{1}{\tau} \int_0^t e_o dt.$$

It is evident that the determining factor in the magnitude of the error term is the time constant τ. If this were infinitely large the error would be zero. It is possible to show that in first approximation the error is proportional to the time used in the measuring operation (the integrating time). With the derived value of $\tau (\approx 280$ min) this approximation is still valid with integrating times of a few minutes. Furthermore, the error will be less than 1% if the integrating time is less than 4 minutes in any of the three flux change functions considered. The sinusoidal variation is perhaps more analogous to the variations experienced in flux measurements. For this case the peak reading is accurate to better than 1% if the time for the cycle is less than 8 minutes.

The short time limit. — When the time of observation is small, then the long time factor term in (16) may be ignored and this equation may be written:

$$e_o = \frac{1}{RC} \int_0^t e_i dt + e_i.$$

At the end of an integration period, e_i becomes zero and therefore the error is zero.

However, if at any time during an integration, the output voltage at that instant is taken as the instantaneous time integral of the input voltage, this error term exists. Evaluation of the magnitude of this term for the three flux change functions considered shows that the error is less than 0.1% when the time of integration is greater than a few seconds. Because the equipment is used with electro-mechanical recorders with response times of several seconds, this limitation is never important: the operations must in any case be performed so that the changes in the output voltages are slow enough for the mechanical system of the recorder to follow them.

Summary. The equipment described serves for the automatic tracing of hysteresis curves. As usual, B and H are measured by the integration of the electromotive forces which are generated in suitably placed coils when the field is changed. The instrument contains two identical integrators, one for B and one for H, providing output voltages which are the time integrals of their input voltages. These output voltages operate the movements of the paper and the pen, respectively, of a self-balancing X-Y-recorder, so that B is plotted automatically as a function of H. The integrators make use of the fact that the voltage across a capacitor is proportional to the time integral of the current flowing into it. The simple RC integrating circuit, consisting of a capacitor, connected via a resistor to the e.m.f. to be integrated, has the disadvantage that it can only be used when the integration time is short compared with the value of the RC product. For practical reasons the value of this RC product is restricted, so that for integration periods longer than about 1/50 second the simple RC-circuit is not suitable. The integrators actually used avoid this difficulty and may be used for the integration times of one or two minutes which occur in practice when investigating non-laminated samples or samples with high coercivity.

The accuracy of the curves obtained is usually better than 0.5%. The scales along the B and the H axes can be set at convenient values at the beginning of the measurement. It is also possible to obtain directly the magnetization instead of the induction as a function of H. One of the features of the equipment is the employment of an novel input device, making it possible to plot this magnetization, while maintaining free and independent choice of the scales along *both* axes. The same device permits automatic correction for the wire size of the B coil.

A "PHOTOFLUX" FLASH-BULB WITH SIMPLIFIED CAP

by R. WESTRA.

771.448.4

The development of the flash lamp has been especially characterized in recent years by a large increase in the specific light output (i.e. the light output in visual lumen-seconds divided by the bulb volume) and by a reduction in price [1) 2)]. A particularly remarkable increase in the specific light output was obtained by the introduction of the "Photoflux" bulb type PF 3, "the smallest flash-bulb in the world", which was put in the market in 1952. This miniature lamp has a maximum diameter of 22 mm and an overall length of 50 mm (*fig. 1c*; in fig. 1*a* and *b* older "Photoflux" lamps have been given for comparison). The PF 3 emits about 5500 lumen-seconds, a quantity of light which is amply sufficient for most amateur purposes. The manufacture of the lamp was made possible by adoption of the manufacturing techniques used for very small incandescent lamps, such as torch bulbs.

Early in 1954 another small flash-bulb made its appearance in America, with a still smaller diameter (20 mm) and a light output which is correspondingly lower (4000 - 5000 lm sec), see fig. 1*d*. Instead of the bayonet lamp cap of 15 mm diameter (BA 15s, see fig. 1*b* and *c*) generally used until then, this lamp has a cap of only 9 mm diameter, provided with a groove by which it is kept in the lampholder. The purpose of the introduction of this smaller and simpler lamp cap was undoubtedly to cheapen its manufacture and to contribute to the popularity of the flash-bulb by way of a correspondingly reduced selling price.

Philips have also gone into the question of further reducing manufacturing costs. In order to analyze the problem and decide which parts merit first consideration with regard to possible simplification, we will enumerate the three main components of which the flash lamp consists:

1) The light-emitting part: a loosely-knit ball of aluminium-magnesium wire of such a length and diameter that on burning in an oxygen atmosphere the wire emits the quantity of light required.
2) The transparent sealed envelope, in the form of a glass bulb which renders the working of the lamp harmless and odourless, and isolates the oxygen filling from the air.

1) G. D. Rieck and L. H. Verbeek, The "Photoflux" series of flash-bulbs, Philips tech. Rev. **12**, 185-192, 1950/51.
2) L. H. Verbeek, The specific light output of "Photoflux" flash-bulbs, Philips tech. Rev. **15**, 317-321, 1953/54.

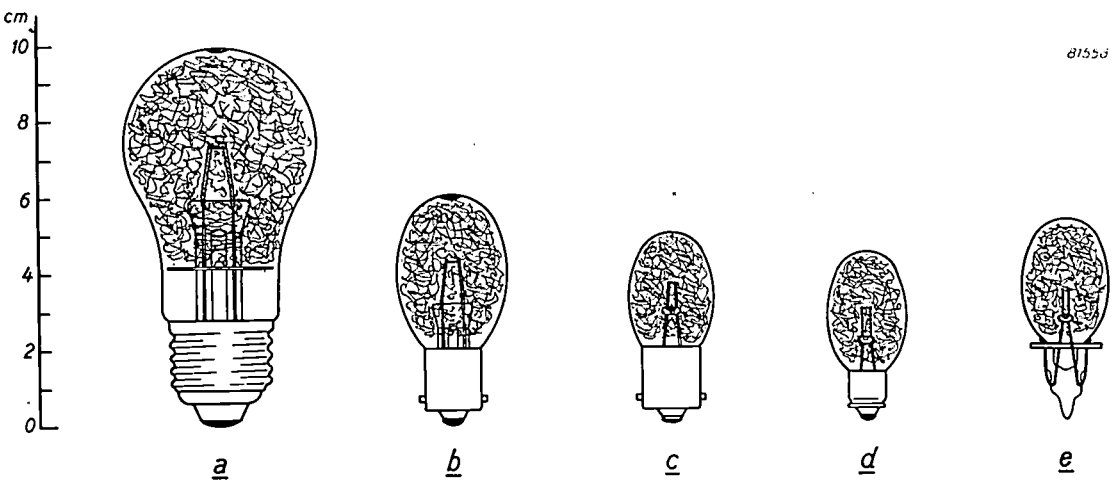


Fig. 1. Different types of flash-bulbs.
a) Lamp in which the lead-in wires for the ignition current are sealed in a "pinch", which also contains the exhaust tube. This construction is now used only for lamps emitting a very great quantity of light, such as the "Photoflux" lamps PF 60 and PF 100.
b) Smaller version of (*a*) with bayonet cap BA 15s. In 1952 Philips replaced this type by the type (*c*). The lamps represented in (*c*), (*d*) and (*e*) are of the "bead-mount" construction (the lead-in wires go through a glass bead).
c) "Photoflux" lamp PF 3 with lamp cap BA 15 s.
d) American flash lamp, with smaller cap.
e) "Photoflux" lamp PF 1, with *glass* cap.

3) The electric ignition mechanism: this consists of a paste applied to a tungsten wire which is provided with two lead-in wires; the lead-in wires are soldered outside the bulb to a metal lamp cap which is cemented onto the bulb. When the tungsten wire is brought to incandescence by means of an electric current, the paste ignites and fires the aluminium-magnesium wire.

A considerable saving was realized by mounting the ignition mechanism, not on a "pinch mount" (fig. 1*a* and *b*) but on a "bead mount" (fig. 1*c*), in which the lead-in wires are kept spaced by a glass bead (see the article referred to in footnote [2]). Further saving must be sought, in the first place, in the non-essential parts of the ignition mechanism. The new American construction (fig. 1*d*), with the smaller cap, already shows a tendency in this direction. Philips have looked for an even more radical step however, and the replacement of the metal cap by a (cheaper) glass construction was therefore considered.

This has, indeed, proved possible, by giving the neck of the bulb such a shape that it can take over the two functions of the metal cap, namely to hold the lamp in the reflector and to make contact between the lead-in wires and the ignition current battery. Fig. 1*e* and *fig. 2* show how this has been achieved in the new "Photoflux" lamp, type PF 1.

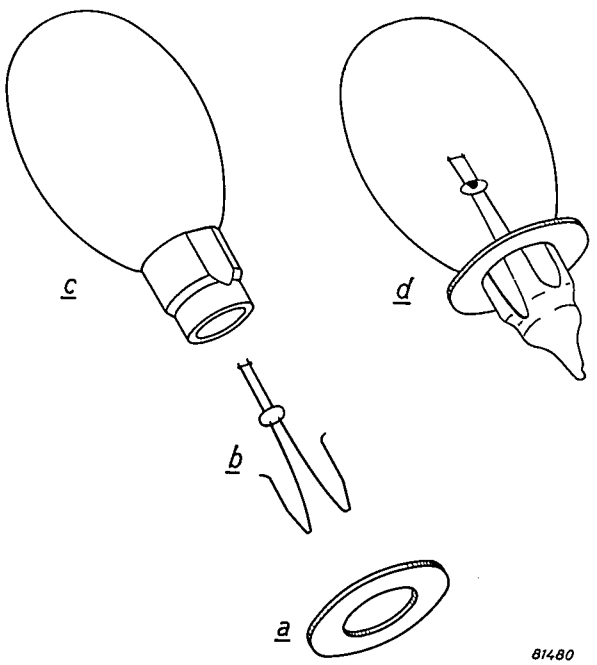


Fig. 2. In the "Photoflux" flash-bulb PF 1, the lead-in wires of the bead mount (*b*) are laid back along the neck of the bulb (*c*), and retained by an insulating ring (*a*). (*d*) The PF 1 complete.

The ends of the lead-in wires outside the bulb are laid back over the neck and are kept tightly in place with a flat ring of insulating material (*a* in fig. 2). The type number of the flash-bulb is marked on this ring. The glass neck is flattened on two sides, so that the lamp is uniquely located in the holder (*fig. 3*), i.e. it can be inserted only in such

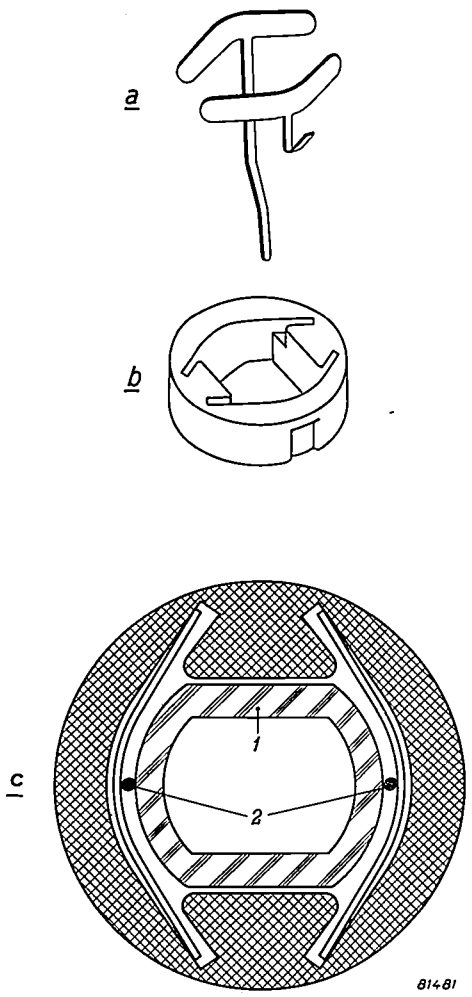


Fig. 3. Lampholder for the "Photoflux" flash-bulb PF 1. *a* flat springs, *b* body of insulating material, *c* holder with lamp inserted (*1* glass neck, *2* current supply wires).

a way that the lead-in wires make contact with flat springs mounted in the holder. On inserting the lamp, the wires scrape along the springs, thereby ensuring good contact. The springs, moreover, secure the lamp in the holder.

The change in construction described is not at the cost of the luminous output of the flash-bulb: on the contrary the PF 1 has an even greater specific light output than the PF 3.

The introduction of a new fitting necessarily has the drawback that an adaptor must be used until flash equipment is available to take the new design.

The adaptor represented in fig. 3 is small enough to be fitted into a BA 15s bayonet cap (*fig. 4*) for which most flash equipment is at present constructed.

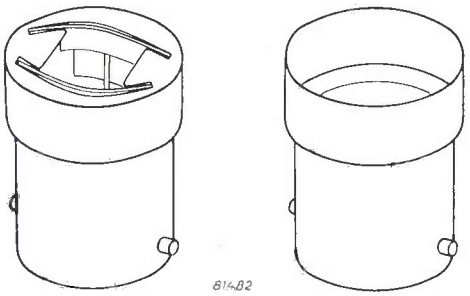
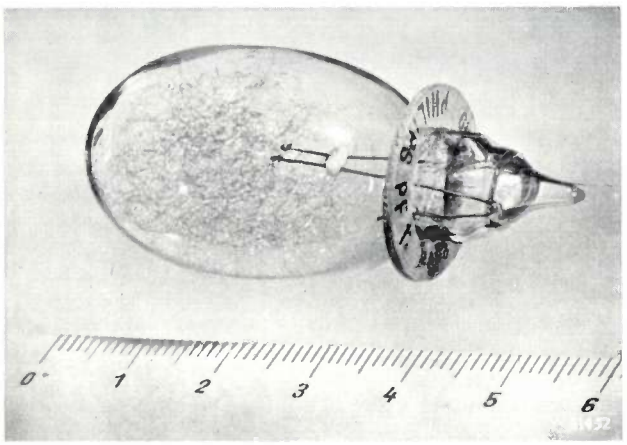
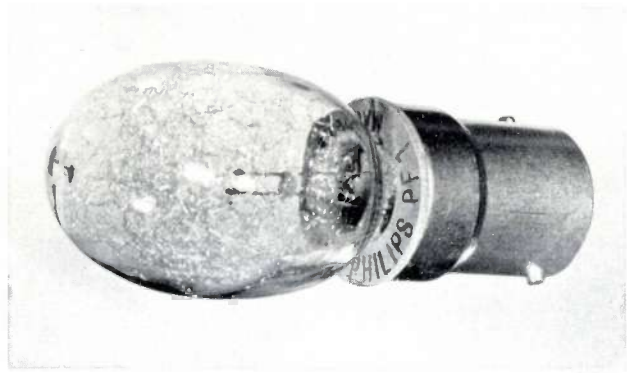


Fig. 4. Left: adaptor, consisting of a BA 15s bayonet cap (right) in which the lampholder of fig. 3 has been placed.





81451

Fig. 5. The "Photoflux" flash-bulb PF 1, with and without adaptor.

Fig. 5 shows the new "Photoflux" lamp, with and without adaptor. The adaptor will of course become superfluous as soon as the makers of flash equipment adapt their lampholders to this latest development in the flash-bulb.

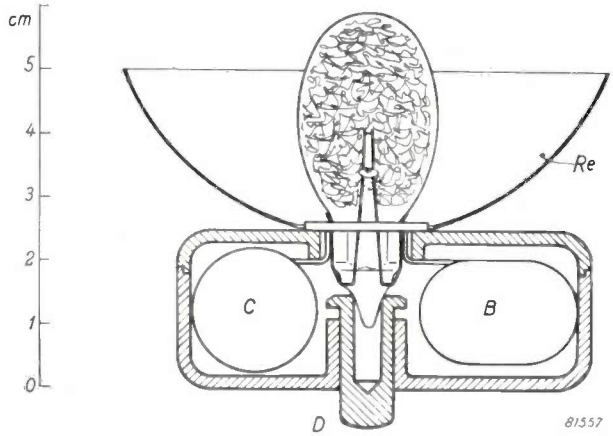


Fig. 6. Flash-bulb equipment with the new "Photoflux" PF 1. The construction of the apparatus shown here is based on ignition with the help of a capacitor (*C*), a method frequently used nowadays. The reflector *Re* has a diameter of 75 mm. *B* is the battery (22.5 V hearing-aid battery). *D* push button for ejecting the used lamp.

In *fig. 6* a section is shown of a flash equipment with the holder for the new flash-bulb. The small dimensions of the lamp also make it possible to use a small reflector, in order to obtain a uniform light distribution. The reflector shown has a diameter of 75 mm. The dimensions of the whole flash apparatus can thus also be made very small, which is of course a further advantage. Used bulbs can be ejected very simply and quickly by means of a push button, thus avoiding the handling of the hot glass envelope.

Summary. A description is given of a "Photoflux" flash-bulb, type PF 1, in which the metal cap used up to now is superseded by the neck of the glass bulb itself; this makes the bulb considerably cheaper. The outer ends of the lead-in wires are bent so that they lie along the neck, and are kept in place by a ring of insulating material. The flash-bulb is secured in the lampholder by the frictional grip of two flat springs, which also serve as connections to the lead-in wires for the ignition current. An adaptor for the new bulb, with a normal BA 15s bayonet cap to fit existing equipment, is described.

THE WEAR OF DIAMOND DIES

by L. SCHULTINK, H. L. SPIER and A. J. van der WAGT.

679.89:549.211:620.178.1

The great value of the diamond is not only due to its rareness and its excellent refractive properties, to which it owes its unique place among precious stones, but equally well to its tremendous mechanical hardness and resistance to wear. These last two properties in particular have won it an unassailable position among materials for tools. Apart from its use for cutting tools, for bearings and as an abrasive, diamond have proved of great value as drawing dies, for producing wire from very hard metals such as tungsten and molybdenum. Even diamonds, however, are subject to wear under continuous use. Special research in this field gives sufficient evidence to assume that the wear on diamond drawing dies can be reduced to a minimum by employing the most suitable cristallographic orientation of the drawing cone.

Introduction

From a technological point of view, diamond is a most useful material since it is the hardest of all known substances. Owing to this property its value as a material for tools is unsurpassed: 80% of the yield of the diamond mines is used for industrial purposes and only 20% finds its way into the jeweller's showcase.

In this paper we shall confine ourselves to the diamond drawing dies for the manufacture of fine wire from very hard metals, such as tungsten. Dies of any other material are not suitable for drawing very hard metals into wires of small diameters except at the cost of enormous wear on the dies.

When a diamond is to be used as a drawing die, a hole of the desired profile and minimum diameter has first to be drilled through it. This can be accomplished by a special process; holes having a purely circular cross-section can be obtained in all diameters. If the diamond is to function satisfactorily as a drawing die, the minimum diameter of the hole should not change and the aperture should retain its circular shape. This ideal state, however, cannot be maintained in practice; even diamonds are subject to some wear during use. After several thousands or tens of thousands of yards of tungsten wire have passed through it, the shape of the hole is found to have changed considerably.

The extent of this wear as well as the ultimate shape of the hole depends on several factors, among which may be mentioned the condition of the outer layers of the material to be drawn, the lubricant used and the temperature. Even if these factors are kept constant, remarkable differences in the nature and the rate of wear may occur. This is due to the fact that a diamond always contains impurities and inclusions that may influence its wear. Moreover, wear will occur in different structural planes, according to the direction of the drilled hole with respect to the crystal orientation, and it is well-known that the resistance to wear of diamond depends greatly on the crystallographic orientation of the faces that are subjected to the wear.

This effect will be dealt with in this article [1]). Before considering the wear during the wire drawing process we shall give a brief summary of the resistance to wear of diamonds in general.

Effect of crystallographic orientation on the resistance to wear of diamond

In the diamond-working industry it is a well-known fact that diamonds can be ground and polished far more easily in certain directions or on certain faces than in other directions. The wear resistance of diamond shows a pronounced anisotropy and the directions of minimum wear resistance are closely related to the crystal structure.

Crystal structure

Diamond crystallizes in the cubic system. The arrangement of the carbon atoms in an elementary cell is characteristic, and other solids having a similar arrangement are often said to have the diamond structure. In such materials the binding forces between the atoms are of the valence type. They have the important property of showing pronounced preferred directions in space. The carbon atoms in the diamond lattice each have four nearest neighbours, arranged tetrahedrally around the

[1]) This subject will be treated in greater detail in paper a by the same authors to be published shortly in Applied scientific Research.

central atom, the four bond directions being thus arranged symmetrically in space, as shown in *fig. 1*. The carbon atoms are situated at the corners and in the face-centres of a cube, as with the ordinary face-centred structure, but apart from these the cell

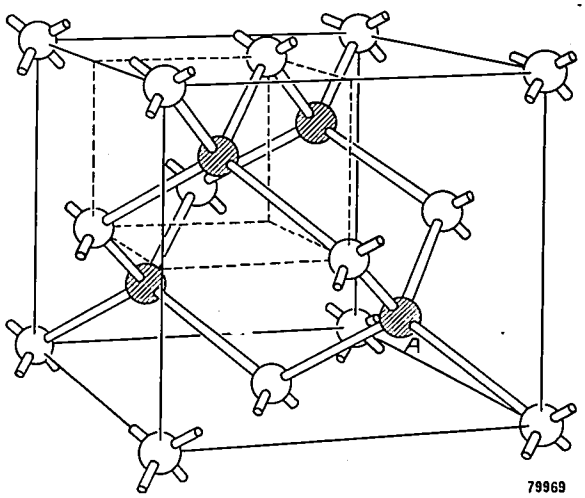


Fig. 1. Cubic unit cell of diamond.

contains another four atoms. If the cube is divided into eight similar small cubes, each having half the edge of the unit cell, then these four additional atoms (hatched in fig. 1) are to be found in the centres of four of the small cubes, which are so arranged that any two cubes have only one edge in common. The above-stated condition that any given carbon atom is tetrahedrally surrounded by four neighbours, is thus satisfied.

The hardness of the diamond is largely due to the fact that great energy is required to liberate one or more carbon atoms from their tetrahedral surroundings, since this involves the breaking of several bonds and disruption of their symmetry, which is essential in valence-bond crystals.

The orientation of the boundary planes of the natural diamond is as a rule simple and nearly always occurs in three forms viz:

a. Along the cube faces, termed (100)-planes in the crystallographic notation.

b. Along the rhombic-dodecahedral faces or (110)-planes. These are planes through a cube edge and parallel to the face diagonal of a plane perpendicular to this edge.

c. Along the octahedral planes or (111)-planes, i.e. planes perpendicular to a body diagonal of the cube.

Wear

When the various diamond faces are ground (neglecting for the moment the *direction* of grinding) it is found that the three planes most commonly occurring show different wear resistance. The rhombic-dodecahedral faces are as a rule most easily ground, i.e. the material is readily removed by a force acting in the rhombic-dodecahedral plane. The cube faces are harder to deal with, and the octahedral faces are the most wear-resistant, which means that a force acting in the octahedral plane removes only very little material.

In a given plane, however, the *direction* of grinding is also of importance. Each type of plane is found to have one or more preferential directions, along which the wear resistance is smallest. In the (110)-plane this preferential direction is along the cube direction in that plane, in the (100)-plane both cube directions are equally preferential, whereas in the octahedral planes the orthogonal projections of the three cube directions in that plane are the directions of preference. *Fig.* 2 shows the fundamental orientations of minimum wear resistance: i.e. the directions along which the diamond crystal is most easily ground.

For directions other than the preferential, the ease of grinding may well be approximated by:

$$S = A \cos^2 \Theta, \quad \ldots\ldots\ldots (1)$$

in which A represents a constant depending solely on the nature of the grinding plane. According to the foregoing, A will be large for (110)-planes and small for (111)-planes. Θ indicates the angle between the actual and the preferred grinding directions.

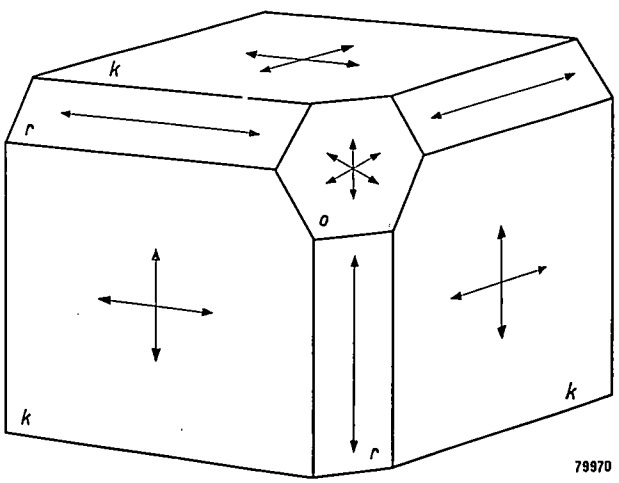


Fig. 2. Schematic diagram of preferred orientations for diamond grinding. The preferred directions on the various crystal faces are indicated by arrows whose lengths are proportional to the ease of grinding. Cube planes are indicated by k, rhombic-dodecahedral planes by r, and octahedral planes by o. (Diagram taken from the article referred to in [2]).)

The cohesion of the atoms in a diamond crystal can be broken not only by grinding, but also by cleavage. The octahedral plane proves a favourable cleavage plane, which means that a force applied so that it acts perpendicular to the octahedral planes will effect a breaking of the bonds relatively most easily. As in the case of grinding, just the opposite applies to forces acting parallel to these planes.

The explanation of this anisotropy in diamond wear, according to Stott [2]) and others is to be found in the nature of the directional bonding forces between the atoms. As an example we shall consider here only the case of the wear resistance of the rhombic-dodecahedral plane ((110)-plane). *Fig. 3* shows the configuration of the atoms in a diamond lattice as seen from a direction at right angles to this plane. The atoms indicated by the black dots lie in the plane of the drawing, and therefore in the rhombic-dodecahedral plane to be ground; the circles show the atoms in the first layer beneath this plane. The connecting lines show the interatomic bonds; the full lines are those parallel to the plane of the drawing and the dotted lines show the bonds with the deeper layer. For the sake of clarity the atom *A* corresponds to that denoted by *A* in fig. 1. It appears that atom *A* can be more readily removed by a force in the direction of arrow *1*, i.e. in the cube direction, than by one acting along arrow *2* which is perpendicular to it. The cube direction is consequently to be preferred in grinding.

Similar considerations will satisfactorily explain other directions of preference with respect to wear and cleavage.

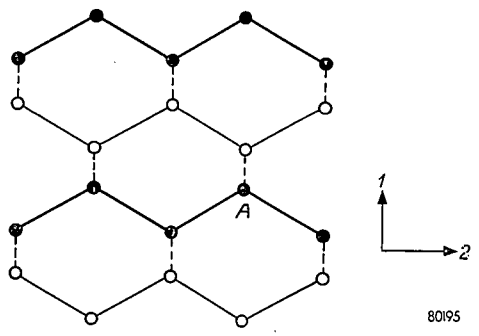


Fig. 3. Configuration of atoms in a diamond lattice as seen perpendicular to a rhombic-dodecahedral plane. The black dots represent the surface atoms in the rhombic-dodecahedral plane; the circles represent atoms in the underlying rhombic-dodecahedral plane. The full lines show the bonds parallel to the plane of the drawing, while the dotted lines represent the bonds between atoms of the two layers. Arrow *1* indicates the cubic direction, and arrow *2* the direction of a face diagonal. Atom *A* (similarly marked in fig. 1) will more easily be removed by a force parallel to arrow *1* than by a force parallel to arrow *2*. (Cf. the article referred to by [2]).)

Wear on a drilled hole in the course of the drawing process

Let us now return to the wear on the drawing dies. There is conclusive evidence to justify the assumption that the boundary surface of the effective part of the drawing hole is composed of submicroscopic elements of crystal planes belonging to each of the above-mentioned categories. Let us first consider a hole drilled in the direction of the cube edge, i.e. at right angles to a cube face. The hole will then probably be bounded by cube planes *k* and rhombic-dodecahedral planes *r* (*fig. 4.*). According to our explanation these *r*-planes, if ground in a cube direction (as occurs, in this particular case, when wire drawing), will wear faster than any other crystal plane. The rhombic-dodecahedral planes will thus show the greatest rate of wear and the hole cross-section will assume a more or less square shape. Wear of the cube faces is also inplied in the assumption of this shape. As these are also relatively easily wearing the process proceeds fairly quickly.

If a hole is drilled perpendicular to a rhombic dodecahedral plane, it will be bounded by crystal planes of all the three types discussed (*fig. 5*). None of the directions of easy wear now coincides exactly with the drawing direction. To ascertain which type of plane has the quickest rate of wear, it is necessary to decide which plane has the highest value of *S* (equation *1*). In this case the cube faces are those subject to the greatest wear. As shown schematically in fig. 5 the aperture will gradually become oval, with the major axis perpendicular to cube faces (which implies parallel to a cube direction). The rate of wear will be slower here than in the first instance, since no grinding takes place along a preferential direction.

If the hole is drilled perpendicular to an octahedral plane, it is surrounded by rhombic-dodecahedral planes, all equivalent as regards their orientation with respect to the drawing direction

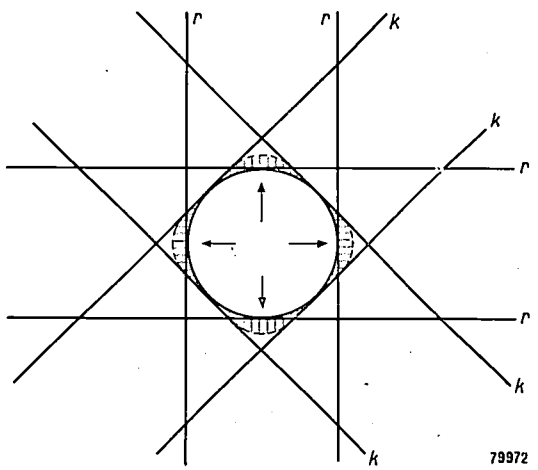


Fig. 4. A hole drilled perpendicular to a cube face is bounded by cube planes *k* and rhombic-dodecahedral faces *r*, as indicated schematically. Wear will occur mainly in the *r*-planes, so that the aperture will eventually assume a square shape.

[2]) W. Stott, National Physical Laboratories, Collected Reports **24**, 3-55, 1938.

(*fig. 6a*). It might, therefore, be expected that the aperture would become hexagonal, all faces showing a uniform wear. It has been found, however, that the drilled hole with this orientation often assumes a triangular cross-section.

A wholly satisfactory explanation of this phenomenon has not yet been found. It is possible that

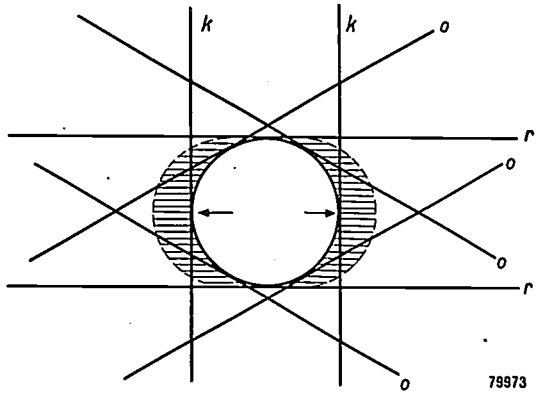


Fig. 5. A hole drilled perpendicular to a rhombic dodecahedral plane is surrounded by crystal planes of all the three main types *k*, *r* and *o* (shown schematically). Wear will occur mainly in the cube planes, so that the aperture will tend to assume an oval cross-section.

negative angle with the drawing direction, will now apparently predominate. More wear will occur on the octahedral planes o "along" which the grinding action takes place than on those marked 0′ "into" which grinding occurs. This provides a qualitative understanding of why the aperture gradually becomes triangular. Owing to the predominance of octahedral planes and to the relatively unfavourable orientation of the grinding direction on them, the rate of wear will be very slow, much slower indeed than in the two previous cases.

The conclusions regarding the symmetry of the ultimate shape of the drawing aperture can also be reached along entirely different lines of reasoning, since this shape will always be determined by the anisotropy of the diamond crystal. Drilling along a three-fold axis will produce a hole which wears to a cross-section of three-fold symmetry, etc. An altogether different matter is the *rate* of wear; information on this can only be gained by ascertaining *which* planes are being ground.

The above considerations lead to the theoretical conclusion that drawing holes should preferably be drilled as nearly as possible perpendicular to an

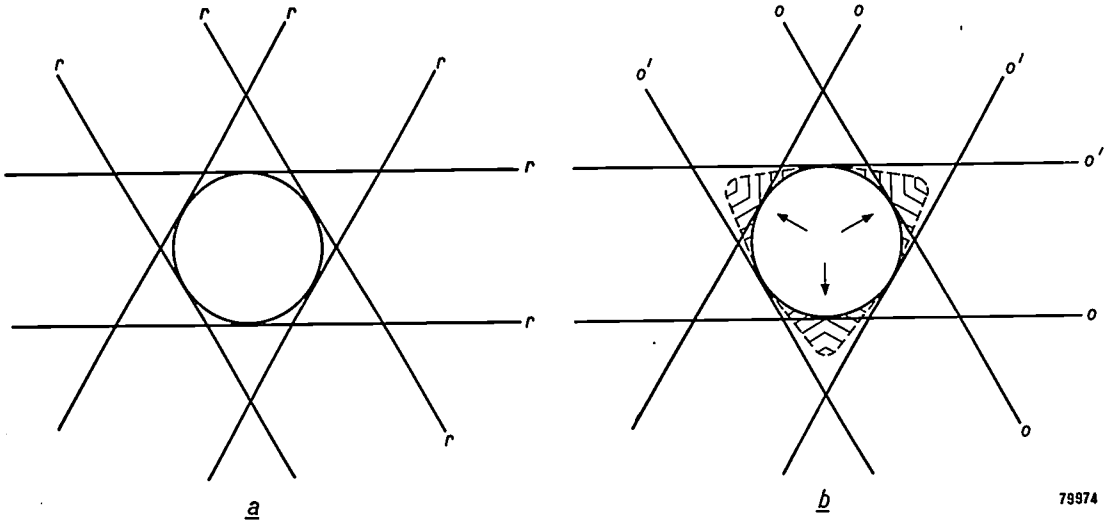


Fig. 6. *a*) A hole drilled perpendicular to an octahedral plane is surrounded by rhombic-dodecahedral planes *r*, all wearing at a uniform rate. *b*) It is supposed that, in the course of the wearing process, these planes largely give way to the octahedral planes *o* and *o′*, of which only those marked *o* will wear because of their favourable orientation with respect to the drawing direction. The ultimate cross-section of the hole will be triangular.

the rhombic-dodecahedral planes surrounding the hole, ultimately give way to octahedral planes (fig. 6*b*). Although not quite parallel to the drawing direction, they lie at a relatively small angle to it. The wear on these octahedral planes, half of which lie at a small positive angle and the other half at a octahedral plane in order to incur the minimum amount of wear.

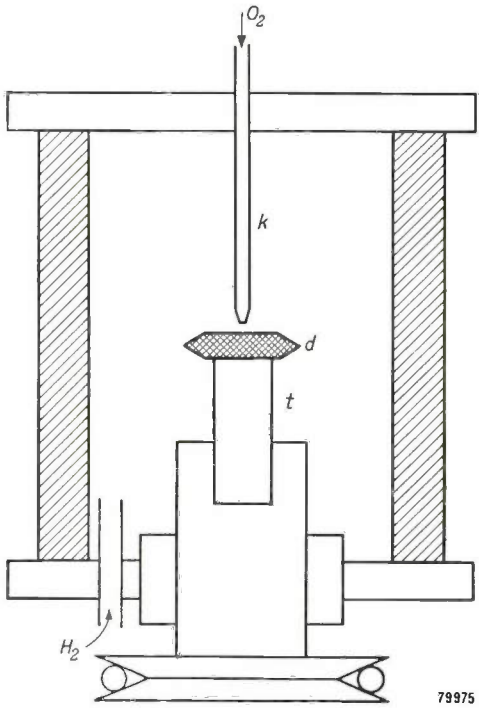
Experimental research into the wear resistance as a function of the crystallographic orientation presents considerable difficulties, particularly owing to the fact that the homogeneity of diamond crystals used for wire-drawing dies is far from perfect.

Only a large-scale statistical investigation can provide the solution. For certain applications, in which large numbers of diamond dies are used, research of this kind may lead to a considerable reduction in diamond wear.

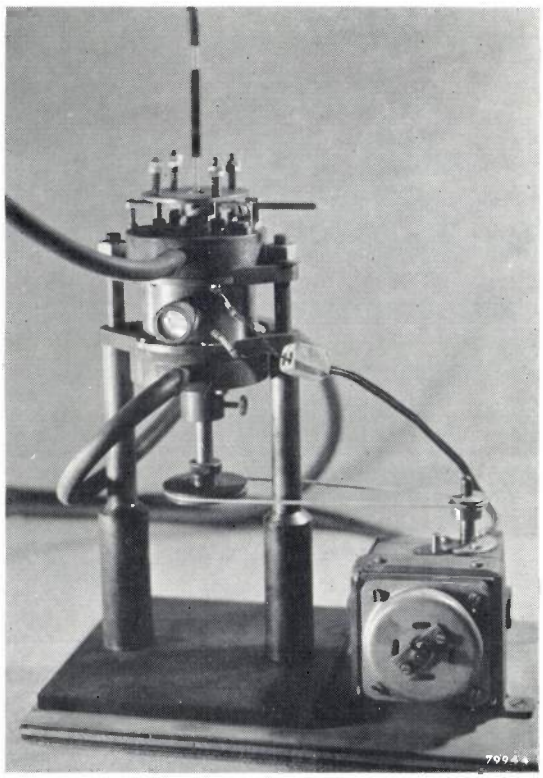
One aspect of this research may be mentioned here, viz. the determination of the orientation of a die from its X-ray diffraction pattern. The recently developed X-ray image intensifier[3]) permits of a direct visual observation of this diffraction pattern; it is no longer necessary to record it photographically.

As a further check on the above-mentioned conclusions, attemps have been made to attack diamond in a different way, namely by a process of etching or, more strictly, of burning. The results thus obtained are found to confirm the theoretical approach.

The diamond under test is fixed in a rotatable jig inside an electric furnace (*fig.* 7). The face to be etched lies horizontally and is faced at a short distance by the narrow jet nozzle of a silica tube set in a movable mount. By suitable movement of the nozzle, combined with rotation of the diamond, the whole surface of the latter can be covered. Inside the oven the temperature in maintained at 800-900 °C, whilst an atmosphere of hydrogen is supplied to prevent combustion of the diamond. A fine jet of oxygen blown out of the moving nozzle onto the diamond then has a pronounced "etching" effect (actually a burning).



a



b

Fig. 7. *a*) Apparatus for the etching of a diamond by burning with oxygen. The diamond *d* is fitted on a rotatable table *t*. A movable silica nozzle faces the upper surface of the diamond. The whole assembly is placed inside a furnace supplied with a hydrogen atmosphere; the temperature inside the oven is approximately 850 °C. A jet of oxygen is blown onto the diamond through the silica tube.
b) Photograph of the apparatus.

Etching of diamond faces

Diamond, as is known, is totally inert to the conventional pickling and etching reagents. In order to study the etch patterns on the three main faces of diamond, a special method has therefore to be applied:

It has been found that the time required to obtain the same etching effect on a (111), a (110) and a (100)-plane is quite different. This is shown in the following table, which applies to a certain uniform depth of etching.

It is clear that the rhombic-dodecahedral plane (110) is most easily etched whilst the octahedral plane (111) presents most difficulties. The removal

[3]) M. C. Teves and T. Tol, Electronic intensification of fluoroscopic images Philips tech. Rev. **14**, 33-43, 1952/53.

Crystal plane	Etching time in minutes
(110)	5
(100)	8
(111)	$11^1/_2$

of carbons atoms, therefore, both by grinding and by burning, is most easily effected from the rhombic-dodecahedral plane and is most difficult from the octahedral plane.

If etching takes place with the diamond rotating and a stationary jet, so that the burning is localized, it is found that the etching patterns, formed on the various diamond faces are, of the same form as the corresponding drilled holes worn in the course of wire drawing. The etching patterns obtained are shown in *fig. 8*. The square shape of the hole burnt into the cube face can be clearly distinguished from the oval crater in the rhombic-dodecahedral face and the triangular one in the octahedral face. These symmetries are entirely similar to those shown ultimately by the drawing holes in the various orientations (figs 4, 5 and 6) and in accord with that to be expected from the crystal anisotropy.

If the completely etched surfaces are examined under the electron microscope, then the cube faces are found to be covered with quadrilateral pyramids (*fig. 9a*). The rhombic-dodecahedral faces resemble something like a Jura landscape (fig. 9*b*), but the orientation of the valleys clearly coincides with the crystal directions: the valleys are oval, with the major axes lying in the cube direction. On the octahedral faces trilateral pyramids have appeared, bounded by cube planes (fig. 9*c*).

These photographs illustrate very well what happens during a polishing or grinding process.

When grinding takes place along the (111)-plane the trilateral pyramids will soon disappear because of the comparatively easy cleavage along the (111)-planes. After all pyramids have disappeared a perfectly smooth plane is left, giving very little foothold for further grinding. The octahedral face is hence highly resistant to grinding.

On the (110)-plane new mountain ridges continue to appear because these cannot be attacked along a direction of easy cleavage and wear will continue, with a far greater uniformity. There will, however, be a great difference according to whether the grinding action occurs across or along the mountain ridges: in the latter case, wear is less readily achieved.

When a cube face isground new pyramids will continue to appear. These new pyramids cannot disappear by cleavage because no plane of easy

Fig. 8. Etch holes on the major planes of diamond;
a) on a cube face.
b) on a rhombic-dodecahedral face.
a) on an octahedral face.

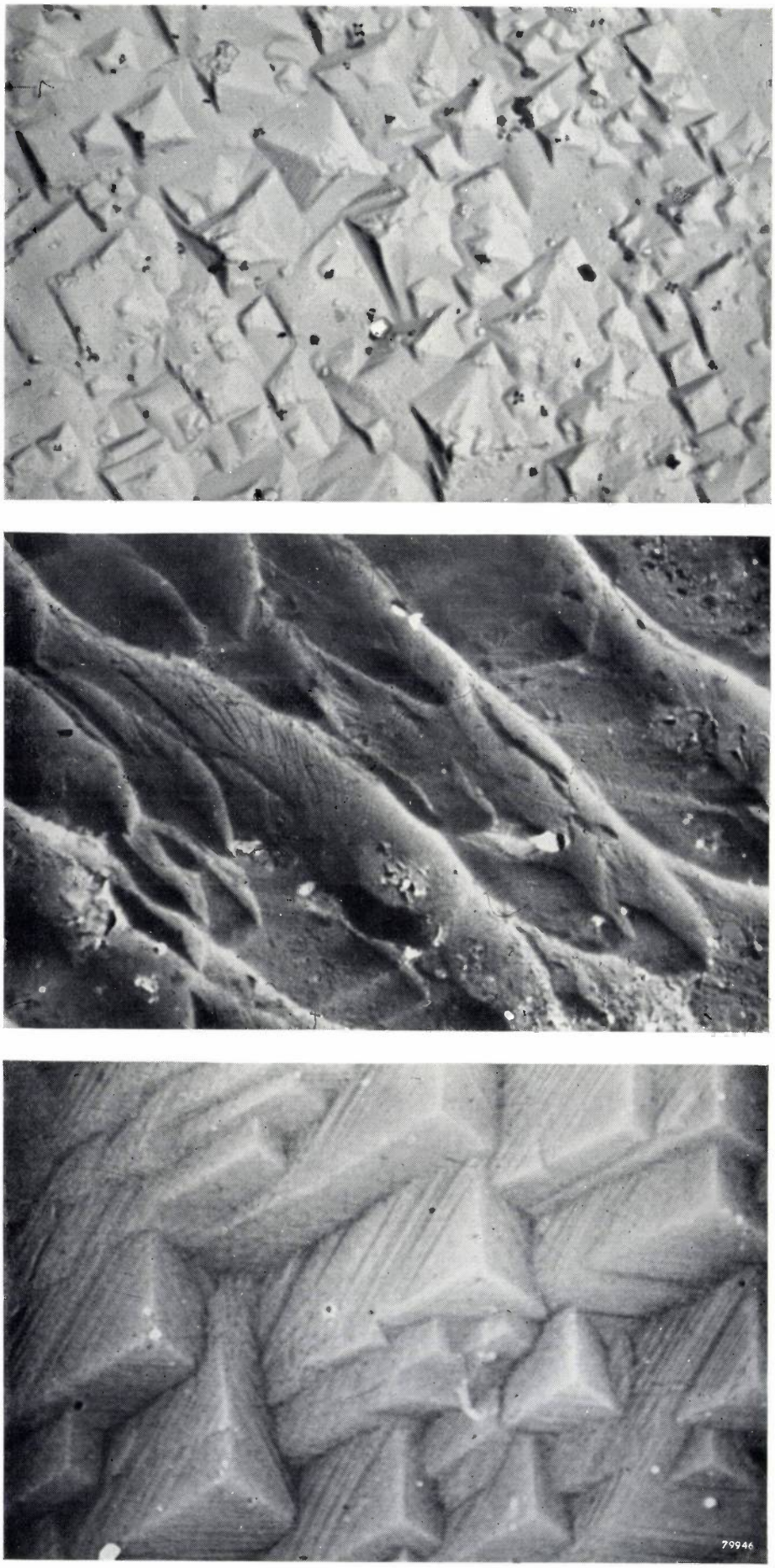


Fig. 9. Electron photo-micrographs of etched crystal faces of diamond. Enlargement 10 000 ×.
a) cube face.
b) rhombic-dodecahedral face.
c) octahedral face.

cleavage is favourably orientated. The resultant rough surface is relatively easily ground.

Summarizing, we may therefore claim that the picture obtained by etching and made visible by the electron microscope, fully confirms the theoretical picture of the wear resistance of the three major planes. Practice, however, must finally decide whether drawing dies with the theoretical optimum orientation of the hole will indeed show the least wear. As already mentioned, this will require an extensive statistical investigation, in view of the effects of factors other than the crystal orientation.

Summary. Theoretically it may be expected that the wear diamond dies will greatly depend upon the crystallographic orientation of the drilled hole, on account of the crystallographic arrangement and the special type of bond between the carbon atoms. Because of this, there are certain preferential orientations of a diamond face along which grinding is relatively more easily achieved. This is confirmed by diamond-working practice. The conclusion is reached that the least wear on a drawing die will occur if the hole is drilled perpendicular to an octahedral plane. This conclusion is checked against the etch patterns obtained by a special technique on the major faces of a diamond. Practical confirmation of these conclusions can only be obtained after a statistical examination of a great number of drawing dies.

CHEATER CIRCUITS FOR THE TESTING OF THYRATRONS

II. LIFE TESTING AT HIGH COMMUTATION FACTORS

by M. W. BROOKER *) and D. G. WARE **).

621.387:621.385.38

In Part I of the present article the authors described a "cheater circuit" which allows the measurement of the grid current of thyratrons under the full rated operating conditions, without (in the case of high-power valves) involving a large power consumption. In Part II, which now follows, the factors which control the rate of "gas clean-up" under practical conditions are considered. The more important of these are due to circuit characteristics embraced by the term "commutation factor". For thyratrons intended for industrial use, where the commutation factor may be quite high, any form of life-testing must be designed to simulate these conditions. This problem too has been solved by the use of a cheater circuit, which in this case is so designed that the test conditions to which the thyratron is subjected, are quite severe.

The causes of gas clean-up in thyratrons

For a given electrode structure in a thyratron, the gas pressure which may be used is determined by the voltage rating of the valve: for the greater the gas pressure used, the lower will be the break-down voltage of the valve. The life of the valve, however, may be prematurely ended if the gas pressure employed is very low, because the gas is liable to disappear, by a process which is termed gas clean-up [1]). The mechanism of gas clean-up has not been accurately ascertained, but it is believed to be due to high-energy gas ions hitting the electrodes with sufficient energy to penetrate the metal surface and thus becoming trapped. Now during the operating cycle of a thyratron there are two occasions on which high-energy ions may be formed:

1) during the ionization time of the valve, if the applied voltage at the moment of firing is high, and
2) immediately after conduction has ceased, if the inverse voltage appears rapidly.

These two occasions will now be considered in more detail.

*) Mullard Radio Valve Co., Ltd., Mitcham England.
**) Formerly with Mullard Radio Valve Co., Ltd.
[1]) This refers only, of course, to valves not containing a liquid phase — e.g. mercury —, which by evaporation makes up for the vapour loss caused by clean-up. For many applications, however, this advantage of mercury is outweighed by the fact that the pressure of the (saturated) vapour rapidly increases with the ambient temperature, which imposes a strict upper limit on this temperature. In such cases, therefore, thyratrons with a gas-filling without a liquid phase are often to be preferred. As filling gases, the rare gases are commonly used, and, in special cases, hydrogen.

Gas clean-up occurring during the valve ionization time

If a cathode-ray oscilloscope is used to observe the waveform produced when a thyratron is operated from an A.C. source, then if conduction is initiated at the peak applied voltage, the waveform will appear as in *fig. 1*. If the frequency of the applied voltage is 50 c/s, then the time elapsing between point A and point D is 10 milliseconds. On the observed trace the portion BC will appear vertical and very faint, showing that the valve passes from the non-conducting state to the conducting state very rapidly. In order to measure the time elapsing between points B and C, elaborate apparatus is required. This time is called the ionization time of the valve. It depends upon the applied anode and grid voltages. For a xenon-filled valve it may be as long as 0.5 microseconds for an anode voltage of 100 V, if the grid not driven positive more than 10 V. It falls to, say, 0.1 μsec at an anode voltage of 500 V, and if the grid is driven 50 V positive, the ionization time with

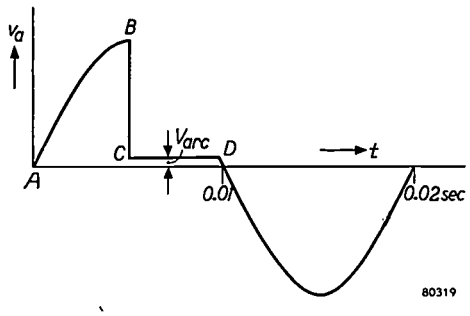


Fig. 1. Anode voltage waveform V_a of thyratron fired near peak of 50 c/s A.C. anode voltage. Ionization processes cause rapid break-down of voltage from B to C.

500 V applied to the anode may be as little as 0.01 μsec [2]).

If it were possible to observe the anode voltage and current waveforms for, say, 1 μsec after conduction is initiated, waveforms like those of *fig. 2* would be seen. It is clear, therefore, that between t_0 and t_1 (an interval of, say, 0.1 μsec) a certain number of high-energy positive ions will be produced. These positive ions will be rapidly accelerated to any negative surfaces and especially to the cathode. On colliding with a metal or glass surface an ion will have a certain probability of penetrating the surface and becoming trapped in the interior of the material. This probability becomes greater the higher the energy of the ion [3]).

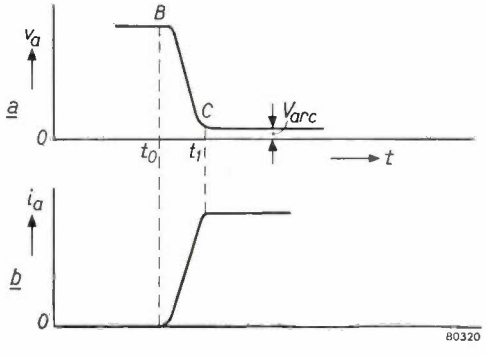


Fig. 2. *a*) Section *BC* of fig. 1, scaled-up horizontally to show gradual fall of voltage during ionization.
b) Corresponding growth of anode current i_a.

It seems probable, therefore, that gas clean-up would depend upon the type of grid excitation employed, for this in turn affects the ionization time of the valve. Thus it would appear that pulse firing of a valve would be beneficial and would produce a lower rate of gas clean-up than other forms of grid control. These expectations are confirmed in practice.

Gas clean-up occurring immediately after current extinction

A thyratron requires a considerable time after conduction has ceased for all ions present to recombine. This time is termed the de-ionization time and usually amounts to 100 μsec or more. Now it is evident that if a large negative anode voltage appears across a thyratron immediately after cessation of conduction, then the positive ions present will be accelerated to the anode and may penetrate its surface and become trapped. Thus gas clean-up may occur at current extinction by the same mechanism as at the commencement of conduction. The valve de-ionization time being long compared with the ionization time, it would appear probable that more gas would be cleaned up at current extinction than at the commencement of conduction.

However, this depends upon the waveforms of current and voltage experienced by the valve. Thus the number of ions present immediately after current extinction will depend upon the rate of decay of current prior to extinction, and the force of impact of an ion will depend on the rate of rise of the inverse voltage.

As a measure of the combined effect of these two factors in producing gas clean-up, the *commutation factor* is useful. This is a property of the circuit rather than of the valve. It can be defined as

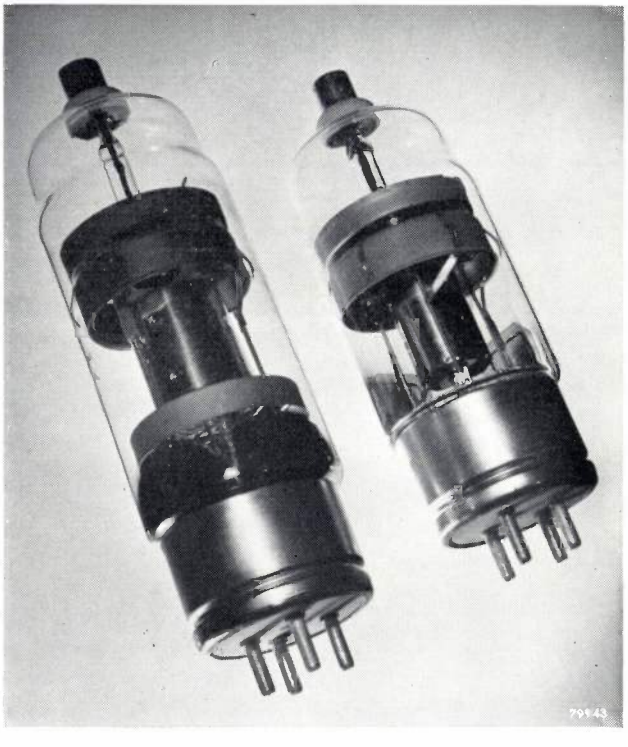


Fig. 3. Mullard thyratrons type 5544 (right) and 5545 (left) both with xenon filling and directly heated cathode. The main data are as follows:

	5544	5545
Maximum peak anode voltage		
forward	1500 V	1500 V
inverse	1500 V	1500 V
Maximum anode current		
peak	40 A	80 A
mean	3.2 A	6.4 A
Maximum commutation factor	130 $VA/\mu s^2$	130 $VA/\mu s^2$
Maximum diameter	66 mm	66 mm
Overall length	178 mm	203 mm

[2]) Milton Birnbaum, A method for the measurement of the ionization and de-ionization times of thyratron tubes, Trans. Amer. Inst. El. Engrs. **67-1**, 209-214, 1948.
[3]) M. J. Reddan and G. F. Rouse, A study of helium gas clean-up in an electric arc discharge, Electrical Eng. **71**, 159-164, 1952.

the product of the average rate of decay of anode current and the average rate of rise of inverse voltage, the units being amps per μsec and volts per μsec [4] [5]).

The higher the commutation factor, the more rapid the gas clean-up. A maximum value for the commutation factor is sometimes included in the published data for a thyratron. In such cases, it is obviously essential for the maker to have some simple and convenient method of life testing valves at varying rates of commutation. Such a method should make it possible to vary the rate of commutation without making major changes in the circuits. It should function when high peak currents are in use, and yet its power consumption should not be excessive. Below, a few straightforward circuits are considered in the light of these requirements, and shown to be not entirely satisfactory. A method of testing based on a "cheater" circuit is then described which has proved successful in all respects. It has been used for the operational testing of Mullard thyratrons types 5544 and 5545 (*fig. 3*; some data on these valves are given in the subscript). The present Mullard thyratrons are developed from a prototype originated elsewhere [6]).

Performance of conventional circuits

The simple circuit of *fig. 4* can meet only one of the requirements for life tests: it can produce any peak-to-mean current ratio from π upwards (π being the ratio for half sine waves, see fig. 3 of Part I [7])). It cannot, however, produce the rapid commutation which is a primary need.

If an inductance is included in the circuit (*fig. 5*). it is possible to obtain a high rate of growth of inverse voltage in the valve after conduction has ceased. Because of the lagging voltage generated in the inductance, the anode of the test valve is held positive after the applied voltage has become negative. Consequently, current continues to flow through the valve for some time. When the conduction ends, however, the inverse voltage across the valve rises abruptly to a high value, corresponding to the applied voltage. Thus, one of the causes of rapid clean-up is present; but with this circuit it is impossible to combine it with the other — a high rate of current decay.

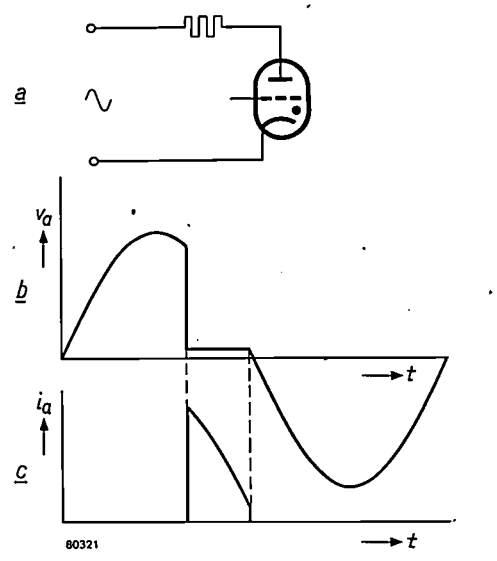


Fig. 4. *a*) Thyratron with purely resistive load. *b*) Anode voltage V_a. *c*) Anode current i_a. At current extinction both current decay and inverse voltage build-up are slow, so no high commutation factors can be obtained.

A simple two-valve full-wave circuit, with a resistance-inductance load (*fig. 6*), gets round this difficulty. The anode current flows through one valve until the other strikes. It then drops to zero almost instantaneously and the inverse voltage

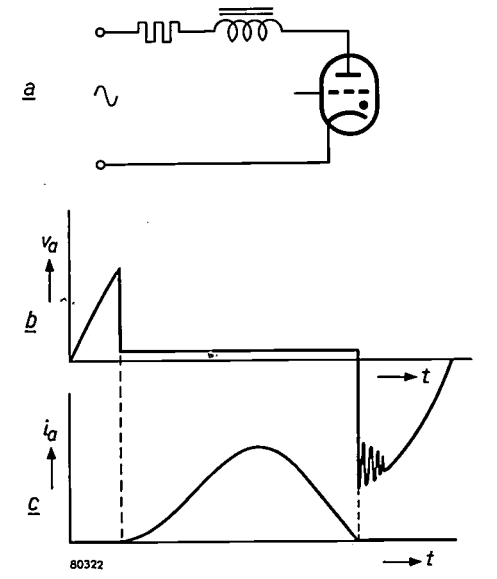


Fig. 5. With a thyratron having a partly inductive load, circuit parameters can be so chosen that conduction ceases near the negative peak of the A.C. voltage. The current decay, however, is slow, so the commutation factor is still not very high. (At current extinction, oscillations are set up in the circuit formed by the inductance and stray capacitance.)

---

[4]) Normally commutation factor is calculated by taking as averageing time of current decay the last 10 μsec before extinction, and as averageing time of inverse voltage rise the interval in which this voltage rises from 0 to 200 V.

[5]) D. V. Edwards and E. K. Smith, Circuit cushioning of gas-filled grid-controlled rectifiers, Trans. Amer. Inst. El. Engrs. **65**, 640-643, 1946. D. E. Marshall and C. L. Shackelford, Commutation factor in thyratron circuit design, Electronics **27**, 198, March 1954.

[6]) A. W. Coolidge Jr., A new line of thyratrons, Trans. Amer. Inst. El .Engrs. **67-1**, 723-727, 1948.

[7]) M. W. Brooker and D. G. Ware, Cheater circuits for the testing of thyratrons, I. Measurement of grid current, Philips tech. Rev. **16**, 43-48, 1954/55, (No. 2).

climbs quickly to its peak. Thus the commutation is extremely rapid. However, there remain two important drawbacks to the use of this circuit. Firstly, the peak-to-mean current ratio is only slightly more than two. Secondly, its power consumption is high, for it is not possible to obtain a high inverse voltage from it while using low-voltage supply.

The simplest conventional circuit that can give both speedy commutation and an adequate peak anode current appears to be a six-valve hexaphase circuit. But there is a serious objection to this as a piece of operational test gear: in the case of six fully loaded type 5545 thyratrons, for example, about 55 kW of power would be dissipated.

It is clear that no simple conventional circuit is suitable for the life testing of thyratrons at high rats of commutation. Once again, a "cheater circuit" has been devised, which permits both rapid commutation and low power consumption.

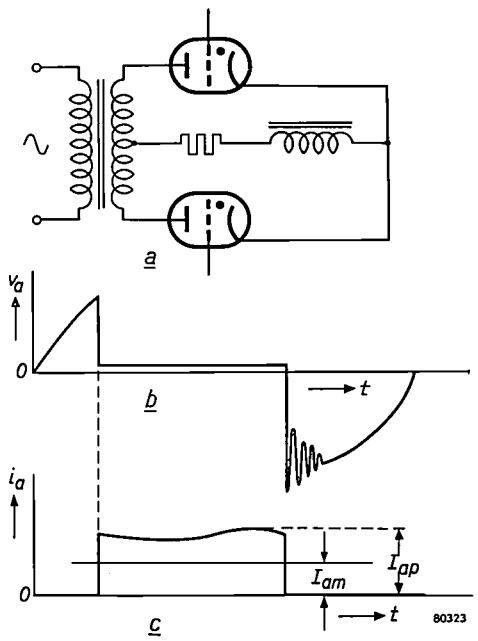


Fig. 6. *a*) Two thyratrons in full-wave circuit with resistance-inductance load. Both inverse voltage rise (*b*) and current decay (*c*) are rapid, but peak-to-mean anode current ratio I_{ap}/I_{am} hardly exceeds 2, and power consumption is high.

The fundamentals of the circuit are shown in *fig. 7a*. The thyratron under test, Th_1, is operated from a 220 V A.C. supply. Its mean anode current can be varied by adjusting the load resistor R_1, and the grid control (which will be described later) is so arranged as to give the correct peak-to-mean

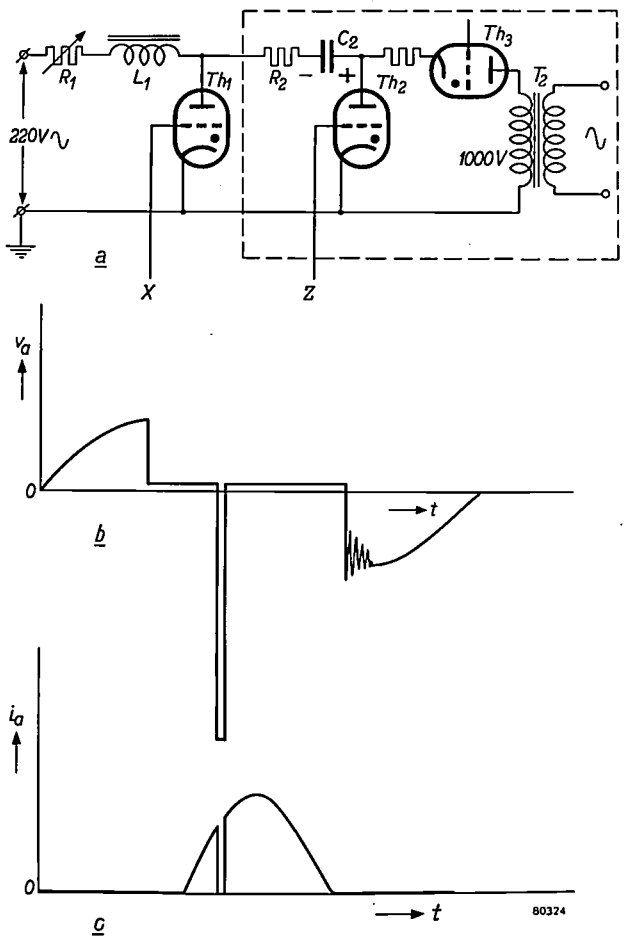


Fig. 7. *a*) Cheater circuit for thyratron testing at very high commutation factors. Th_1 valve under test, with resistance-inductance load R_1-L_1. The dotted line encloses a circuit for producing high negative voltage pulses at the anode of Th_1 that extinguish Th_1. This circuit comprises capacitor C_2 charged to approx. 1500 V from H.T. transformer T_2 via thyratron Th_3. When thyratron Th_2 is triggered, C_2 discharges through Th_1 (which is then conducting) and Th_2. Valve Th_1 is extinguished by the discharge, as shown by voltage and current waveforms for Th_1, at (*b*) and (*c*).

Cheater circuit details

An apparatus will now be described which has been used for testing type 5545 thyratrons.

In this circuit the high commutation factor required is obtained by feeding a short pulse of high negative voltage to the anode of the test valve while it is conducting. This causes a sudden break in conduction; the current falls rapidly to zero and an inverse voltage builds up at a very high rate.

anode current ratio. The rest of the circuit shown has only one function: to produce the pulses of negative voltage at the anode of the test valve. These pulses are drawn from the 1000 V transformer winding which produces a voltage in anti-phase with the mains voltage.

The sequence of events in the circuit is as follows. While the test thyratron Th_1 and an auxiliary thyratron Th_2 are not conducting, the capacitor C_2 is charged via a second auxiliary thyratron Th_3

it reaches approximately +1500 V with respect to neutral. When the flow of current through Th_3 has ceased, the test valve Th_1 is triggered. Then, while Th_1 is conducting, Th_2 is triggered. This completes the circuit of the capacitor C_2, which immediately discharges via the resistor R_2 and the valves Th_1 and Th_2.

The result is that suddenly the anode of the test valve is driven 1500 V negative (fig. 7*b*) and the flow of current through it is stopped abruptly (fig. 7*c*). Thus an extremely rapid decay of current and growth of inverse voltage occur simultaneously at Th_1, i.e., a high commutation factor has been obtained.

Whether or not the valve resumes conduction as the negative pulse decays, depends on the de-ionization time of the valve and on the method of control. However, this does not affect the operation of the test circuit, and neither would a resumed conduction unduly increase the power consumption, as the supply voltage is low.

Snubber circuits

Using the procedure described above, the commutation factor was found far too high for the original purpose, in fact it was about 10 000 VA/μsec², instead of the 80 VA/μsec² required. A special "snubber circuit", or "cushioning circuit", was therefore introduced into the apparatus to limit the commutation factor (see the first article mentioned in footnote [5]). The use of the normal resistor-capacitor combination in this rôle proved impracticable because the amount of energy available in the capacitor was so limited. Instead, therefore, an inductor L_3 and a capacitor C_3 were employed as in *fig. 8a*. They form a series resonant circuit via Th_2 while Th_1 is conducting. The voltage which appears across the test valve Th_1 as a result of this is shown in fig. 8*b*.

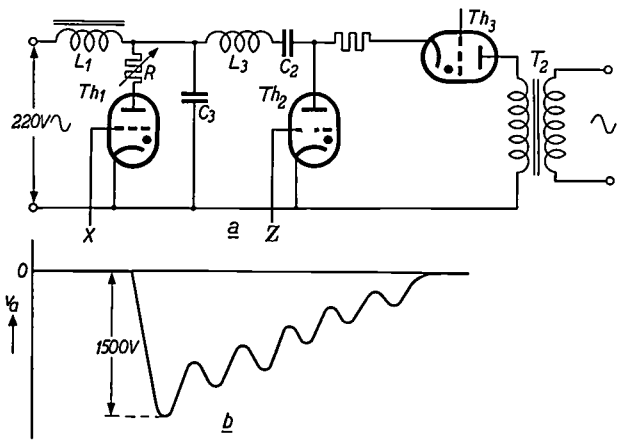


Fig. 8. *a*) Cheater circuit as in fig. 7*a*, but with "snubber" circuit L_3-C_3 for limiting the commutation factor to 80 VA/μsec². *b*) Voltage v_a appearing across the test valve.

Owing to the oscillatory nature of the discharge, the inverse voltage rises higher than it would if oscillations were absent; a peak inverse voltage of 1500 V is easily produced.

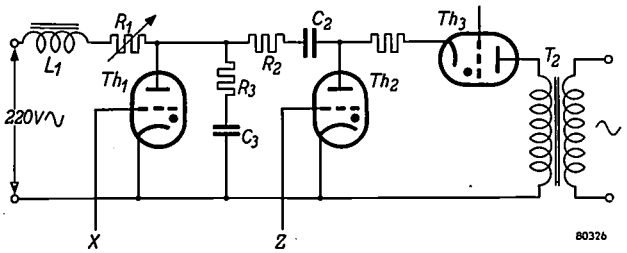


Fig. 9. Cheater circuit as in fig. 7*a*, but with an *RC* snubber circuit (R_3-C_3), used when commutation factors higher than 80 VA/μsec² are required.

Using a suitable snubber circuit, it proved possible to obtain the required commutation factor of 80 — a current decay rate of 2 A/μsec and an inverse voltage rise of 40 V/μsec. When later higher commutation factors were required, they were obtained by reverting to the usual resistor-capacitor snubber circuit (*fig. 9*).

Grid control requirements

Three separate grid control circuits were required, one for each thyratron. Taking first the test valve, Th_1, there were two important conditions which had to be fulfilled.

Firstly, it was undesirable that the grid of a type 5545 thyratron should be positive with respect to the cathode while the anode was highly negative. This would lead to a sustained discharge to the grid, and the positive ions produced would be drawn to the anode and might lead to undue clean-up or possibly to arc-back.

Secondly, it was necessary to guard against certain unwanted effects likely to be met during the early development of the 5545 thyratron i.e., grid emission and leakage troubles.

To ensure stable operation despite these defects in the valve, it seemed wisest to use a large negative bias on the grid and to trigger the valve with a short pulse. With this aim the circuit of *fig. 10* was employed. The first stage of this circuit is identical to that of fig. 6 of Part I [7]), but now a cathode follower is used as an output stage, from which pulses are fed to the grid of Th_1.

The grid control requirements of thyratron Th_2 were settled by the need to prevent conduction in Th_2 at the same time as in Th_3, i.e., the grid of Th_2 must be negative before the start of the positive half cycle of the H.T. transformer, and the grid must remain negative until some period during the conduction of Th_1. This was obtained by the use of a D.C. bias with a trigger pulse fed from a flip-flop circuit as shown in *fig. 11*. This circuit is stable with section *1* extinguished and section *2* conducting. The trigger pulse from the circuit shown in fig. 10 is used to render section *1* conducting and to extinguish section *2*. The variable resistor R_6 is used to vary the period during which section *2* is extinguished. When section *1* is extinguished again, a positive pulse is fed to a cathode-follower, and from thence (point *Z*) to the grid of Th_2. A choke is connected in the cathode circuit of the final stage of this

grid pulse network, thus reducing the rate of rise of current through the cathode-follower. This ensures that an oscilloscope can be triggered (from point U) before the grid of Th_2 becomes sufficiently positive to produce conduction. In this

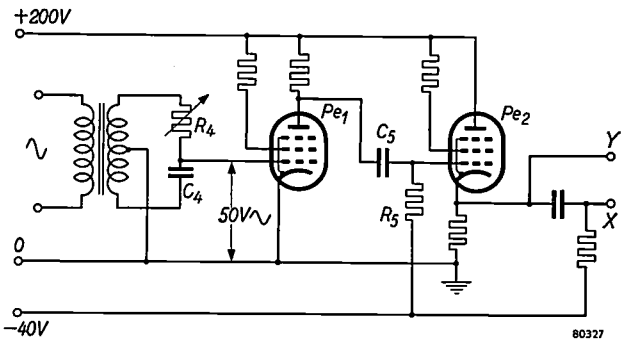


Fig. 10. Circuit for supplying triggering pulses, Pentode Pe_1 is strongly overloaded by A.C. grid voltage of 50 V, 50 c/s, (derived from phase-shifting network R_4-C_4), so that square-wave anode current is produced. This waveform is differentiated by C_5-R_5, giving positive and negative pulses. The positive pulses control a cathode-follower (pentode Pe_2). Terminal X is connected to grid of test valve (Th_1 in figs. 7*a*, 8*a* and 9), terminal Y to a flip-flop (fig. 11) firing auxiliary valve Th_2.

way the whole period of current decay and inverse voltage rise can be observed on the oscilloscope. The cathode-ray oscilloscope trigger is a necessity in order to avoid "jitter", but even so it is difficult to observe the oscilloscope trace for periods of one or two microseconds after the time of current extinction, due to the transient effects of stray capacitances and inductances.

For the capacitor-charging thyratron Th_3 the grid control circuit is much simpler. The main requirement is that it should not trigger Th_3 until Th_2 has had ample time to de-ionize; otherwise the capacitor C_2 would be short-circuited and would not charge properly. A delay of about 60° in the triggering of Th_3 is enough for this purpose, and it is obtained by the use of a simple phase-shifting network.

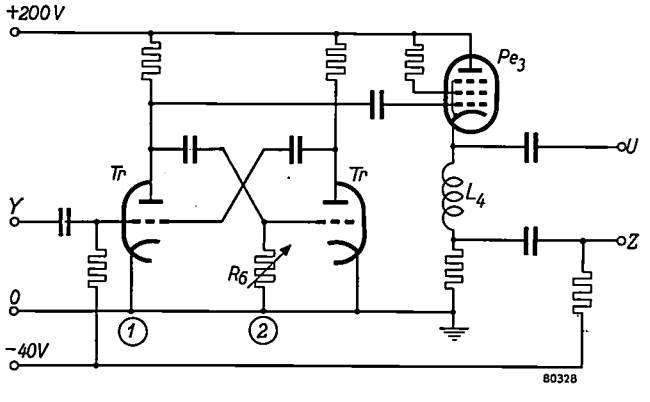


Fig. 11. Flip-flop circuit. It is stable with section *1* of *Tr* (Mullard ECC 33 double triode) non-conducting, section *2* conducting. A positive pulse at Y (from circuit shown in fig. 10) trips the flip-flop, which returns to stable position after a time variable by resistor R_6. The cathode-follower output stage contains the pentode Pe_3. The pulse appearing at terminal U triggers the oscilloscope, while the pulse at Z fires the auxiliary thyratron Th_2 (figs. 7*a*, 8*a* and 9). Choke L_4 ensures that the pulse at U appears earlier than the pulse at Z.

Results obtained with the cheater circuit

The apparatus has been used to life-test valves of type 5545 for periods up to 2200 hours. This is quite a considerable time, taking into account that the valves were run at commutation factors of 450-750 VA/μsec², i.e., several times the published valve rating (130 VA/μsec²). Special valves were made with pressure gauges attached comprising a quartz fibre, which could be set into vibration. The damping of this vibration is a measure of the gas pressure in the valve. Using such gauges, the gas pressure could be measured at intervals throughout the life of the valve.

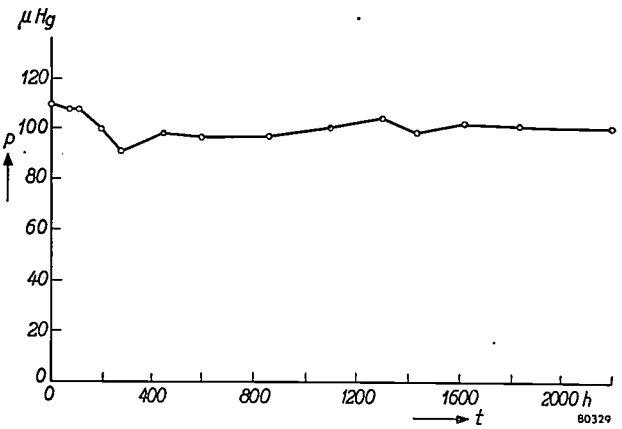


Fig. 12. Gas pressure p in xenon-filled Mullard MT 5545 thyratron during a test for gas clean-up. The first 1270 hours the valve was run with a commutation factor $f_c = 450$ VA/μsec², the rest of the time with $f_c = 750$ VA/μsec².

The results obtained with a valve of type 5545 are shown in *fig. 12*. The accuracy of the measurements is about 5%. It can be seen that a small amount of gas clean-up occurs during the first 250 hours and thereafter the gas pressure remains substantially constant. These characteristics were maintained even at commutation factors as high as 750 VA/μsec².

Work is in progress on other valve types to compare the suitability of various valve structures, but the above results already bear out the claims made for this valve construction by its designers [6]).

Summary. The life of a thyratron may be unduly shortened by gas clean-up. This effect is strongest when the commutation factor (product of rate of current decay and rate of inverse voltage built-up) is high. The need was felt, therefore, for life tests at a high commutation factor. However, conventional circuits satisfying this condition have a power consumption which with large thyratrons is almost prohibitive. This difficulty was circumvented by the use of a "cheater circuit" that simulates the desired working conditions with a great economy of power. A life test of a xenon-filled Mullard MT 5545 thyratron in such a circuit is discussed. There appears to be no appreciable gas clean-up in over 2000 hours, at commutation factors up to 750 VA/μsec², which is far in excess of the published valve rating.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the administration of the Philips Research Laboratory, Eindhoven, Netherlands.

2089: J. I. de Jong and J. de Jonge: The action of bromine on derivatives of urea in neutral aqueous solution (Rec. Trav. chim. Pays-Bas 72, 169-172, 1953).

It was found that only unsubstituted amide groups in derivatives of urea are rapidly attacked by bromine in neutral solution, while the reactions with alkylated amide groups is very slow. Dimethylol urea has a symmetrical structure.

2090: K. H. Klaassens and C. J. Schoot: Derivatives of p-diethoxybenzene, II. 1,4-diethoxy-2-aminobenzene-5-diazoniumborofluoride (Rec. Trav. chim. Pays-Bas 72, 178-182, 1953).

The preparation of the compound named above is described and its structure is confirmed.

2091: Y. Haven: Dielectric losses in sodium chloride crystals (J. chem. Phys. 21, 171-172, 1953, No. 1).

By plotting log (tan δ) against log frequency it is possible to distinguish between dipole losses (Debye losses) and conductivity losses. Measurements were made at 100 °C on very pure NaCl and on contaminated NaCl (0.003 mole percent divalent ions, 0.05 mole % Ca^{2+}, 0.14 mole % Ca^{2+}, 0.008 mole % Mn^{2+}, and 0.014 mole % Mn^{2+}). The Debye losses appear to increase with increasing contamination.

2092*: H. P. J. Wijn: Magnetic relaxation and resonance phenomena in ferrites (Thesis, Leiden 1953).

After a short review of the preparation and properties of sintered ferrites (Ch. 1), the measuring methods are described that have been used to determine the magnetization curves and the distortion as a function of frequency. The conditions that must be fulfilled in order to limit the systematic error in the measuring results are discussed. For ferrites with a high initial permeability the magnetization curve depends on frequency such that, up to rather high inductions, there is a linear relation between the induction and the magnetic field in the ferrite. This frequency dependence is not found for ferrites having an initial permeability below about 400, unless they are fired at a very high temperature (1500 °C). In Ch. 4 the influence of porosity and external stresses on the magnetization curve is examined, and the conclusion is reached that when (at high frequency) a magnetization curve becomes a straight line, the irreversible domain-wall displacements no longer take place. Ch. 5 deals with the frequency dependence of the initial permeability of ferrites and it was established that at high frequency a resonance phenomenon underlies the decrease of the permeability. Moreover a ferromagnetic resonance of the type described by Snoek is found for a ferrite under a large external stress such that only simultaneous spin rotations make a contribution to the permeability. The resonance frequency is related in a simple way to the natural ferromagnetic resonance frequency of the same ferrite. In this chapter some results of measurements are also given that point to a new relaxation phenomenon, which is largely responsible for the residual losses in ferrites.

After a brief survey of the theories found in the literature concerning the dispersion mechanism of domain wall displacements and spin rotations in Ch. 6, the next chapter compares the measured results with these theories. The following conclusions are reached:

a. The relaxation of irreversible domain-wall displacements is brought about by a friction process characterized by a friction coefficient having the order of magnitude of unity. The origin of this friction is unknown.

b. The theory of ferromagnetic resonance permits many of the properties of the initial permeability at high frequency to be clarified.

c. The relaxation process causing a dispersion in the initial permeability of ferrites at low frequency is closely connected with an electron diffusion. The activation energy of this relaxation process is the same as that found from the temperature dependence of the resistivity of ferrites.

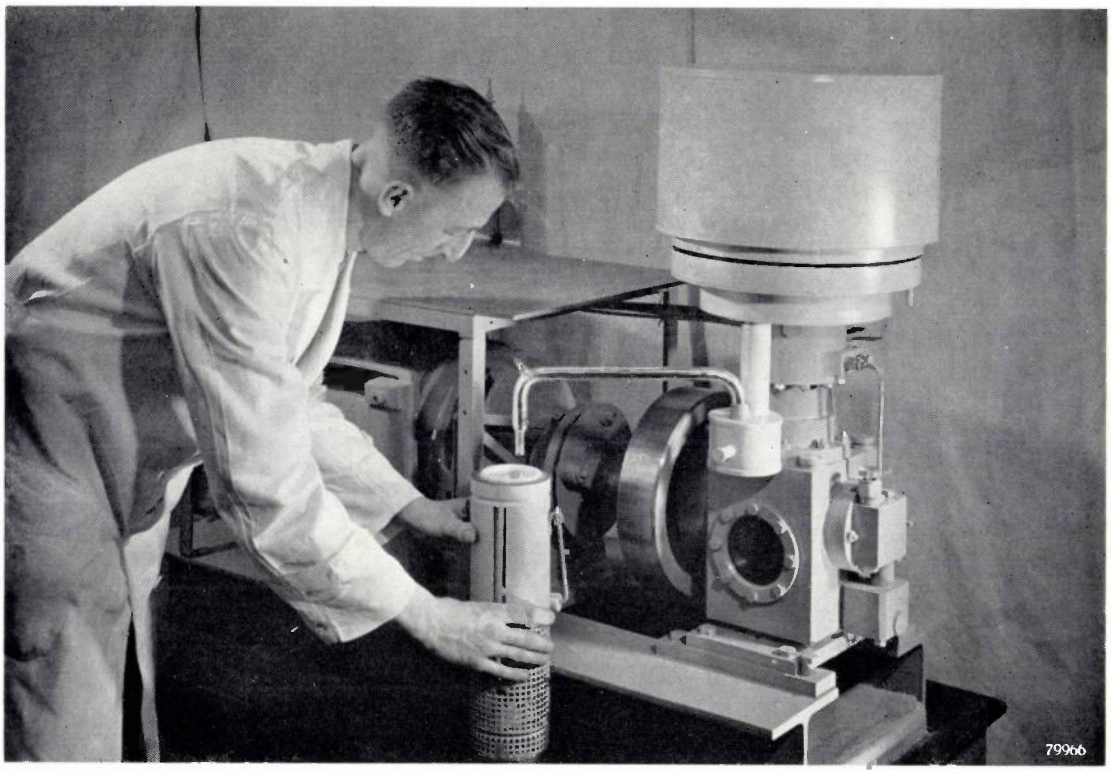
Though in some ferrites at high frequency B is a linear function of H, it is seen in the last chapter that many losses still occur in the ferrite, giving rise to a phase shift between field and induction. Finally it is shown that in the case of ferrites, the determination of the distortion from the hysteresis loss can give erroneous results.

VOL. 16 No. 4, pp. 105-140 OCTOBER 1954

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS



CONSTRUCTION OF A GAS REFRIGERATING MACHINE

by J. W. L. KÖHLER and C. O. JONKERS.

621.573

A theoretical study of the gas refrigeration cycle, recently published in this Review, was based on research in the Philips laboratories in Eindhoven closely parallel to work on the hot-gas engine. The present article deals with the application of this cycle and describes a practical machine for its realization. The results achieved with this machine will be discussed, especially with a view to possible applications of this new refrigeration cycle.

For the practical application of the gas refrigeration cycle — the principles of which were discussed in the previous issue of this Review [1]) — a number of problems have first to be solved. Before discussing these problems it may be useful to recapitulate briefly the principle of the gas refrigeration cycle.

A quantity of gas is compressed at room temperature, after which it is cooled to a low temperature. At this temperature it is permitted to expand and the cold thus produced is utilized. After the expansion, the gas is re-heated to room temperature, and the cycle is completed. The cooling and re-heating of the gas takes place by an exchange of heat in a regenerator. The working fluid is at all times in the gaseous state and will be considered here as a perfect gas.

[1]) J. W. L. Köhler and C. O. Jonkers, Fundamentals of the gas refrigerating machine, Philips tech. Rev. **16**, 69-78, 1954/55 (No. 3), hereafter referred to as I.

As mentioned in I, it has been found in the course of research that machines based on this principle have the most favourable properties when used to produce cold between —80 °C and —200 °C. The existence of this optimum range, however, cannot be derived from the discussion of the idealized process in I. This follows from the fact that the efficiency of the ideal cycle (no losses) is at all temperatures equal to that of a Carnot cycle working within the same temperature limits; *as regards efficiency therefore, the ideal process cannot be surpassed at any temperature.* The provision that the working fluid must behave as a perfect gas to a sufficient approximation, does not constitute a restriction even at —200 °C, for it can be satisfied by using hydrogen or helium even at considerably lower temperatures.

Limitations of the process

It has been found that the limitation of the process to the above temperature range for efficient working is due to deviations from the ideal cycle which occur in practice. They result in an increase of the required shaft power and a decrease in the refrigerating capacity; the smaller the values of these quantities, the more serious are the relative effects of the losses. *Fig. 1* shows that we may expect that *increase of the required shaft power limits the process at high temperature while decrease of the refrigerating capacity limits it at low temperature.* In this way we arrive at the aforementioned temperature range, i.e.

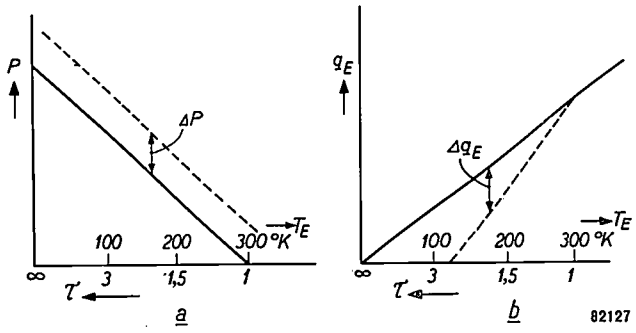


Fig. 1. Graphs of the shaft power P and the refrigerating capacity q_E of the gas refrigerating machine as functions of the temperature ratio $\tau = T_C/T_E$ (T_C = temperature of the cooler $\approx$ 300 °K, T_E = the desired low temperature). The full lines apply to the ideal machine (see fig. 10 in I), the dotted line to a practical machine subject to various losses.

the optimum working range, in which the actual efficiency differs least from the ideal efficiency — see *fig. 2*. The limits of this range are not fixed precisely, since they greatly depend on present and future technological possibilities. We shall now examine in more detail these deviations from ideal behavior which give rise to the optimum range.

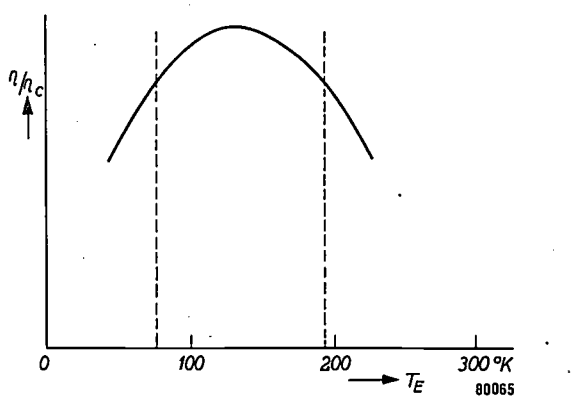


Fig. 2. The "figure of merit" η/η_c of the gas refrigerating machine as a function of the temperature (η = efficiency of the actual machine, η_c = Carnot-efficiency, cf. I). The diagram shows how the influence of the losses limits the useful temperature range of the machine.

Losses increasing the shaft power

The drive of every machine is subject to frictional losses, so that a certain amount of extra power is required merely to keep the machine going. In the gas refrigerating machine the absolute value of this "mechanical loss" is practically independent of the freezing temperature [2]), as may be seen from the dotted curve in fig. 1*a*; the relative effect of this loss is therefore greatest at a high temperature.

Another effect of a similar character to the mechanical loss is what is termed the "adiabatic loss". In I it was assumed that in the cylinders the heat transfer between the gas and the surrounding walls is so complete that the heat of compression and the cold of expansion can be discharged during every phase pof the process. This, however, is very difficult to realize. For this reason heat exchangers have been incorporated between the cylinders and the regenerator (a "cooler" at the compression side and a "freezer" at the expansion side), which establish the thermal contact between the interior of the machine and the outside. In these heat exchangers the refrigerant gas is caused to flow through narrow channels, so that a good thermal contact with the walls is attained.

If we assume the thermal contact in these heat

[2]) The pressure ratio, which determines the forces and hence the losses set up in the drive, is in fact, only slightly influenced by the freezing temperature. Incidentally it is to be noted in this connection that in the compression refrigerator the mechanical losses will increase as the evaporator temperature decreases.

exchangers to be perfect, then we obtain a situation as shown in *fig. 3*, in which the temperatures of the various parts of the machine have been indicated schematically. The temperature of the cooler is determined by the temperature of the cooling water, and the freezer temperature by the temperature at which the cold is to be utilized, e.g. the boiling point of air. The gas enters the cylinders at the temperature of the heat exchangers; after this the gas temperature varies adiabatically with the pressure inside the machine. This means that the average temperature T_{Ca} in the compression cylinder will be higher than the temperature T_C of the cooler (due to compression of the gas), whilst the average temperature T_{Ea} of the expansion cylinder will be lower than T_E.

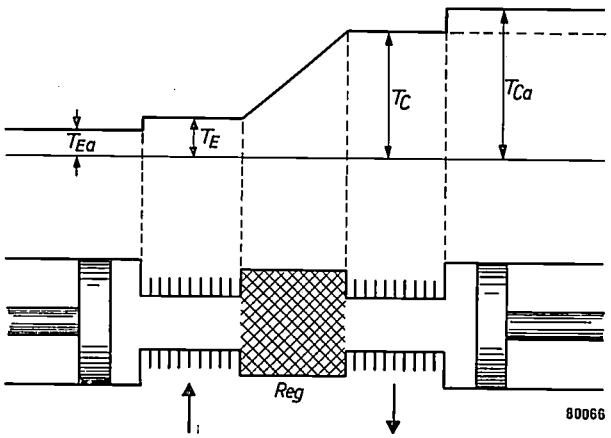


Fig. 3. Temperature distribution in the gas refrigerating machine. The machine is equipped with two heat exchangers effecting the thermal contact between the refrigerant and the ambient atmosphere. The first (the "freezer") is at the low temperature T_E, the second (the "cooler") is at room temperature T_C. Between the two, in the regenerator (*Reg*), there is a gradual transition between the two values. The two arrows indicate the direction of flow of heat in the two cases.

Owing to the adiabatic behaviour of the refrigerant in both the expansion and the compression cylinder, the average temperature T_{Ea} in the former is slightly lower than T_E, whilst the average temperature T_{Ca} in the latter is slightly higher than T_C. This is the cause of the "adiabatic loss".

We then see that $T_{Ca}/T_{Ea} = \tau_a > \tau$, which means that the machine works internally at a greater temperature ratio than externally. From fig. 1*a* it can be concluded that the shaft power is then greater than in the isothermal case. It can be shown very simply that the value of this additional shaft power is very little influenced by the freezing temperature, so that the effect of this loss, as for frictional losses in the drive, is relatively greater at high temperatures.

By analogy with the expression derived for the shaft power in the isothermal process, $P_{iso} = q_E(\tau-1)$, (cf. I, eq. 14), we can put:

$$P_a \approx q_E(\tau_a - 1).$$

The value of the refrigerating capacity q_E is found to be practically the same for both processes, owing to the fact that the decrease due to the greater temperature ratio, as would be expected according to fig. 1*b*, is compensated by the greater pressure ratio inherent in the adiabatic process. The additional shaft power thus amounts to:

$$\Delta P_a = P_a - P_{iso} \approx q_E(\tau_a - \tau) = q_E\tau\left(\frac{\tau_a}{\tau} - 1\right).$$

The ratio τ_a/τ depends very little on the freezing temperature T_E, because the pressure ratio changes very little with the temperature; we can therefore put:

$$\frac{\tau_a}{\tau} \approx 1 + \beta,$$

where β is independent of T_E. Consequently

$$\frac{\Delta P}{P_{iso}} \approx \frac{q_E\tau\beta}{q_E(\tau-1)} = \frac{\beta T_C}{T_C - T_E}.$$

This makes it clear that the relative effect of adiabatic loss is greatest when T_E differs least from T_C.

The adiabatic loss is approximately proportional to the pressure ratio. It can therefore be limited by reducing the pressure ratio. This, however, impairs the refrigerating capacity, so that it must not be carried too far.

Losses reducing the refrigerating capacity

The cold parts of the machine can never be completely protected against loss of cold (or influx of heat) through conduction. The lower the freezing temperature, the greater becomes the influence of this "insulation loss".

If G is the thermal conductivity of the particular section causing the loss, then the loss of cold per unit time amounts to;

$$\Delta q_G = G(T_C - T_E) \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

This loss should be compared with the refrigerating capacity q_E, which is approximately proportional to T_E, i.e. $q_E \approx CT_E$. Hence

$$\frac{\Delta q_G}{q_E} \approx \frac{G}{C}\frac{T_C - T_E}{T_E} = \frac{G}{C}(\tau - 1). \quad . \quad . \quad (2)$$

The relative influence of the insulation loss at very low temperatures is therefore practically inversely proportional to the freezing temperature T_E (cf. the dotted line in fig. 1*b*); this effect is mainly due to the decrease of the refrigerating capacity with T_E.

The insulation losses through conduction play only a minor part in the gas refrigeration machine, as the cold parts can be built compactly. The losses

caused by a non-perfect regenerator, however, will also have the character of insulation losses, and it is these losses that are more difficult to combat. *This "regeneration loss" is the major limitation of the refrigerating capacity at very low temperatures* It can be roughly estimated as follows:

The quantity of heat Q_r which the gas has to discharge on its way from the compression space to the expansion space, is stored in the mass of the regenerator. For this a loosely packed mass of thin wire is used. This mass of wire shows a continuous temperature change in the direction of the gas flow; the form of this temperature distribution is roughly represented in fig. 3.

The value of Q_r is given by the equation

$$Q_r = W_g(T_C - T_E), \quad . \quad . \quad . \quad . \quad . \quad (3)$$

where W_g represents the thermal capacity of the gas flowing through the regenerator per half cycle. In practice, however, slightly less than the full amount of heat Q_r is transferred to the regenerator mass, viz. $\eta_r Q_r$, η_r being the efficiency of the regenerator ($\eta_r = 1$ for ideal regeneration). The difference between Q_r and $\eta_r Q_r$, which is denoted by ΔQ_r, is the regeneration loss:

$$\Delta Q_r = Q_r - \eta_r Q_r = (1-\eta_r)\, W_g\, (T_C - T_E), \quad (4)$$

or, per second:

$$\Delta q_r = w_g\, (1-\eta_r)\, (T_C - T_E), \quad . \quad . \quad . \quad (4a)$$

where w_g represents $nW_g/60$, and n is the number of r.p.m. of the machine.

This regeneration loss wastes part of the cold produced and thus reduces the refrigerating capacity. If we compare (4a) with (1), we notice that the regeneration loss has formally the character of an insulation loss if $(1-\eta_r)w_g$ is considered as the thermal conductivity of the regenerator.

The relative regeneration loss now becomes:

$$\frac{\Delta q_r}{q_E} \approx \frac{w_g}{C}\,(1-\eta_r)\,\frac{T_C - T_E}{T_E} = \frac{w_g}{C}\,(1-\eta_r)\,(\tau - 1). \quad (5)$$

w_g/C is very little dependent on T_E; its value is approximately 7. The following *table*, computed for a value of $\eta_r = 0.99$ (which can actually be obtained in practice), shows the influence of the regeneration loss in various cases).

Temperature range	T_E in °K	$\frac{\Delta q_r}{q_E}$ in %
liquid air	75	21
liquid hydrogen	20	98
hot-gas engine	900	4.7

At the boiling point of air the loss (21%) is acceptable. At 20° K, however, the loss has risen to 98%, so that the refrigerating capacity is almost completely vitiated by the regeneration loss. For comparison, the value for the hot-gas engine has also been given; this shows that there the regeneration problem is less important than in the refrigerator.

There are two main factors which cause the efficiency of regenerators to deviate from unity, viz, imperfect heat transfer between the gas and the regenerator mass and the finite heat capacity of this mass. Because of the imperfect heat transfer the gas will not exactly follow the temperature of the regenerator mass, so that it is not cooled down or warmed up to the correct temperature. Because of the finite heat capacity of the regenerator, its temperature changes while the gas flows through it; this has the same effect.

The relationship between heat transfer, heat capacity and the efficiency of regenerators has been dealt with by *Hausen* [3]).

Miscellaneous losses: dimensioning of the gas refrigerating machine

Apart from the losses already dealt with, which limit the field of application of the gas refrigeration cycle, there are some other causes that impair the efficiency in practice, which we shall briefly discuss now.

In fig. 3 it is explicitly assumed that the cooler and the freezer are ideal. If this is not the case, then temperature differences occur, both at the inside and at the outside of the heat exchangers, and in addition, a temperature difference arises between inside and outside, owing to the thermal resistance of the material of the heat exchanger. In this way the gas temperature in the cooler becomes higher and that in the freezer lower than required, so that the machine has to operate internally at a higher temperature ratio. This loss is somewhat similar to the adiabatic loss, though of less importance [4]).

Consequently, in designing a gas refrigerating machine, the aim will be to attain the largest possible transfer of heat in cooler, regenerator and freezer. The attainment of the highest heat-transfer however, involves new difficulties, since the components mentioned each constitute a resistance to the gas flow from the compression space to the expansion space and back. To overcome this flow-resistance a certain difference in pressure between the two spaces is required, for which additional shaft power has to be supplied, and which results in a reduction of the refrigerating capacity. This "flow loss" is

[3]) H. Hausen, Z. angew. Math. Mech., 9, 173-200, 1929.

[4]) Contrary to the adiabatic loss, however, this loss does impair the refrigerating capacity.

intimately connected with the extent of the heat transfer and with the size (total diameter) of the ducts. It will be clear that in practice a compromise is necessary, aiming at a reasonable heat transfer at acceptable values of the flow loss and the dead space.

It is not easy to give clean-cut directions for determining the values of the parameters relevant to the refrigerating machine; moreover, the choice is, as always, influenced by the designer's personal preference. We must, therefore, confine ourselves to some general observations.

In practice the machine must be designed to satisfy the demand for a certain refrigerating capacity at a prescribed freezing temperature. The refrigerating capacity, which is the output of cold per second, has been given by formula (13) in I, which reads:

$$q_E = 5.136\ \bar{p}\ V_0 \frac{\delta}{1+\sqrt{1-\delta^2}} \sin\Theta \frac{n}{1000}\ \text{watt}, \quad (6)$$

where

$$\tan\Theta = \frac{w \sin\varphi}{\tau + w\cos\varphi}.$$

In the expression (6) only the variables $\bar{p}$ (average pressure), V_0 (volume of the expansion space) and n (number of r.p.m.) can be chosen without restriction. τ is prescribed by the freezing temperature, whilst δ, φ, and w are more or less fixed for a well-designed refrigerator, viz. $0.3 < \delta < 0.4$, $60° < \varphi < 120°$, and w has the order of magnitude 1.

In view of the fact that the refrigerator should preferably be as small as possible for a given capacity (and hence V_0 should be small), $\bar{p}$ and n have to be selected as high as possible. In raising n one is handicapped by increasing losses (mechanical loss and flow loss) and by criteria concerning the operating life of the machine. In practice, therefore, it is only the value of $\bar{p}$ that can be freely varied. *Increasing the pressure level, is indeed a most effective means of drastically reducing the size of the gas refrigerating machine,* contrary to the evaporation refrigerator, where this cannot be done. The raising of the working pressure is limited by mechanical considerations, such as the required strength of the partition walls and the load applied to the drive.

Still another factor precludes undue raising of the pressure level. According to formula (6) the refrigerating capacity should increase linearly with rising pressure. In practice, however, losses occur which cause the refrigerating capacity to increase less than linearly. The higher the pressure, the greater the relative influence of the loss, until finally the refrigerating capacity will even decrease if the pressure is made still higher (cf. *fig. 4*). It was because of this effect that our experimental machines were initially unable to reach temperatures below −160 °C (the loss increases at lower freezing temperature), a behaviour which baffled us for a long time. We finally came to the conclusion that the interaction between the regenerator and the rest of the machine is responsible for this behaviour. As already mentioned, the average temperature of

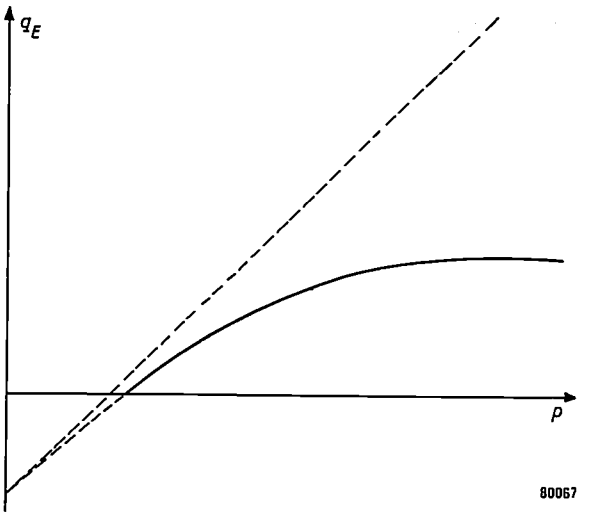


Fig. 4. Variation of the refrigerating capacity q_E with the pressure p in the refrigerator. According to the theoretical formula (equation) (13), part I) the dotted straight line would be expected. The measured curve deviates from this straight-line as shown by the full ciurve. Clearly there is a surprising decline of the refrigerating capacity at increasing pressure in the high-pressure range.

the regenerator mass varies periodically as a result of the periodic accumulation and discharge of heat. This affects the normal pressure cycle of the machine and results in a smaller refrigerating capacity (as well as decreased shaft power). This effect is governed by the ratio of the thermal capacity of the regenerator mass to that of the gas contained in it; the smaller the value of this ratio, the more is the linearity impaired. The ratio becomes smaller as the denominator increases, which happens if the pressure level in the machine is raised or the freezing temperature lowered: in both cases the quantity of gas in the regenerator is increased. These considerations adequately explain the non-linear behaviour, and calculations of the effect agreed well with the actual measurements. According to this explanation, the effect can be minimized by choosing the highest possible thermal capacity per cm³ for the regenerator mass. This requirement is in addition to that of classical regenerator theory (cf. footnote [3]), according to which the total thermal capacity of the regenerator must exceed a certain prescribed value.

It is now convenient to deal with the actual refrigerating machine developed in the Eindhoven laboratories. The way in which the volume variations are effected in this machine, is somewhat different from what has been described in I. This type of machine is termed a "displacer machine"; its working principle is outlined below.

The displacer machine

It has been found that the displacer-type of mechanism, which was used in some of the old hot-air engines, has certain particular advantages for the refrigerator. The principle is illustrated in *fig. 5*. The volume variations are no longer obtained with the aid of two equivalent pistons, but by means of a main piston and an auxiliary piston, which is termed the "displacer". The main piston *1* moves in a cylinder *2* and varies the volume of the entire working space. This working space is divided by the displacer *3*, which like the main piston, has a harmonic motion: there is thus a space *4* between main piston and displacer, and a space *5* above the displacer, both spaces varying harmonically. The displacer motion is such that space *4* is lagging in phase with respect to space *5*, so that space *4* is the compression space and space *5* is the expansion space (cf. I); this is illustrated in the graph in *fig. 6*.

The spaces *4* and *5* are in open communication with one another via the annular heat exchanger surrounding the displacer. The gas pressures above and below the displacer are nearly the same (hence

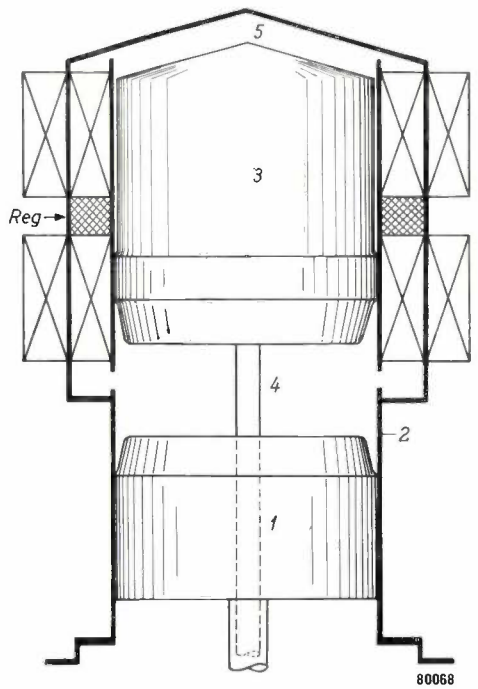


Fig. 5. Principle of the displacer machine. *1* = main piston, moving in cylinder *2*. *3* = displacer, causing a periodic flow of gas back and forth between the spaces *4* and *5*. The gas then flows through the annular heat exchangers surrounding the displacer.

the name "displacer"). Owing to this, there is only a slight leakage of gas from the expansion space so that the accompanying loss of cold is very slight, which is a substantial advantage of the

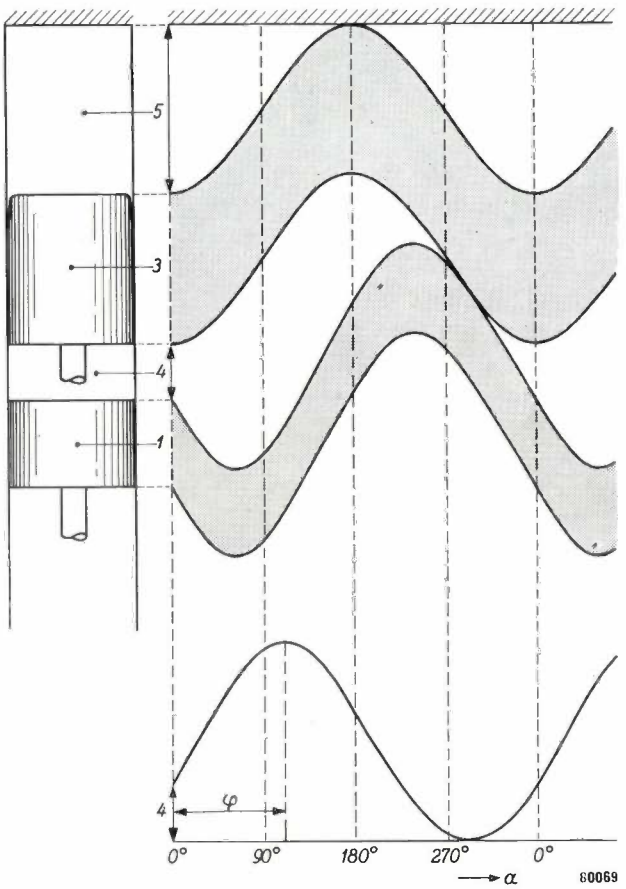


Fig. 6. Variations of the volumes of spaces *4* and *5* of fig. 5 as functions of the shaft angle α. The constant volumes taken up by the displacer body (*3*) and by the body of the main piston (*1*) are shown in grey. The fixed head of the cylinder is shown shaded. In the lower part of the diagram the volume of space between displacer and master piston has been plotted separately, in order to demonstrate its sinusoidal variation and the phase shift with respect to space 5.

displacer-type machine. In a machine having two normal pistons this loss of cold is considerably greater, due to the far greater pressure differences. Another advantage of the displacer machine is the far smaller mechanical loss, thanks to the small frictional loss of the displacer.

Description of the gas refrigerating machine

The machine is shown in *fig. 7*. The components already mentioned are marked by the same numbers. The main piston is driven, via two parallel connecting rods *6*, by the cranks *7* of the crankshaft *8*. The displacer rod *9* passes through the centre of the main piston to the crankcase, where it is coupled, via the connecting rod *10*, to a third crank *11* of the crankshaft. The angle between the cranks *7* and *11* has been so chosen that the motion of the displacer has the desired phase difference with respect to that of the main piston. The gas flows out of the compression space through the ports *12* to the

space containing the cooler, the regenerator and the freezer; the upper end of the freezer communicates with the expansion space.

The displacer consists of a piston body *16* and the "cap" *17*. The piston body carries piston rings and fits in the cylinder liner as a normal piston, and has

This construction, the principle of which was applied in the earliest hot-air engines, provides an elegant solution of the problem of how to seal a cold space by means of a moving piston at the ambient temperature, without loss of cold. It is remarkable that the designers of expansion engines for the liquefaction of air have overlooked this solution, which could have saved them many a difficulty.

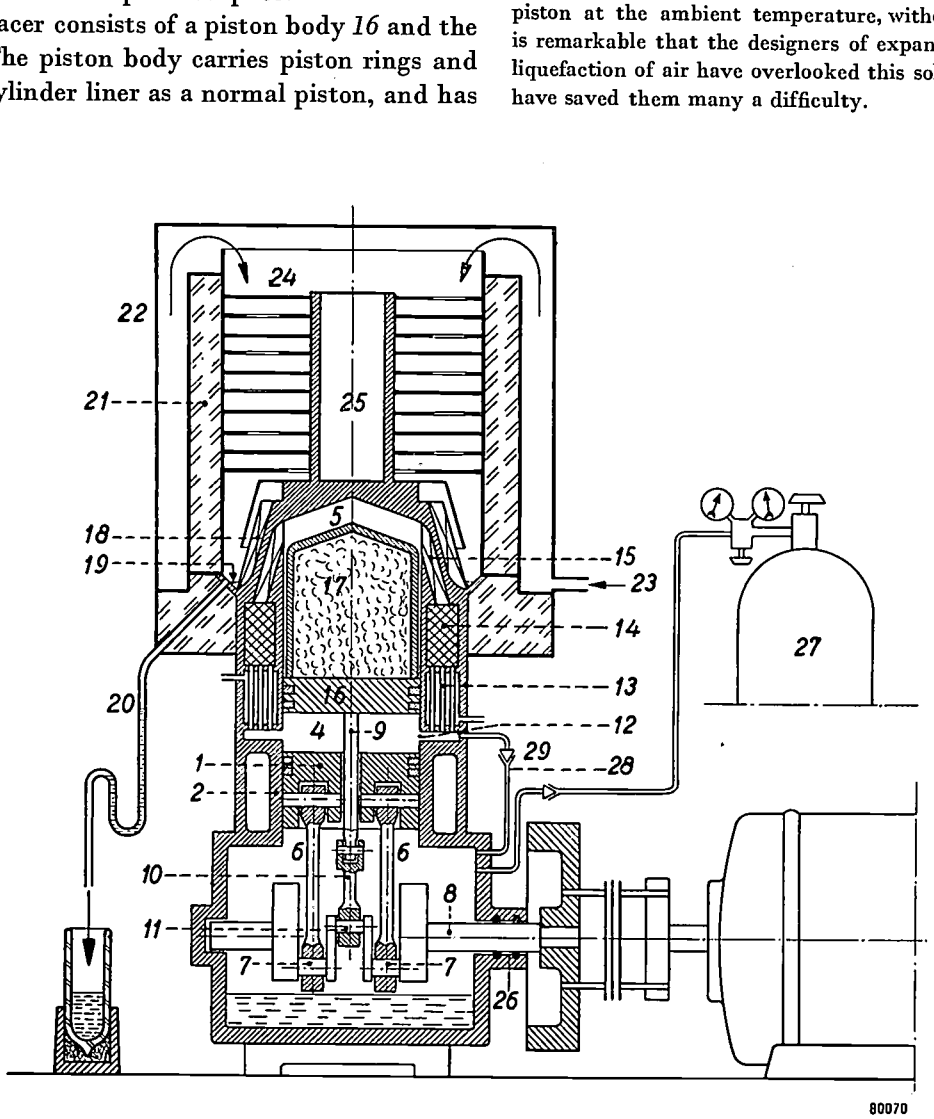


Fig. 7. Simplified cross-section of a gas refrigerating machine designed for the liquefaction of air. The figures *1* - *5* have the same meaning as in fig. 5. Further figures represent: *6* = two parallel connecting rods with cranks *7* of the main piston. *8* = crankshaft. *9* = displacer rod, linked to connecting rod *10* and crank *11* of the displacer. *12* = ports. *13* = cooler. *14* = regenerator. *15* = freezer. *16* = displacer piston and 17 = cap. *18* = condenser for the air to be liquefied, with annular channel *19*, tapping pipe (goose-neck) *20*, insulating screening cover *21*, and mantle *22*. *23* = aperture for entry of air. *24* = plates of the ice separator, joined by the tubular structure *25* to the freezer (*15*). *26* = gas-tight shaft seal. *27* = gas cylinder supplying refrigerant. *28* = supply pipe with one-way valve *29*.

about the same temperature as the liner, which is surrounded by cooling water. The cap is made of a heat-insulating material and is filled with a loose woolly substance in order to preclude gas circulation within the cap. The cap has a slightly smaller diameter than the piston body, so that it does not touch the cylinder. All these provisions considerably reduce the loss of cold through conduction to the warm parts of the machine.

The outside of the freezer forms the "condenser" *18*, against which the air can condense. The liquified air is collected in an annular channel *19* and can be tapped via a pipe *20*. The condenser is surrounded by an insulating screening cover *21* and a mantle *22*. Fresh air can enter through the aperture *23* and flow through holes in the plates *24*, which are thermally connected via the tubular structure *25*, to the freezer. The water vapour and the carbon dioxide

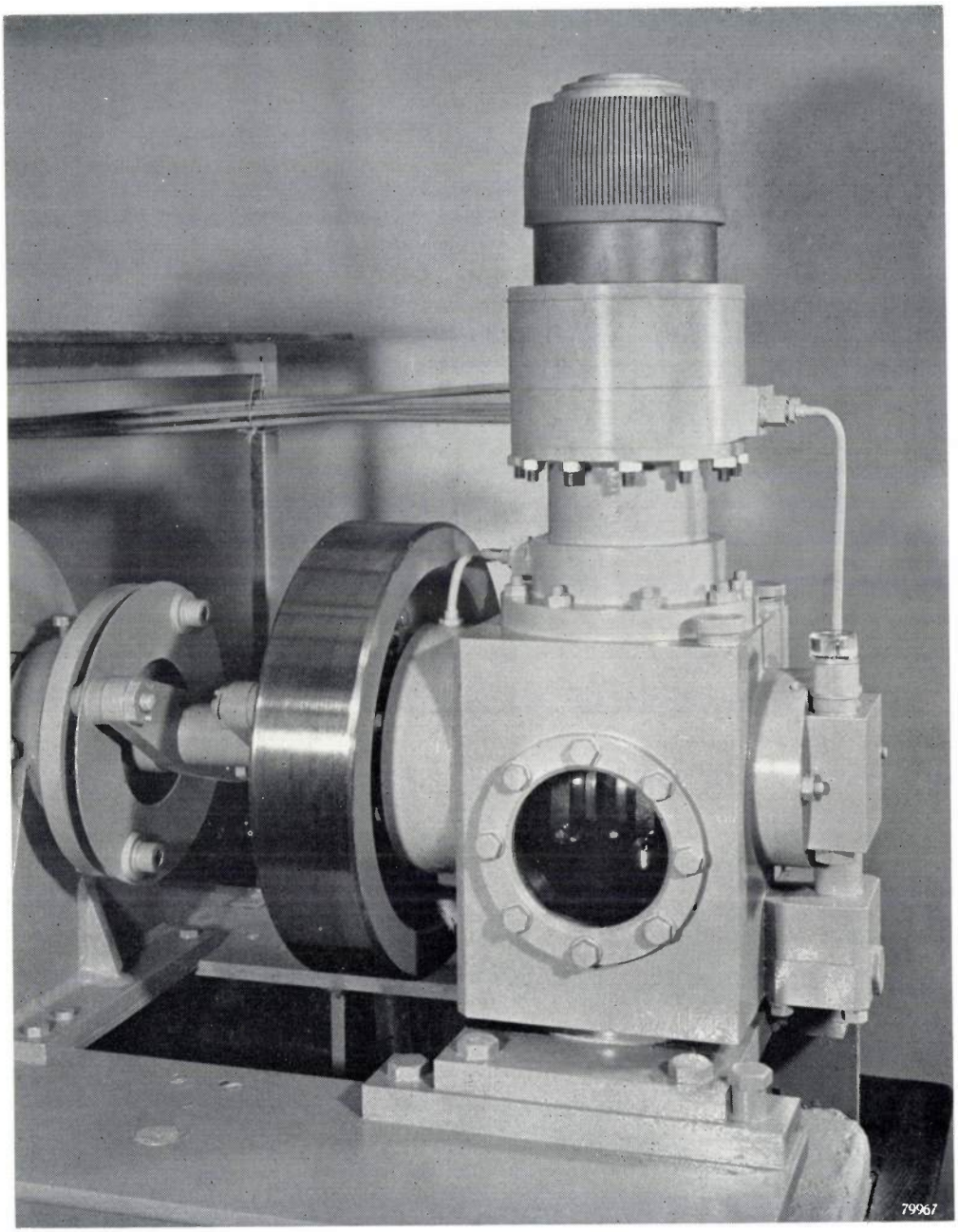


Fig. 8. Photograph of the gas refrigerating machine.

contained in the air will precipitate as a frost on these cold plates (termed the "ice separator"), so that the condenser *18* is not contaminated by ice or solid CO_2.

The complete refrigerator is shown in *fig. 8*. *Fig. 9* is a photograph of the cooler, the regenerator and the cylinder head. We shall now proceed to discuss some of the details of the machine.

Constructional features

The crankcase is closed and contains the gas serving as the refrigerant; the gas pressure is approximately equal to p_{min} in the working space. The crankshaft is led out via a shaft seal (fig. 7, *26*). The filling gas is supplied from the cylinder *27* to the crankcase; from there, pipe *28* with one-way valve *29* leads to the working space.

Whenever leakage round the main piston (which reduces the pressure level in the working space) causes p_{min} to drop below the pressure in the crankcase, the gas will flow back to the working space through *28*. In principle the machine is gas-tight; and only if incidental leakage has caused the gas pressure to fall below the necessary level is the machine replenished from the supply cylinder.

The crankcase has to be of comparatively heavy

construction to cope with the high gas pressure applied. A pressurized crankcase offers some substantial advantages over a crankcase at atmospheric pressure. With a pressure p_{min} in the crankcase, the pressure difference exerted on the main piston is

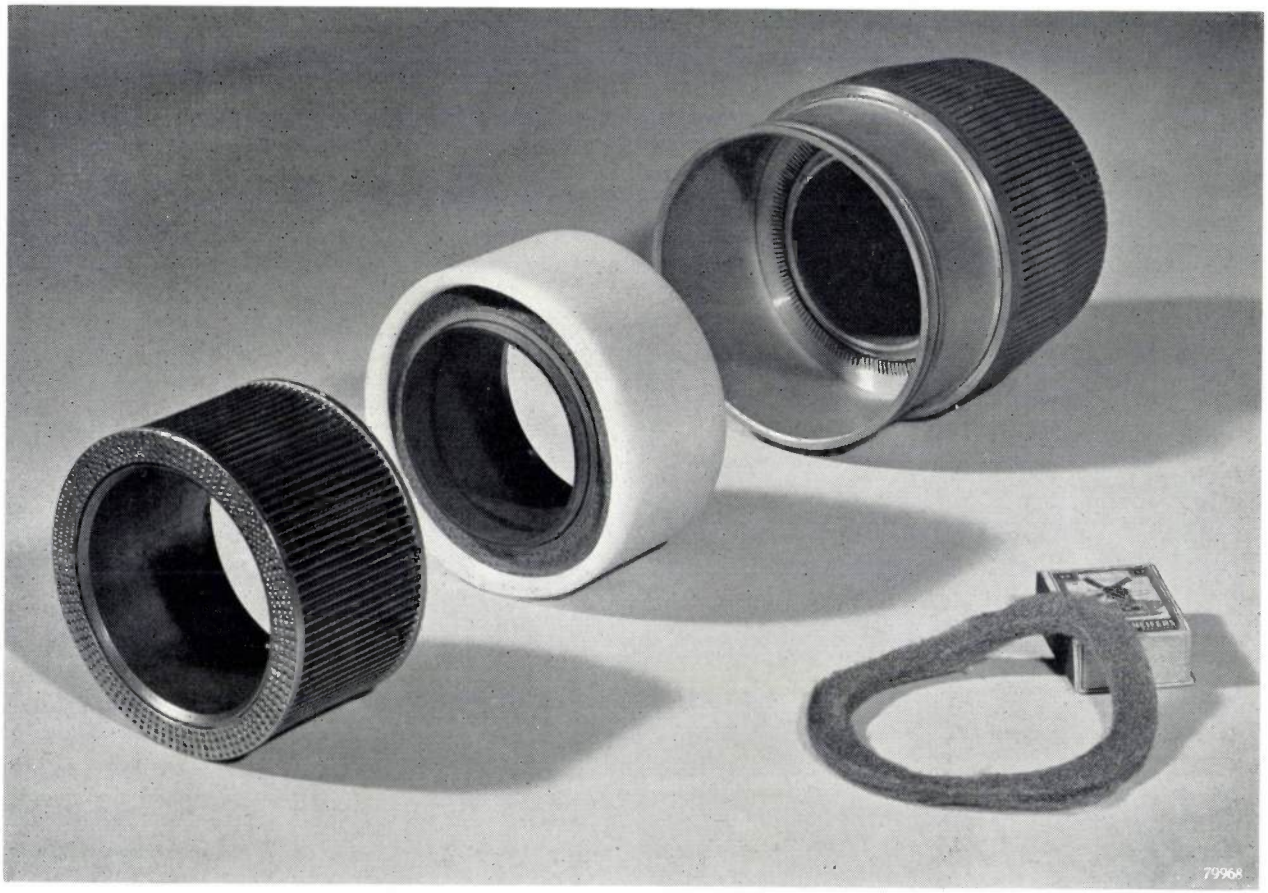


Fig. 9. Left to right: cooler, regenerator and cylinder head (of which the freezer and the condenser form integral parts) of the gas refrigerating machine.

The cooler is built-up of a large number of thin pipes, parallel to the axis of the cylinder. The refrigerant gas flows through these pipes, which are surrounded by the cooling water; this water also cools that part of the cylinder wall along which the displacer moves (see figs. 5 and 7).

The regenerator, like that of the hot-air engine, consists of a mass of extremely fine metal wire. Several rings of this material (one ring is shown on the right) are stacked together and kept in shape by two concentric rings of heat-insulating material (in this case nylon). The outer wall of the machine at the location of the regenerator should also be a poor conductor of heat in the axial direction to avoid "short-circuiting" of the regenerator. It has therefore been made of a steel of great strength but poor thermal conductivity, which, although extremely thin, is capable of withstanding the pressure inside the refrigerator.

The freezer is formed by the massive cylinder head, in which a large number of very fine slots have been cut. The outside of the cylinder head (i.e. the condenser) is also provided with a large number of slots in order to improve the heat exchange with the air to be liquefied. The cylinder head is made of copper to ensure the slightest possible temperature difference between the exterior and interior of the refrigerator.

which would otherwise not be possible. Finally it provides the means of drastically curbing leakage of gas from the machine since, unlike a piston, a rotating shaft can be provided with a perfectly gas-tight seal.

far smaller (varying between 0 and $(p_{max} - p_{min})$), than in case of operation under atmospheric pressure, when the difference would vary between p_{min} and p_{max}. The forces acting upon the drive and moving parts are thus smaller and so are the frictional losses. The gas leakage round the piston is likewise less; also the amount of gas thus lost is automatically returned to the working space,

A particularly tricky problem was the fact that the working space had to be completely free of oil, since this would freeze in the colder parts of the regenerator and clog them up. After extensive research it was found possible to give the piston such a shape that despite ample lubrication of all parts subject to friction, no oil can enter the working space. This subject will be dealt with in a separate article.

The pipe from which the liquid is tapped off forms a goose-neck, thus providing a liquid seal. Non-purified air is, therefore, prevented from entering along this pipe and contaminating the condenser. Liquid air can be tapped off notwithstanding the fact that the gas pressure around the condenser is slightly lower than that of the ambient atmosphere because of the flow resistance between the inlet opening and the condenser; the liquid in pipe *20* simply assumes a level above the highest point of the goose-neck. As a consequence of this, fresh air is, as it were, sucked into the machine at just the rate at which it can be condensed (i.e. in accordance with the refrigerating capacity). *No special means are thus required to supply the condenser with air; the refrigerator itself sucks in the required quantity*. As a result of this, the temperature of the condenser is fixed at the condensation point of atmospheric air, viz. — 194 °C. This clearly demonstrates the essential simplicity of the installation, owing to the fact that the air is condensed at atmospheric pressure.

Some data and results

The following list provides some data on the machine:

Cylinder bore	70 mm
Piston stroke	52 mm
Crankshaft speed	1440 r.p.m.
p_{max}	35 kg/cm²
p_{min}	16 kg/cm²
p_{max}/p_{min}	2.2.

The refrigerant used is hydrogen or helium; air is obviously out of the question in view of the fact that at any excess pressure it would be liquefied above the normal boiling point (at 35 kg/cm² air liquefies at — 144 °C).

In the laboratory, the following measurements were made when using cooling-water at a temperature of 15 °C:

Yield	with dry air . . .	5.8 kg/h
	with moist air . . .	4.8-5.8 kg/h
Shaft power		5.8 kW
Specific shaft power [5] . .		1.0 kWh/kg of air
Starting-up period		approx. 13 minutes
Period of continuous operation	dry, CO_2-free air	several days
	moist air [6]	20-30 hours

[5]) For liquefiers, kWh/kg is a familiar measure of the "efficiency" of the installation; in order to avoid ambiguity concerning the word "efficiency", however, we have introduced the designation "specific shaft power" for this quantity (which may be read as kW per kg air/hour). It applies to the liquefaction of dry air.

[6]) The period of continuous operation depends on the volume of the ice separator. Whenever this is clogged up, it has to be defrosted, which takes, inclusive of the new starting-up period, approximately 1 hour.

Fig. 10 shows the measurements for higher freezing temperatures; the diagram also shows the figure of merit curve (cf. fig. 2).

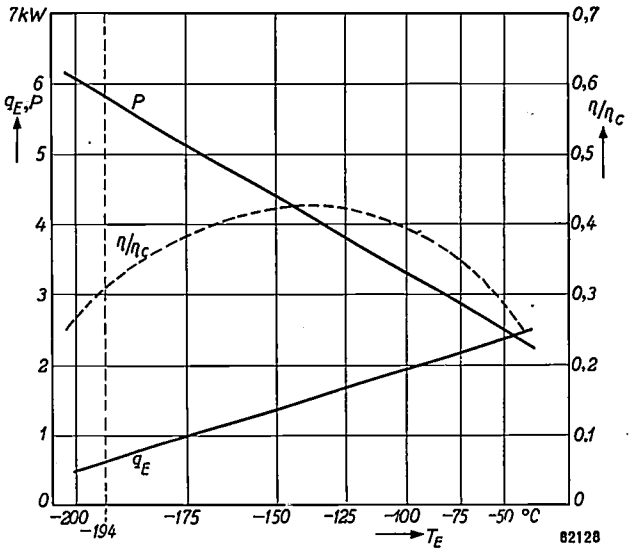


Fig. 10. Measured shaft power P and refrigerating capacity q_E of the gas refrigerating machine plotted as functions of the freezing temperature T_E. The figure of merit η/η_c (cf. fig. 2) derived from this, has also been plotted.

The value of the specific shaft power, which is a measure of the coefficient of performance ("efficiency"), requires some further explanation. The values given in the literature for the specific shaft power of various liquefiers vary within wide limits. With very large installations (producing 100-1000 kg/h), using expansion turbines, a specific shaft power of approximately 0.7 kWh/kg can be obtained; for smaller installations (producing a few kg/h) the figures lie between 1.5 and 3 kWh/kg. We see, therefore, that the efficiency of the gas refrigerating machine, in spite of its small capacity, approaches very nearly that of the large installations.

The scope of this article does not allow an analysis of the cause of this relatively high efficiency; this would require a comparison of the gas refrigeration cycle with other systems, which are quite different. It may suffice here to mention two points in this connection. First of all the pressure ratio (approx. 2.2) is small, so that the adiabatic losses (which play a part in the compressors of the other systems) are small. In the second place there is the circumstance that the power released upon expansion, is recovered by extremely simple and therefore efficient means, viz. by the gas pressure acting upon the piston.

The efficiency of the process, moreover, can be improved a great deal. No less than half the cold produced by the machine is spent in cooling the air; the remainder is used for condensation. By means

of a second refrigerator the air may be pre-cooled to an intermediate temperature, so that part of the necessary cold is obtained with a higher efficiency: consequently the overall efficiency is improved. It has been calculated that by means of an intermediate stage of this kind the specific shaft power can be reduced below 0.8 kWh/kg.

The gas refrigerating machine compares favourably with the conventional types not only in efficiency but also by its simplicity of operation and by the fact that it is not subject to contamination by dirt and dust (no expansion valves, etc.). An additional advantage is that the liquid air is completely free of oil, owing to the fact that the air need not be compressed.

Moreover, the refrigerating capacity can be varied both by changing the speed of rotation and by varying the gas pressure in the working space: hence a very simple *control* is achieved.

We shall conclude this article by giving a rough survey of the possible applications of the gas refrigerating machine. It will already be clear that the machine can be of great use in laboratories and factories for the production of liquid air. For use in laboratories particularly it will be of great value that other gases, such as nitrogen, argon, oxygen or methane can also be conveniently liquefied by this refrigerator. In this way liquid baths with a well defined boiling point can be made readily available. Preliminary tests have revealed that in certain cases the machine can be successfully used for gas separation, e.g. the fractionation of air. Finally, it can be employed as a pre-cooler for hydrogen that is to be liquefied on a small scale with the conventional type Linde-liquefier *).

The optimum working range of the machine, as previously mentioned, lies between —80 °C and —200 °C, where it can serve as a source of cold for any desired purpose. The machine has a good efficiency (cf. fig. 10) and is considerably less complicated than the cascade machine used up to now in this field. A further great advantage is that any temperature in this range is obtained with the one machine in a single stage. The simplicity and convenience of this refrigerating machine will undoubtedly act as a stimulus in low temperature research.

*) See this issue, p. 116. - Ed.

Summary. In its practical form the gas refrigerating machine, is subject to certain losses. Some of these losses increase the shaft power (mechanical loss and adiabatic loss), others cause a reduction of the refrigerating power (insulation loss and regeneration loss). The former group of losses define the upper limit and the latter group the lower limit of the useful temperature range of the gas refrigerating machine (—80 °C to —200 °C). After a brief consideration of these and other losses and their quantitative significance, this article deals with the factors that influence the design of such a refrigerator. Of considerable practical importance is the fact that the volume of a refrigerating machine of given refrigerating capacity can be drastically reduced by using a high pressure level. A description is given of a small machine in which the pressure varies between 16 and 35 kg/cm², at a crankshaft speed of 1440 r.p.m. This machine is built on the displacer principle, which was applied in some of the early hot-air engines, and is used for liquefying air. Ownig to the fact that the air does not require a preliminary compression, this liquefier is of very simple design and is easily operated. It has an output of approx. 5.5 kg liquid air per hour and the "specific shaft power" amounts to 1.0 kWh per kg of liquid air — a value which is extremely low for such a small and simple machine, and which may be even further reduced without undue difficulty. In conclusion the article deals with the properties and further potential applications of the gas refrigerating machine.

LIQUEFACTION OF HYDROGEN

Photo Walter Nürnberg

The investigation of the solid state has made remarkable advances in the last two decades. This work has been particularly fruitful in electronic and electrotechnological applications (ferromagnetic materials, semiconductors, dielectrics, luminescent substances, etc.).

In these investigations it is sometimes important to work at very low temperatures, e.g. when it is required to minimize thermal agitation of atoms and molecules, which may otherwise complicate or mask a phenomenon. In this connection, a simple hydrogen liquefier using a liquid nitrogen pre-cooler is installed in the Philips Research Laboratories. Liquid hydrogen has a boiling point of 20.4 °K at standard atmospheric pressure. The white tubes shown in the photograph contain the equipment for the cleaning and pre-cooling of the gas and also the liquefier proper. The operator is in the process of syphoning some liquid hydrogen into a Dewar flask.

AN ELECTRONIC D.C. MILLIVOLTMETER

by A. L. BIERMASZ and A. J. MICHELS.

621.317.321.027.21

Electronic meters for measuring alternating current, comprising an amplifier, a rectifier and a moving-coil instrument, are sufficiently well known. When such meters are preceded by D.C.-A.C. voltage convertors, they can also be used for the measurement of D.C. voltages. In this way a meter of very high input resistance is obtained and the use of D.C. amplifiers is avoided. A meter designed along these lines is described in the following article. The objection to many electronic meters — that they can be used only where A.C. mains are available — is eliminated by using dry batteries for the supply voltages.

This article describes an electronic D.C. voltmeter suitable for a wide variety of measuring ranges between 1 mV and 300 V. It can be used in place of ordinary moving-coil voltmeters and mirror galvanometers. Compared with the latter it has the advantages of much greater robustness, better ability to withstand overloads, and smaller inertia. An advantage of this instrument compared with conventional voltmeters is the very much higher input resistance, which greatly widens the useful scope. The input resistance is of the order of 1 MΩ in the lower measuring ranges and 100 MΩ in the higher values. For a full-scale deflection of say 1 mV, the current taken is 1.5×10^{-9} A, which corresponds to an input power of not more than 1.5×10^{12} W. In contrast with most conventional electronic voltmeters the present instrument is operated from dry batteries, so that its usefulness is not limited to places where A. C. mains are available; one application in particular where this is a great advantage is in the use of strain gauges out of doors [1]). Another useful feature of this battery-operated instrument is that measurements can be taken between two points which are both at a high potential with respect to earth (with mains-operated meters the voltage is limited by the breakdown field strength of the insulation between the windings of the power transformer); it is, of course, necessary to insulate the meter itself.

When faced with the problem of constructing an electronic D.C. meter the designer will probably first turn his thoughts to a combination of D.C. amplifier and moving-coil instrument. With an ordinary valve in the input stage a high input impedance is then obtained (10^9 ohms). In view of the required sensitivity (full deflection on 1 mV) the unavoidable drift is then fairly great, viz. about 0.3 mV per hour [2]). Moreover, D.C. amplifiers necessitate carefully stabilized supply voltages; in the laboratory this is no obstacle, but in transportable, battery-operated equipment it would certainly be a problem.

A better method consists in converting the D.C. voltage to be measured into an A.C. voltage, and to measure this with an ordinary electronic A.C. meter consisting of an A.C. voltage amplifier, a rectifier and a moving-coil instrument.

There is a choice of two methods of converting the D.C. voltage into an A.C. voltage, viz. the vibrating capacitor or the vibrating contact.

The first of these has already been described on various occasions in this Review [3])[4]), the principle being as follows. The D.C. voltage to be measured is applied to the vibrating capacitor with a high resistance is series with it. The capacitor itself consist of two plates, one fixed and one moving, with air dielectric. An moving coil loudspeaker unit fed from a valve oscillator vibrates the moving plate, thus producing a periodic variation in the capacitance. In the same way as in capacitive microphones, an alternating voltage is thus produced across the capacitor, the amplitude of which is proportional to the applied D.C. voltage.

This method, too, has the advantage of a very high input resistance (about 10^{10} ohms). It will give

[1]) A. L. Biermasz and H. Hoekstra, Philips tech. Rev. **11**, 23-31, 1949/50.

[2]) A very much higher input impedance (10^{14} Ω) can be obtained with the electrometer triode (H. van Suchtelen, Philips tech. Rev. **5**, 54-59, 1940), but the drift in this instrument is appreciably greater.

[3]) C. Dorsman, A *p*H meter with a very high input resistance, Philips tech. Rev. **7**, 24-32, 1942.

[4]) J. van Hengel and W. J. Oosterkamp, A direct-reading dynamic electrometer, Philips tech. Rev. **10**, 338-346, 1948/49.

good results provided that the voltage to be measured is not too small. If this is of the order of millivolts, however — and our meter is to be suitable for measuring fractions of a millivolt — difficulties are experienced owing to the fact that the vibrating capacitor produces an appreciable A.C. voltage even in the absence of any D.C. voltage. This effect can be attributed to the small differences in the work function of the materials of which the capacitor plates are made [5]). By making the plates of the same metal, which must be quite pure, we can reduce this effect to a few millivolts, but this is still not low enough for our purpose and, moreover, the voltage is likely to rise owing to the entry of dust and damp into the components; an instrument used in the open air would be more than usually exposed to these.

Preference has therefore been given to the vibrating-contact method. The contact periodically short-circuits the input of the A.C. amplifier and thus produces a square-wave voltage, alternating between the value to be measured and zero.

The output current from the amplifier has to be measured with a moving-coil meter and must accordingly be rectified, this latter being achieved by means of a second vibrating contact which opens and closes exactly in phase or in antiphase with the first-mentioned contact.

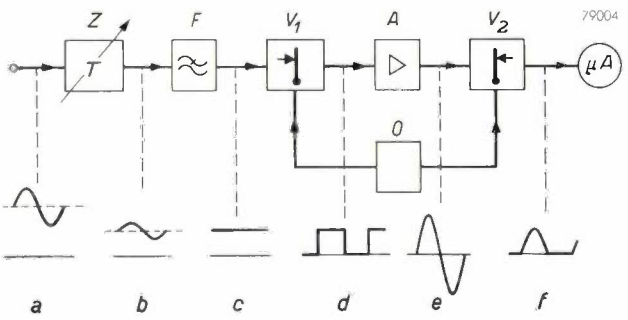


Fig. 1. Block diagram of the D.C. voltmeter GM 6010. *Z* variable attenuator (T-network). *F* low-pass filter. V_1 vibrator to convert the D.C. voltage to a square-wave voltage. *A* A.C. voltage amplifier. V_2 vibrator to rectify the output current of the amplifier. *μA* microammeter. *O* oscillator for driving the vibrator.
a represents the D.C. voltage to be measured, including ripple. *b* attenuator output voltage. *c* voltage at output of filter. *d* square-wave voltage. *e* amplifier output voltage. *f* half-wave rectified current passing through the meter.

The instrument further includes a variable attenuator for adjustment of the meter to the desired measuring range, and a filter for suppressing any A.C. voltage that may be superimposed on the D.C. voltage (see block diagram, *fig. 1*).

[5]) See p. 28 of article [3]).

The complete unit [6]), the Type No. of which is GM 6010, is depicted in *fig. 2*; *fig. 3a* shows the battery compartment and fig. 3*b* the chassis.

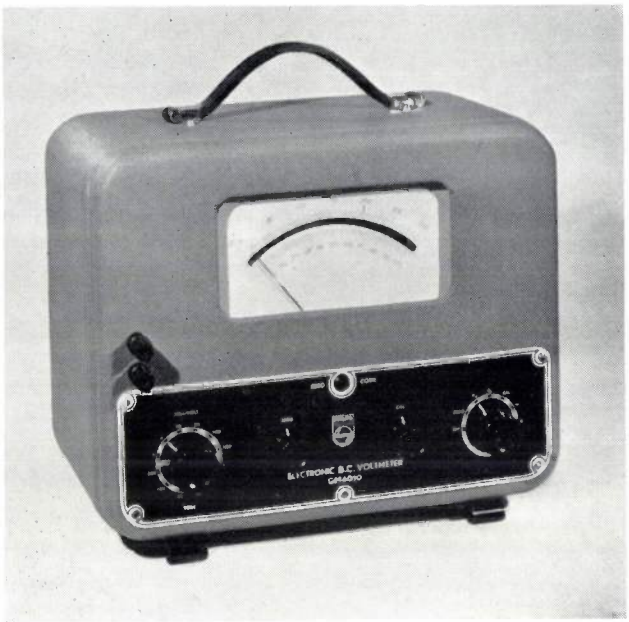


Fig. 2. The millivoltmeter GM 6010. Left to right, attenuator control, zero adjustment (compensating current), gain correction control, and multi-position switch for zero correction, measurement of the battery voltages, calibration, and reversal of polarity.

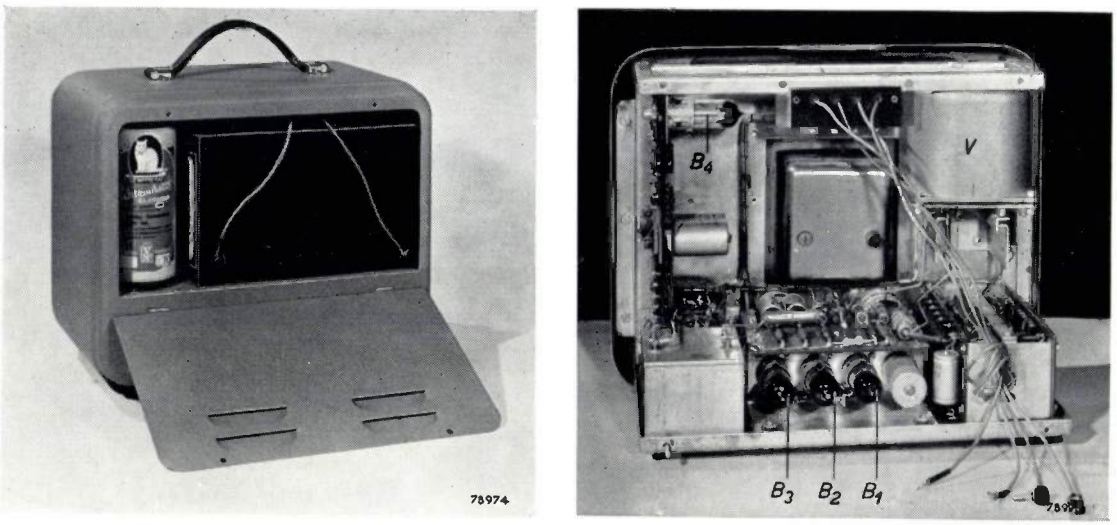
The vibrator

The contacts

The functions of the two vibrators illustrated in fig. 1 are combined in a single flat spring which alternately makes contact with contact screws mounted on each side of it (*fig. 4*).

In order to avoid errors in measurement, the making and breaking must take place without any chattering, and the time taken by the spring to travel from the one contact to the other must be short (about 2%) compared with the complete period. To ensure that no chattering will take place, the velocity at which the spring strikes the contact points must be low [7]). To ensure this, the contact screws are mounted near the clamped end of the spring where the amplitude of vibration is small and of a very definite value (about 10 μ); the amplitude of the free end of the spring is then about 0.5 mm. This arrangement also achieves a time of travel short compared to the time the contacts are closed.

[6]) The design of this meter was commenced by J. M. L. Janssen, who has since left the service of the Company.
[7]) Cf. J. A. Haringx, Vibration of contact springs, Philips tech. Rev. 7, 155-158, 1942.



a b

Fig. 3. *a*) The meter GM 6010 showing the battery compartment. *b*) the chassis; B_1, B_2 and B_3 are the amplifier valves; B_4 is the oscillator valve. *V* is the vibrator.

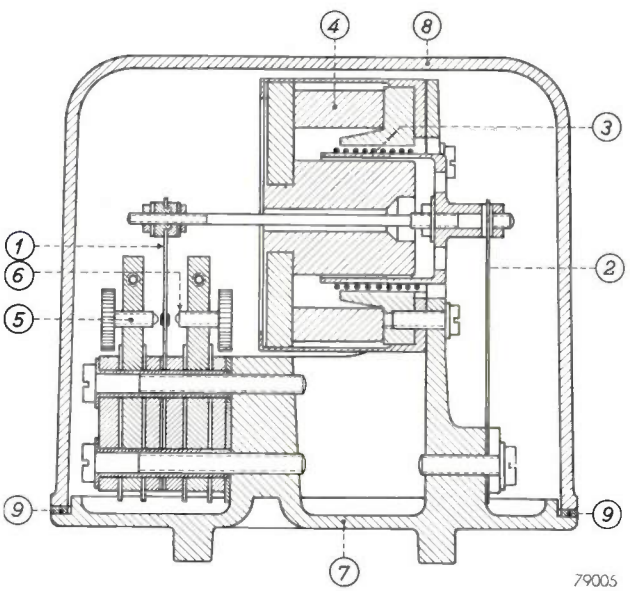


Fig. 4. Cross section of the vibrator. *1* and *2* are flat springs clamped at one end and maintained in vibration by an electrodynamic driving system comprising coil *3* and permanent magnet *4*. The spring *1* makes contact alternately with screws *5* and *6*. *7* base plate; *8* housing; *9* rubber gasket.

Whether contact is made and broken without chattering and whether the time of travel is sufficiently short, can be checked by means of the circuit shown in *fig. 5a* in conjunction with a cathode-ray oscilloscope, which should then produce a trace such as that shown in fig. 5*b*.

In the design of the vibrator the following possible sources of measuring errors were taken into account:

1) thermo-e.m.fs due to local heating at the contact points [8]).
2) contact potentials set up by differences in the chemical composition of the metals used for the vibrating contacts.
3) electrical double layers formed by the entry of dust and moisture.

[8]) The faces of the contacts in unused vibrators are relatively soft. Vibrators are "run-in" for some time in the factory: during this process the microscopically small projections from the faces are melted off, which suggests increases in temperature which are quite considerable though of short duration. This running-in makes the contact faces smoother and harder, so that no further changes in shape take place.

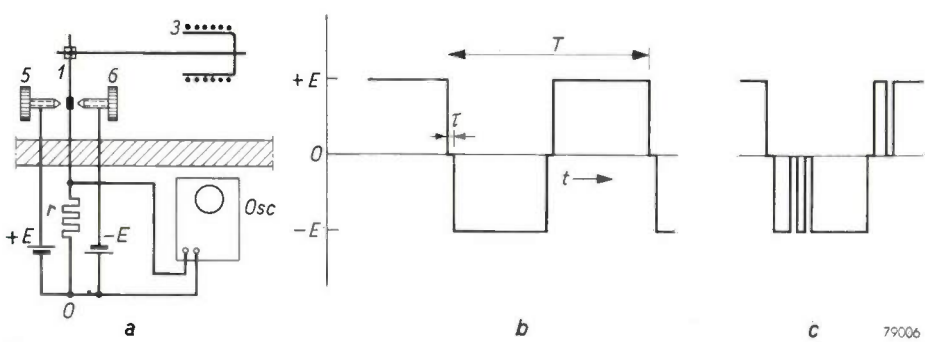


Fig. 5. *a*) Circuit for testing the vibrator; voltages $+E$ and $-E$ occur alternately across the resistor *r*. The oscilloscope *Osc* should show the image given in (*b*), without interfering pulses as in (*c*) and with a change-over time τ equal to roughly 2% of the period *T*. Other references as in fig. 4.

These sources of error are minimized by taking the following steps: the contact pressure is made small (i.e. very flexible contact spring), gold is used as contact material (this is better than platinum or rhodium, for example), and the vibrator is made as airtight as possible.

In this way it has been found possible to reduce the residual voltage arising from the causes listed above to 20 or 30 μV, i.e. to a value that is some thousandths of that occurring with a vibrating capacitor.

The meter is provided with an adjustment for eliminating the small deflection resulting from this residual voltage (2 or 3 scale divisions); this is done by passing a small variable compensating current through the meter in the opposite direction.

Maintaining the vibration

The simplest system whereby the vibrator can be kept in motion is that of the ordinary trembler bell. All that is needed is a solenoid with an additional contact on the vibrating armature, the whole being fed from the source of filament current. This system does not give satisfactory results however, as the armature spring readily assumes modes of vibration other than the fundamental and therefore cannot be made to work without chattering.

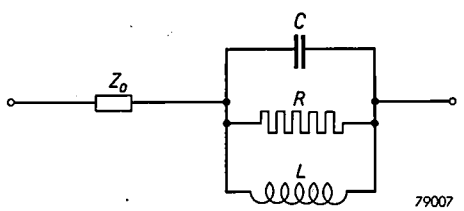


Fig. 6. Equivalent circuit of the vibrating coil. Z_0 = impedance of coil at rest.

We have accordingly adopted an electrodynamic drive of the kind shown in fig. 4. Alternating current for the driving coil is supplied by a Hartley oscillator using a DAF 41 valve. The oscillatory element is the coil itself, the equivalent circuit of which is depicted in *fig. 6*. The total impedance Z of this element should be high at the resonance point, in order to satisfy the condition of oscillation even when the mutual conductance S of the oscillator valve is low; that is: SZ = constant.

The parameters C, R and L in this circuit can be derived from the mechanical data. Let us denote the moving mass by m, the spring constant by c, the damping factor by k, the velocity by v and the angular frequency by ω. The force acting on the moving system in the stationary condition is then:

$$F = (j\omega m + k + \frac{1}{j\omega c})\, v. \quad \ldots\ldots \quad (1)$$

If we further denote the terminal voltage of the coil by e, the induced electromotive force by e_i, the impedance of the coil at rest by Z_0 and the current by i, then:

$$e = iZ_0 - e_i. \quad \ldots\ldots\ldots \quad (2)$$

Also:

$$F = ai \quad \ldots\ldots\ldots\ldots \quad (3)$$

and:

$$e_i = -av, \quad \ldots\ldots\ldots \quad (4)$$

where $a = Bl$, B being the magnetic induction in the air-gap, and l the length of wire in the coil.

Elimination of F, v and e_i from (1), (2), (3) and (4) gives the impedance Z of the coil:

$$Z = \frac{e}{i} = Z_0 + \frac{1}{j\omega \frac{m}{a^2} + \frac{k}{a^2} + \frac{1}{j\omega c a^2}}.$$

From this and from fig. 6 it follows that $C = m/a^2$, $R = a^2/k$ and $L = ca^2$. At resonance, $\omega m/a^2 = 1/\omega\, ca^2$, so that the resonance frequency is determined by mc, and Z becomes $Z_0 + a^2/k$, where Z_0 is approximately the D.C. resistance of the coil. For easy oscillation a^2/k should be high, i.e. a high (strong magnetic field, many turns), and k small (weak damping).

For our purpose the self-inductance L in the equivalent circuit is 0.25 H and the capacitance C is 20 μF, which gives a resonance frequency of about 70 c/s. At this low frequency the impedance of the coil at rest, Z_0, is almost that of a resistance of 800 ohms. The total impedance of the coil vibrating at its own natural frequency is roughly 7000 ohms, so that oscillation will occur even when the mutual conductance of the valve is low. This high impedance was attained by adopting various measures to reduce the damping; the air damping has been kept low by using a streamlined coil former with perforations. Owing to the flexible suspension, moreover, there is very little loss through a transfer of energy to the framework. The input power of the coil is only a few milliwatts.

The amplifier

A three-stage amplifier is used, with DAF 41 valves in the first and second stages and, in view of the desired linearity, a DL 41 [9]) as output valve (*fig. 7*).

The anode circuit of the output valve includes a transformer whose core is provided with an air-gap to avoid D.C. saturation, which would affect the linearity of the scale. By means of a parallel capacitance, the primary side of the transformer is tuned roughly to the frequency of the vibrator, so that, although the input voltage is of square wave-form, the output voltage is practically sinusoidal.

This parallel capacitance affects the phase of the output voltage. To ensure effective rectification, it

[9]) The DL 41 valve is discussed in Philips tech. Rev. **10**, 346-351, 1948/49.

is of such a value that the current is zero during the time that the vibrator passes from one contact to the other.

on the D.C. voltage and which would otherwise introduce errors in measurement. The filter consists of resistors and capacitors and is designed to give

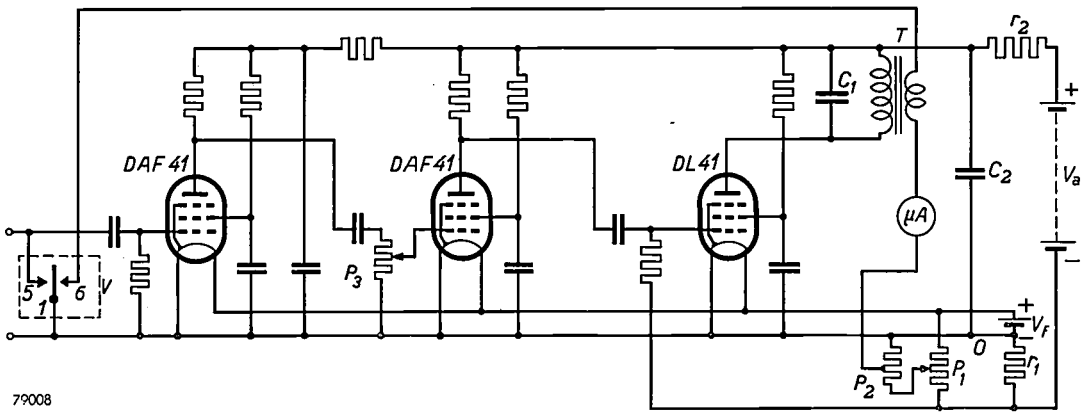


Fig. 7. Simplified schematic diagram of the amplifier and vibrator. *V* vibrator, with spring *1*, shorting contact *5* and rectifying contact *6*. By means of the capacitor C_1, the output transformer *T* is tuned roughly to the frequency of the vibrator (approx. 70 c/s). V_a H.T. voltage (90 V nominal). V_f L.T. voltage (1.4 V nom.) The resistor r_1, through which the anode and screen currents pass, provides bias for the output valve DL 41 and also makes available at the slider of the potentiometer P_1 a positive or negative voltage with respect to the point *0*. By means of P_1 (through the fixed potential divider P_2) a current is passed through the μammeter which is just sufficient to compensate the deflection due to residual voltage. The gain is corrected by means of potentiometer P_3. A 3900 ohm resistor r_2 is connected in series with the H.T. battery to reduce the effect of variations in the internal resistance of the battery. C_2 is a decoupling capacitor.

A moving-coil meter is connected to the secondary side of the transformer, in series with the vibrator, and the current passing through it is therefore half-wave rectified. The advantage of this is that the instrument thus also indicates the polarity of the voltage to be measured; when the polarity is reversed, the other half-cycle of the output current is passed and the meter needle deflects in the other direction. A switch is provided to permit this reversal of polarity.

Other components

The attenuator

This comprises three groups of resistors R_1, R_2 and R_3, arranged as a T-network (*fig. 8*), and a switch controlling them gives a choice of 12 ranges, with sensitivities of 300 V, 100 V, 30 V, ..., 3 mV, 1 mV, full scale deflection. The input resistance is highest (100 MΩ) in those positions in which 1 V or more is required for maximum deflection, and lowest (0.6 MΩ) in the most sensitive position of the switch. The maximum error in the attenuator is 2%.

The filter

The filter follows immediately after the attenuator (fig. 8). The object of the filter is to suppress any A.C. voltage (ripple) which might be superimposed on the D.C. voltage and which would otherwise introduce errors in measurement. The filter consists of resistors and capacitors and is designed to give an attenuation of 500 × for 50 c/s voltages. This means that even a ripple voltage equal to 10 times the D.C. voltage to be measured has little or no effect on the result (provided that the ripple is not greater than 5 V; if this happens, the A.C. voltage arriving at the input of the amplifier via stray capacitances would overload the amplifier).

Early on during the design, the capacitors in the filter presented difficulties. These will now be mentioned.

Dielectric after-effects and frictional electricity

In the original design, standard capacitors with paper dielectric were used in the filter. Now, when the attenuator is set in advance to the required measuring range, a potential of not more than 1 mV occurs across the capacitors. With incorrect handling,

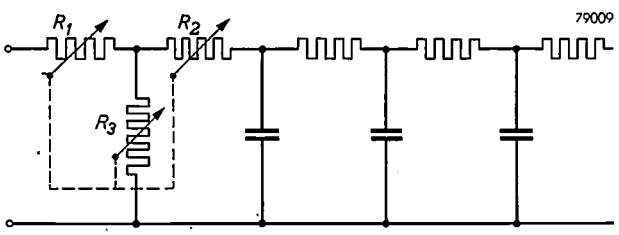


Fig. 8. Attenuator and filter. R_1, R_2 and R_3 are three groups of resistors in a T-network, operated by a single 12-position switch.

however, for example when a 100 V potential is to be measured and the instrument is set for maximum 1 mV, the voltage on the capacitors can be very much higher. The instrument is able to withstand this (because the gain drops considerably on overloads), but the real difficulty was found to be that the discharge of the capacitors takes place very slowly, necessitating an hour's wait before the instrument is again ready for use.

This can be demonstrated by the following experiment. A paper capacitor, of capacitance $C = 0.22$ μF, is charged up to 100 V and is then discharged through a resistor $R = 1.68$ MΩ. At first the voltage will drop in accordance with the anticipated exponential curve with time constant $RC = 0.37$ sec, until a value of about 1 mV is reached. The further voltage drop, which is shown plotted in *fig. 9*, takes place much more slowly: as an approximation, exponentially with a time constant of 165 sec, i.e. roughly 500 RC. It takes about an hour for the voltage to drop to 10 μV (= one division of the scale of the GM 6010). Short-circuiting of the capacitor does not help matters, because as soon as the short is removed, roughly the same voltage occurs again (see A and B, fig. 9).

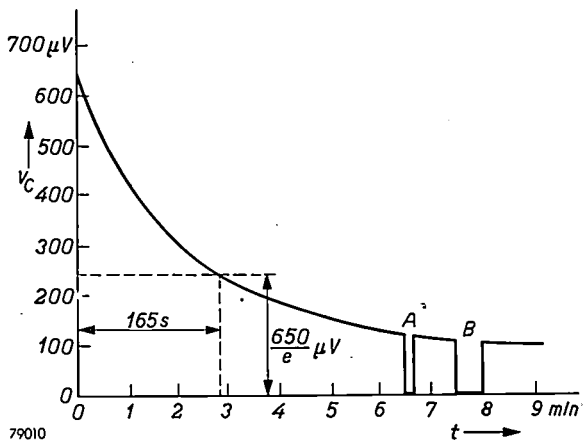


Fig. 9. Dielectric after-effect in a paper capacitor charged to 100 V and discharged through a resistor (time constant $RC = 0.37$ sec). The figure shows the drop in the capacitor voltage from the moment it reaches 650 μV. v_C then diminishes much more slowly than before, more or less exponentially, with a time constant of about 165 sec. Even when the capacitor is short-circuited for different lengths of time (at A and B), the voltage v_C returns to roughly the same value after the short is removed.

This is due to a kind of after-effect; ions are produced in the dielectric which can disappear only slowly. Some materials exhibit this property to a much smaller extent than others; in polystyrene, for example, the effect is very slight indeed and persists for only about 1/50 the time as in paper. For this reason filter capacitors with polystyrene dielectric are used in the GM 6010.

As the insulating material used for the terminal blocks would also be likely to exhibit this effect, polystyrene is used for this purpose as well.

Another source of interference is to be found in frictional electricity generated on the insulating material on the connecting leads, particularly if this is of plastic. Static charges may be set up by friction, flexion or vibration, and these will produce a deflection. Because of this, rubber-covered flex, or, better still, bare leads, if necessary supported by ceramic insulators, are used.

Calibration

For preference, electronic meters should be equipped with means for self-calibration. This applies all the more when the instrument is operated from dry batteries, whose voltage may vary considerably, and when the amplifier, as in the GM 6010, does not use negative feed-back.

In the first place the supply voltages of the instrument are checked to see that they are within the appropriate limits. This is done using the moving coil meter which, with the calibration control set to certain positions, functions as an ordinary voltmeter. The H.T. battery should give a reading between 95 and 75 V, and the filament cell between 1.55 and 1.05 V.

Calibration is effected as follows. A potential divider of fixed resistors delivers a certain fraction $(1/A)$ of the L.T. voltage V_f. With the calibration control rotated one stage, this voltage V_f/A is applied to the input of the vibrator, the gain being then so adjusted by means of the gain control (potentiometer P_3, fig. 7) that the meter indicates V_f, i.e. the same deflection as with direct measurement of the L.T. voltage. It is then known that the gain is equal to the fixed value A on which the scale calibration is based.

Summary Description of an electronic D.C. millivoltmeter (type GM 6010), in which the D.C. voltage to be measured is converted by a vibrator into a square-wave voltage which is subsequently amplified by an A.C. amplifier. A second contact on the vibrator rectifies the output current from the amplifier, and this rectified current is passed through a moving-coil meter. The vibrator is driven by an electrodynamic system in which the coil constitutes part of a valve oscillator. The amplifier and oscillator are operated from dry batteries, and the use of the instrument is therefore not limited to places where A.C. mains are available. As the meter is independent of the mains, voltages can be measured between points which are both at a high potential with respect to earth. The vibrator circuit is preceded by a variable attenuator and a filter; the former has 12 positions corresponding to ranges of 300 V, 100 V, 30 V, ..., 3 mV, 1 mV. The input resistance lies between 100 MΩ for ranges of 1 V and upwards and 0.6 MΩ for 1 mV; the filter suppresses A.C. voltages that may be superimposed on the D.C. voltage to be measured. Polystyrene is used for the dielectric of the filter capacitors as well as for the terminal blocks, as this material shows very little electrical after-effect. The instrument is provided with means for self-calibration.

THE "NORELCO" X-RAY DIFFRACTOMETER

by W. PARRISH *), E. A. HAMACHER †*) and K. LOWITZSCH *).

539.262:548.733:
621.387.424

The X-ray spectrometer, described in this Review a number of years ago, has been completely re-designed. The new instrument, which has now been commercially available for some years, will be the subject of this and a following article. The present description is not intended for the instruction of potential users (these are served more fully by special publications of the Company): its main purpose is to reveal the technical basis and implications of the new design.

Introduction

The application of X-ray diffraction analysis as a tool for technical and scientific investigations has gained a firm footing in an ever-increasing number of industrial and university laboratories. The last decade has witnessed the successful introduction into this field of a new method, the direct measurement and recording of line intensities in X-ray diffraction patterns by means of a Geiger counter tube. An instrument based on this principle and known as the "Norelco" Geiger counter X-ray spectrometer was designed and marketed by the North American Philips Company in 1945, and described in this Review about six years ago [1]). *Fig. 1* offers a schematic picture of this instrument; the caption recapitulates some details concerning the forming of the X-ray diffraction pattern and the focusing method.

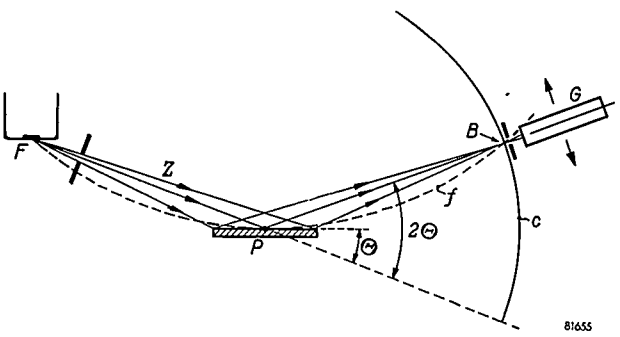


Fig. 1. Focusing arrangement of the Geiger counter X-ray diffraction instrument. A divergent X-ray beam Z coming from the focus F irradiates the surface of a flat specimen positioned at P. All the rays diffracted at Bragg angle Θ by suitably oriented crystallites on the specimen surface converge approximately to a single line, at B. Here the receiving slit of a Geiger counter tube G is placed. The counter tube is moved along the goniometer circle c about the axis P, in order to scan angles 2Θ; the flat specimen is rotated about the same axis at half the angular speed, thus always remaining tangential to the "focusing circle" f through F, P, B (Bragg focusing). (The "focusing" effected is, of course, not true focusing in the optical sense.)

This article describes a completely re-designed version of the spectrometer, whose performance and facilities represent a great advance on those of the first instrument. The new instrument, which from now on will be termed an X-ray *diffractometer* [2]), is pictured in *fig. 2*. It comprises three parts, which may be purchased separately and each of which can be used in conjunction with other equipment if desired: a basic diffraction unit (X-ray tube with high voltage generator and controls), a Geiger counter goniometer, and an electronic circuit panel with automatic recorder. The Geiger counter goniometer is shown separately in *fig. 3*.

Although this and a subsequent article will give a self-contained description of the diffractometer, it will be based on a comparison with the former instrument, as this should enable the reader more readily to grasp the significance of a number of details of the design.

Features of the diffractometer

The old and the new instrument alike offer the features characteristic of the method, viz., the instantaneous indication and recording of line intensities and a very considerable saving of time in cases where only part of the diffraction pattern need

*) Philips Laboratories, Irvington-on-Hudson, N.Y., U.S.A. We regret to record the death of Mr. Hamacher on March 25, 1954.

1) J. Bleeksma, G. Kloos and H. J. di Giovanni, X-ray spectrometer with Geiger counter for measuring powder diffraction patterns, Philips tech. Rev. **10**, 1-12, 1948.

2) The instrument has been manufactured and marketed for some time by North American Philips Company, Inc., New York, U.S.A., again under the name of X-ray spectrometer. It has now been agreed to reserve the name X-ray spectrograph or spectrometer for the proper classical use in the measurement of X-ray spectra (and for a modified form of the X-ray diffractometer applied to X-ray fluorescence analysis).

Features of the instrument have been previously described by W. Parrish and E. A. Hamacher at A.S.X.R.E.D. meetings 1947-1949, in Science **110**, 368-371, 1949 and in Trans. Instr. Meas. Conf. Stockholm 1952, p. 95.

A similar X-ray diffractometer is now in production in the Philips Works at Eindhoven; the electronic circuit of the latter instrument, although built on the same principles, is somewhat different in detail.

Fig. 2. The "Norelco" Geiger counter diffractometer consists of three independent parts: a basic diffraction unit (X-ray tube with high voltage generator and controls, at left), the high angle precision Geiger counter goniometer placed on top of it, and an electronic circuit rack with automatic 10″ strip chart recorder (at right).

be analysed. In the new design, however, a number of substantial improvements were obtained, which may be enumerated as follows:

1) Higher resolution in the diffraction pattern.
2) Better accuracy in the measurement of diffraction angles and line intensities.
3) Higher diffraction angle range including the "back reflection" region up to angles of $2\Theta = 165°$.
4) More universal and more flexible use of the instrument. Two goniometers and (for example) two normal photographic powder diffraction cameras can be used simultaneously in conjunction with one basic diffraction unit.

A more precise and quantitative indication of these improvements will be given below, but this short list may serve us as a guide in the description that now follows.

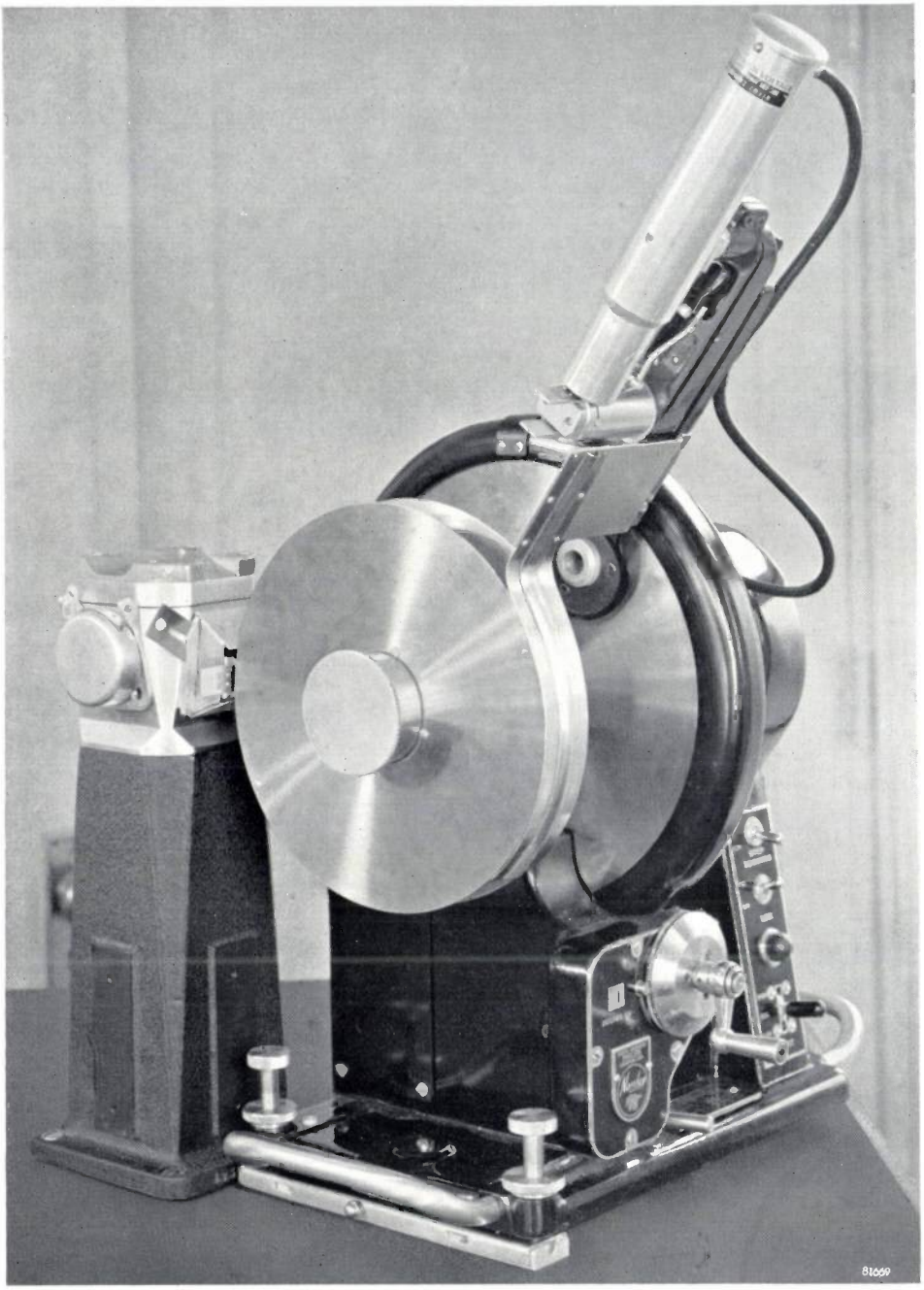


Fig. 3. The "Norelco" Geiger counter goniometer. The Geiger counter tube rigidly mounted on the scanning arm of the goniometer is seen at the top of the photograph. A handle for manual scanning and dials for reading degrees and fractions of degrees diffraction angle 2Θ are seen near the base of the instrument.

Arrangement of X-ray tube and goniometer

For the sake of clarity we start with the last point, though this improvement was obtained by rather straightforward measures, and is not necessarily the most important for all users of the instrument.

In normal photographic diffraction techniques it has been common practice for a long time to use an X-ray tube with four windows so that four diffraction cameras can be operated simultaneously. The running time of the tube and the time of the operator are thus more usefully employed. *Fig. 4* shows the arrangement. It should be noted that the windows of the X-ray tube may be divided into two pairs which transmit beams of different cross-section. The line-shaped focus produced on the target of the X-ray tube (10 mm $\times$ 1.6 mm in the "Norelco" tube) is viewed at a small glancing angle to the target surface, since at a smaller apparent size it will exhibit a higher X-ray brilliance. Through the two opposite windows looking in the *long* direction of the focus at a glancing angle of say 6°, a small, approximately square focus of 1.6 $\times$ 1.6 mm will be

seen, whereas the two remaining windows looking at the same glancing angle in the direction *perpendicular* to the line focus will offer a long and very thin apparent focus, of e.g. 10 mm × 0.16 mm.

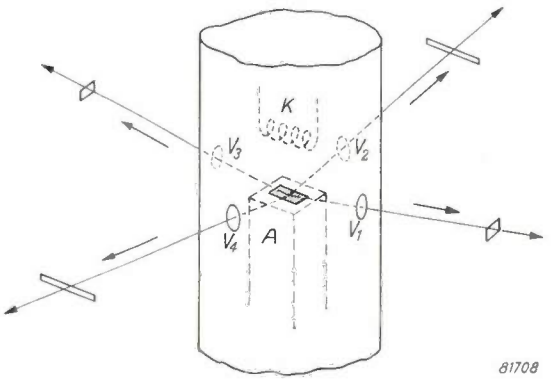


Fig. 4. Arrangement for using four diffraction devices simultaneously with a four-window X-ray tube. Through one pair of windows (spot focus windows V_1, V_2) the source is seen as a small square surface, through the other pair as a very narrow line (line-focus windows V_2, V_4). (In reality the anode A is positioned above the cathode K.)

Using the "line-focus" windows a large part of the focus area will remain unused in filling the narrow collimator sytem of a usual Debye-Scherrer camera with radiation, whereas when using the first mentioned "spot-focus" windows, radiation of the full focus area may be employed. The spot-focus windows are therefore better suited for photographic cameras such as the Debye-Scherrer and single crystal goniometers, and the line-focus windows for focussing cameras such as the symmetrical back-reflection and crystal monochromators.

In the old X-ray spectrometer both the X-ray tube and the goniometer circle, along which the Geiger counter tube travels, were mounted in a horizontal position and the X-ray optics of the arrangement was based on the employment of a "spot-focus" window (cf. the article quoted in [1]).

In the new instrument, on the other hand, the X-ray tube is mounted with its axis vertical and the Geiger counter tube scans in the vertical plane, thus offering geometrical conditions favorable for using four windows, according to fig. 4. Moreover, the X-ray optics of the present design is based on the use of a line-focus window, thus leaving the two spot-focus windows available for photographic devices. Finally, the former spectrometer for the sake of simplicity and economy contained an air-cooled X-ray tube rated for only 35 kV, 125 W; the basic diffraction unit employed with the new instrument, is equipped with a normal water-cooled X-ray diffraction tube rated for 50 kV, 800 W (copper target), which produces a high X-ray intensity, sufficient for the photographic recording of weak diffraction lines.

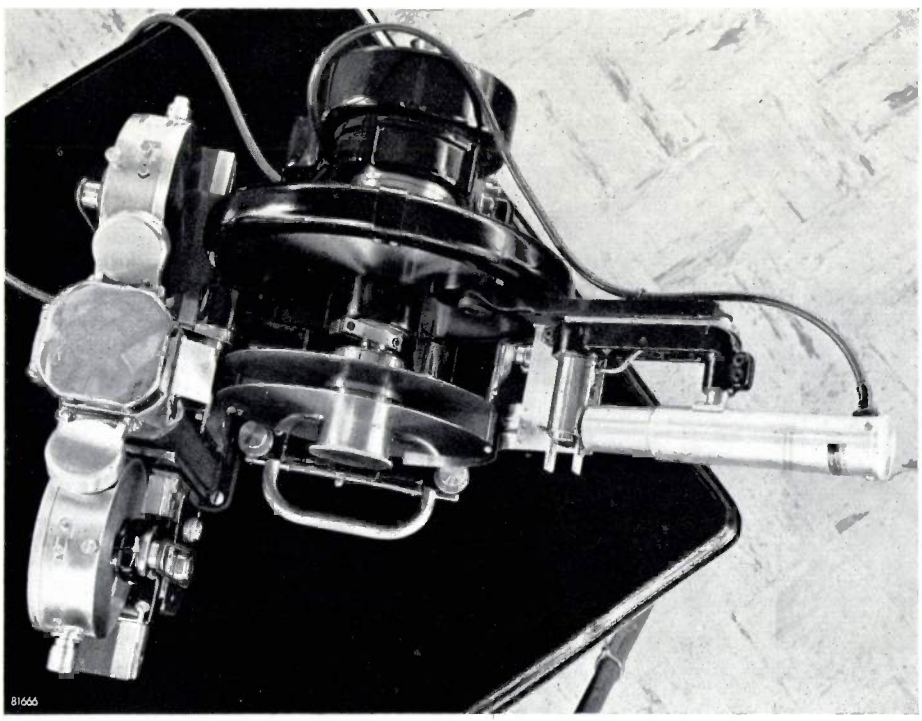


Fig. 5. Top view of basic diffraction unit, with three diffraction devices in simultaneous operation. On the left the anode end of the X-ray tube may be seen; on either side of it are two "Norelco" powder diffraction cameras; to the right is one of the new Geiger counter goniometers. Opposite the fourth window of the X-ray tube (i.e. to the far left in the photograph) another diffraction instrument can be placed.

The photograph *fig. 5* demonstrates the practicability of a simultaneous use of all four windows. In this instance one Geiger counter goniometer is used on one of the line-focus windows and two Debye-Scherrer cameras on the spot-focus windows; the fourth place is vacant.

The new X-ray optics, which is an essential innovation of the design, will presently be discussed at some length, but first it should be briefly explained how the high angular range of the diffractometer is achieved.

The high angular range

In the old instrument the angular range was limited mechanically since the movement of the Geiger counter tube beyond angles of $2\Theta \approx 90°$ was impeded by the anode end of the X-ray tube cover. With a water-cooled X-ray tube, as employed in the present diffractometer, the total anode surface required for cooling is rather small, so that the tube cover can be made to extend only slightly beyond the focus. This permits the Geiger counter tube to continue its travel to $165° 2\Theta$.

The increased X-ray tube rating again is essential for providing this facility, as the diffraction lines at these high angles (back reflection lines) are usually very weak.

It should be mentioned incidentally, that the useful X-ray intensity is raised not only by increasing the rating of the X-ray tube, but also by providing the tube with mica + beryllium windows instead of the formerly used Lindemann glass windows [3]). The transmission of different types of windows is indicated in *Table I*. It is seen that the transmission of mica 0.012 mm thick is about 86% for the CuKα-line (1.54 Å) which is used in most cases. The transmission of a complete window consisting of such a sheet of mica and a thin beryllium plate is about 83%. A Lindemann window 0.25 mm thick transmits only 61% of CuKα. For a softer radiation which must be used in some diffraction investigations, e.g. CrKα radiation, the advantage of the mica + beryllium window becomes even more important.

The vacuum tight mica + beryllium windows are relatively cheap and easy to fabricate, making four-window X-ray diffraction tubes an economic proposition.

Table I. Calculated transmission (%) of different types of window for X-rays of different wavelengths.

Window \ X-ray spectral line		CuKα	CrKα
Lindemann glass	0.5 mm	38	5
Lindemann glass	0.25 mm	61	22
Mica	0.012 mm	86	66
Beryllium	0.12 mm	96	90
Mica 0.012 mm + Be	0.12 mm	83	60

[3]) The use of a mica entrance window in the Geiger counter tube to give a very high sensitivity has been described in the article quoted in [1]), where the application of mica exit windows for the X-ray tube was anticipated. A description of one type of X-ray tube (contact therapy tube) equipped with a mica + beryllium window was given in Philips tech. Rev. **13**, 75-77, 1951/52.

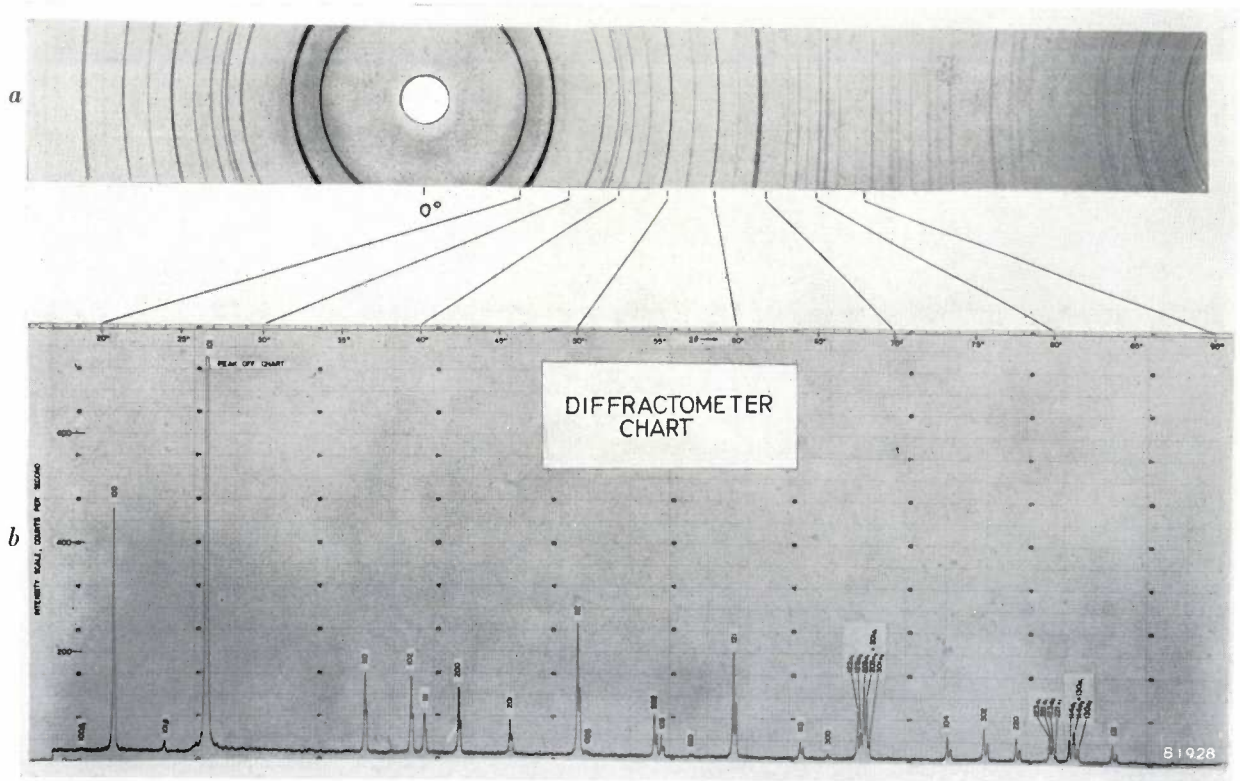


Fig. 6. *a*) X-ray diffraction pattern of quartz powder photographed with a 114.6 mm diameter Debye-Scherrer camera. Filtered CuKα radiation, 40 kV_p, 20 mA, exposure time 2 hours. *b*) Automatically recorded chart obtained with the new Norelco X-ray diffractometer using the same X-ray tube conditions. Recording time 4.8 hours.

X-ray optics of the instrument

Perhaps the most notable improvement obtained in the diffractometer is its very high resolution, i.e. the extreme sharpness of the diffraction lines recorded. When using CuKα radiation and recording the diffraction pattern of a well crystallized powder specimen (whose diffraction lines have a very small natural width) with a receiving slit of about the same width as the source, the width at one half peak height of the recorded Kα_1 lines is about 0.1° 2Θ in the front reflection region. The separation of the two lines produced by the CuKα_1 and Kα_2 radiations, which in normal photographic techniques can be seen only at rather large diffraction angles ($2\Theta > 110°$) is visible at about $2\Theta > 30°$ in the diffractometer recordings, cf. *fig. 6*.

This high resolution is primarily obtained by reducing the geometrical width of the X-ray source: the focus of the X-ray tube measuring 10 mm × 1.6 mm is viewed at an angle of about 3° to the target surface perpendicularly to its length (i.e. a "line" X-ray tube window is being used as stated above). The effective source width with this arrangement is 0.08 mm as against 0.2 mm in the old spectrometer. *Fig. 7* serves to illustrate the difference between the former and the present method.

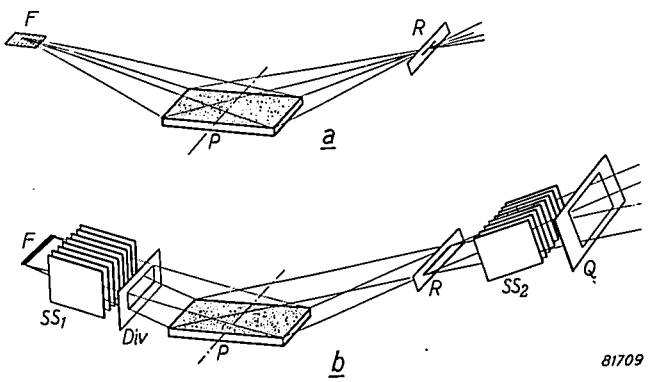


Fig. 7. *a*) X-ray optics of the former instrument (the "spectrometer" described in [1])). A "spot-focus" source F is used. The receiving slit R must be rather short in order to get a narrow recorded line profile when scanning. P = flat specimen. *b*) X-ray optics of the new instrument. A "line-focus" source F is used. The receiving slit R has the same length as the source. This is made possible by the use of the parallel slit systems SS_1 and SS_2, in conjunction with a chlorine quenched Geiger counter tube, as explained in the text. *Div* = aperture limiting slit, Q = scatter eliminating slit.

The substitution of the effective line-source in fig. 7*b* for the spot-source in fig. 7*a* was made possible by two new elements in the design, viz. the introduction of a Geiger counter tube containing chlorine as a quenching agent and the insertion of two "parallel slit systems" in the X-ray beam as shown in fig. 7*b* and *fig. 8*. Let us first consider what would happen if the parallel slit systems were omitted. The long effective X-ray source may be regarded as a large number of point-sources aligned in a horizontal direction parallel to the axis of the specimen and of rotation of the counter tube.

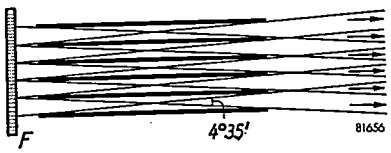


Fig. 8. Parallel slit system consisting of a number of molybdenum foils 0.025 mm thick, transmitting X-rays directed parallel to the foils and suppressing rays entering obliquely. F = focus. (The effect may be compared to that of the well-known Potter-Bucky grid for suppressing scattered radiation in X-ray diagnostics.)

Considering a number of imaginary vertical planes perpendicular to the axis of the goniometer, each point-source in its own vertical plane will give rise to a pattern of very narrow diffraction spots. In fact, however, owing to the horizontal divergence of the X-rays emitted from the focus each point-source will produce horizontal diffraction *lines* and thus contribute to the patterns in the planes of all the remaining point-sources. As these lines are shaped as ring sections (with a curvature largest for angles 2Θ approaching 0° and 180°, as is well-known from Debye-Scherrer photographs, cf. fig. 6*a*), the superposition of the contributions of the aligned spot-sources would result in considerable asymmetric line broadening as illustrated in *fig. 9*. This is avoided by the insertion of the two parallel slit systems, each of which consists of a number of thin (0.025 mm) molybdenum foils placed in the X-ray beam in such a way that the narrow spaces between adjacent foils may be regarded as the above-mentioned imaginary planes. Each individual spot of the line

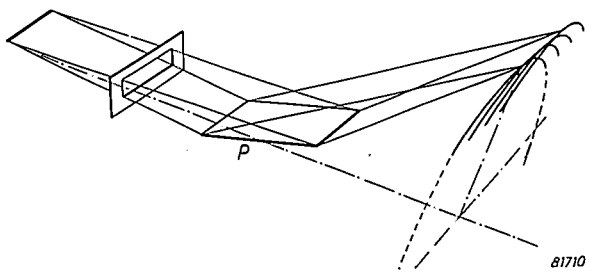


Fig. 9. If the parallel slit systems in fig. 7*b* are omitted, each point of the line source in a given specimen position produces a diffraction line of its own, covering the whole length of the counter window and being curved as a ring with its centre of curvature in the 0° 2Θ direction. The curved lines of all source points are integrated to form a broadened diffraction line of asymmetric profile.

source is now substantially prevented from contributing radiation outside its "own" plane, as these oblique radiations are strongly absorbed on their long path through the foils; the horizontal divergence of rays in each system is restricted to the very small value of 4° 35′, resulting in the very sharp lines demonstrated by fig. 6*b*. Even at small diffraction angles ($2\Theta = 5$ to 10°), where the curvature of lines is rather pronounced, the measured line profiles still have a high symmetry; see *fig. 10*.

The Geiger counter tube positioned with its window opposite a diffraction line will indicate the line intensity by integrating the X-ray energy received along the length of the line (parallel to the axis of rotation of the specimen). This is where the

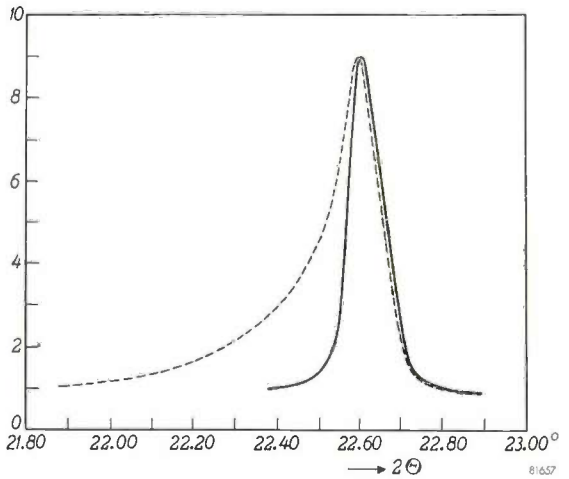


Fig. 10. Profile of a diffraction line measured with the Geiger counter diffractometer, without the parallel slit systems (broken line) and with the parallel slit systems (full line).

second factor mentioned above, i.e. the addition of chlorine to the gas filling of the counter tube, enters the picture. The tubes formerly used contained argon for absorbing the X-ray quanta (and giving rise to photo-electrons) and methylene bromide as a quenching agent. These tubes had a rather limited "sensitive volume": only those quanta that passed within about 1.5 mm from the axial anode-wire had a high probability of being counted. Hence the integration was effectively accomplished over only a 3 mm length of the line. Reflections from the whole specimen height (10 mm) were nevertheless included in this integration owing to the rather large horizontal divergence permitted. With the X-ray optics of the goniometer illustrated in fig. 7*b*, however, the limitation of the line length to 3 mm at the entrance of the Geiger counter tube would mean that reflected rays emanating from the ends of the specimen surface are not detected and that the ends of the X-ray source do not contribute to the integrated line intensity. It is an essential feature of the revised X-ray optics that the chlorine counter tubes possess a much larger sensitive volume, covering the complete 10 mm width of the beam in the goniometer (*fig. 11*): this ensures that full use is made of the

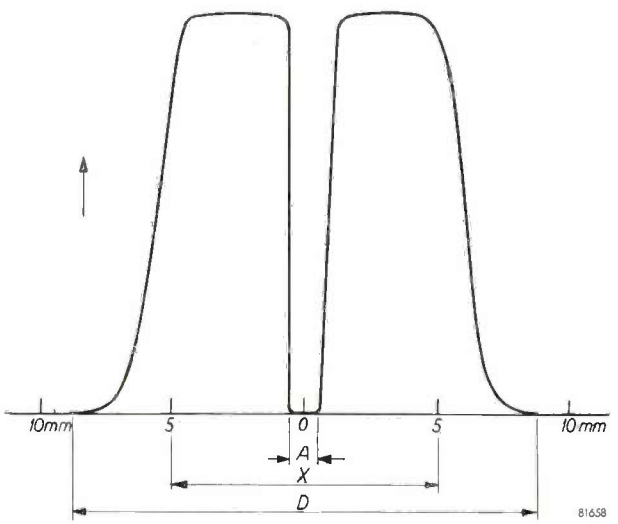


Fig. 11. Radial distribution of sensitivity (in arbitrary units) of chlorine-quenched Geiger counter tube (type 62019). A = width of the axial anode wire and bead. X = width of X-ray beam when using the new X-ray optical system. D = inside diameter of the Geiger tube.

focus as well as of the specimen area (the latter is important for good averaging of the crystallite reflections).

The chlorine counter tubes also offer other advantages: for example, a very long life, a rather low working voltage and a long "plateau", but we shall not dwell on these points in this article [4]). A photograph of the counter tube in its present form is reproduced in *fig. 12*.

Although the Geiger counter tube in its present form has excellent characteristics as a radiation detector, other types of detectors have been developed for use in the diffractometer, viz. special forms of the proportional counter and the scintilla-

Fig. 12. The Geiger counter tube, type 62019, used in the "Norelco" diffractometer. The end window where the X-rays enter the tube (in a direction parallel to the tube and its axial anode wire) is made of mica 0.012 mm thick. The gas filling consists of argon to a pressure of approximately 55 cm Hg with a few mm Hg chlorine as a quenching agent. The operating voltage is 1400 to 1500 volts, the plateau is at least 300 volts long.

[4]) See N. Warmoltz, Geiger Müller counters, Philips tech. Rev. **13**, 282-292, 1951/52.

tion counter. Both of them have the advantage of a very short "dead time", allowing counting at rates many times higher than with the Geiger counter. Moreover, they offer the possibility of pulse height discrimination, which in some cases may be useful. An example is the X-ray diffractometry of radioactive samples, where the background caused in the diffraction pattern by the radioactivity can be reduced very effectively when using a scintillation counter [5]). The latter has a higher quantum efficiency than all other radiation detectors (nearly 100%) for the wavelengths normally used in X-ray analysis.

It is interesting to compare the old and the new instrument from the viewpoint of the number of counts per second obtained for a given diffraction line. It should be pointed out that with a given real focus area a crystallite of the specimen "seeing" the whole focus will receive the same radiated energy per second regardless whether the focus is viewed from its broad or from its narrow side and whatever viewing angle is chosen [6]) (within certain limits). This well known basic fact would mean in our case that the increased resolution would not entail a sacrifice of line intensity except in so far as the horizontal divergence of the rays has been somewhat decreased, permitting a crystallite to see only part of the focus. This loss, however, is amply made good by the much higher radiation output per cm^2 of the focus (higher specific loading of the focus made possible by the water-cooling).

The goniometer; alignment procedure

In order to obtain full profit from the high resolution achieved it was necessary to ensure that diffraction angles could be measured with adequate accuracy. A large goniometer radius was therefore chosen (radius 17 cm instead of 13 cm with the former instrument [7])), resulting in a large dispersion of the diffraction pattern. At the same time a completely new mechanical design was adopted for the goniometer, enabling the setting and direct reading of the Geiger counter tube position to an accuracy of 0.01° 2Θ or better.

A cross-section drawing of the goniometer is shown in *fig. 13*. It consists essentially of a precision 10″ worm wheel and worm drive (W_1 and W_2). The wheel carries the Geiger counter tube rigidly attached to it in a radial position. The main dial permitting direct reading of 2Θ to 0.01° is fixed on the worm, complete dial revolutions being registered by a subsidiary gear and a mechanical counter. This accuracy is achieved by "cold working" the wheel and permanently loading the worm against it [8]). The specimen under investigation, which is made in the shape of a flat plate 20 × 10 mm (P in fig. 13), is placed in a holder carried by a hollow shaft positioned within the hollow bearing of the worm wheel and coupled to the latter by a set of accurate herringbone gears. These rotate the specimen at exactly one-half the angular speed of the Geiger tube arm.

To ensure the above-mentioned accuracy of the diffraction angle measurements, the zero angle position of the goniometer must be set to a precision of better than 0.01° 2Θ. This is achieved as follows. A narrow slit or pin-hole is placed in the specimen holder (*fig. 14a*). The reference or banking surface of the holder contains the specimen rotation axis, which is made in the factory to coincide with the goniometer axis to better than 0.01 mm. The goniometer arm is turned slowly in 0.01° steps across the X-ray beam transmitted by the slit or pin-hole. The position of maximum intensity thus found is the zero angle position, provided the slit was placed

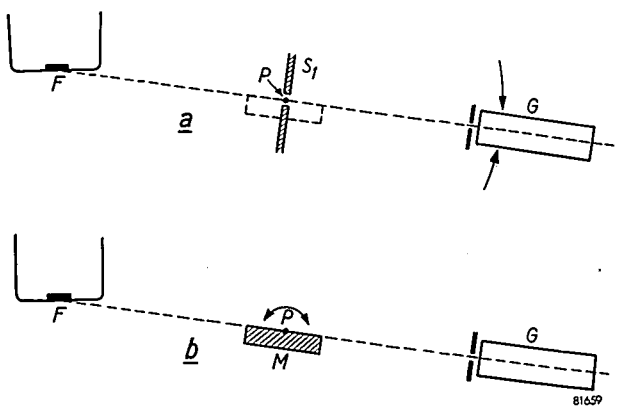


Fig. 14. Alignment and calibration procedure for the goniometer. *a*) A narrow slit S_1 is placed at the specimen rotation axis P. The Geiger counter tube G is turned to the position in which the largest intensity of the direct X-ray beam transmitted by S_1 is recorded (F = focus of X-ray tube). This is the zero angle position of the goniometer.
b) A flat machined piece of metal M is fastened on the specimen holder and rotated so that maximum intensity of the direct X-ray beam is received by the Geiger counter tube in its zero angle position. The specimen reference plane is then exactly parallel to the zero angle direction, and the exact ratio 1:2 of specimen angle and counter tube angle is obtained when scanning.

[5]) T. R. Kohler and W. Parrish, X-ray diffractometry of radioactive samples, to be published in Rev. sci. Instr. Cf. also: J. Taylor and W. Parrish, Absorption and counting efficiency data for X-ray detectors, also to be published in Rev. sci. Instr.

[6]) This is due to the fact that the impinging electrons penetrate only very slightly into the target whereas the X-rays produced at a depth where the electrons are stopped emergc from the target practically unimpeded even in directions nearly parallel to the target surface. Thus from all directions (in front of the target) the same *volume* of target material is seen to contribute to the radiation. See for example Philips tech. Rev. 3, 261, 1938.

[7]) It should be remembered that most diffraction cameras have *diameters* of 5.7 or 11.4 cm. Increasing the radius beyond 17 cm did not seem desirable because of increasing air absorption for long wavelengths and air scatter; moreover, at 17 cm, the size of the instrument is still reasonable.

[8]) This process is conducted in such a way that the resultant surface hardening almost eliminates further wear.

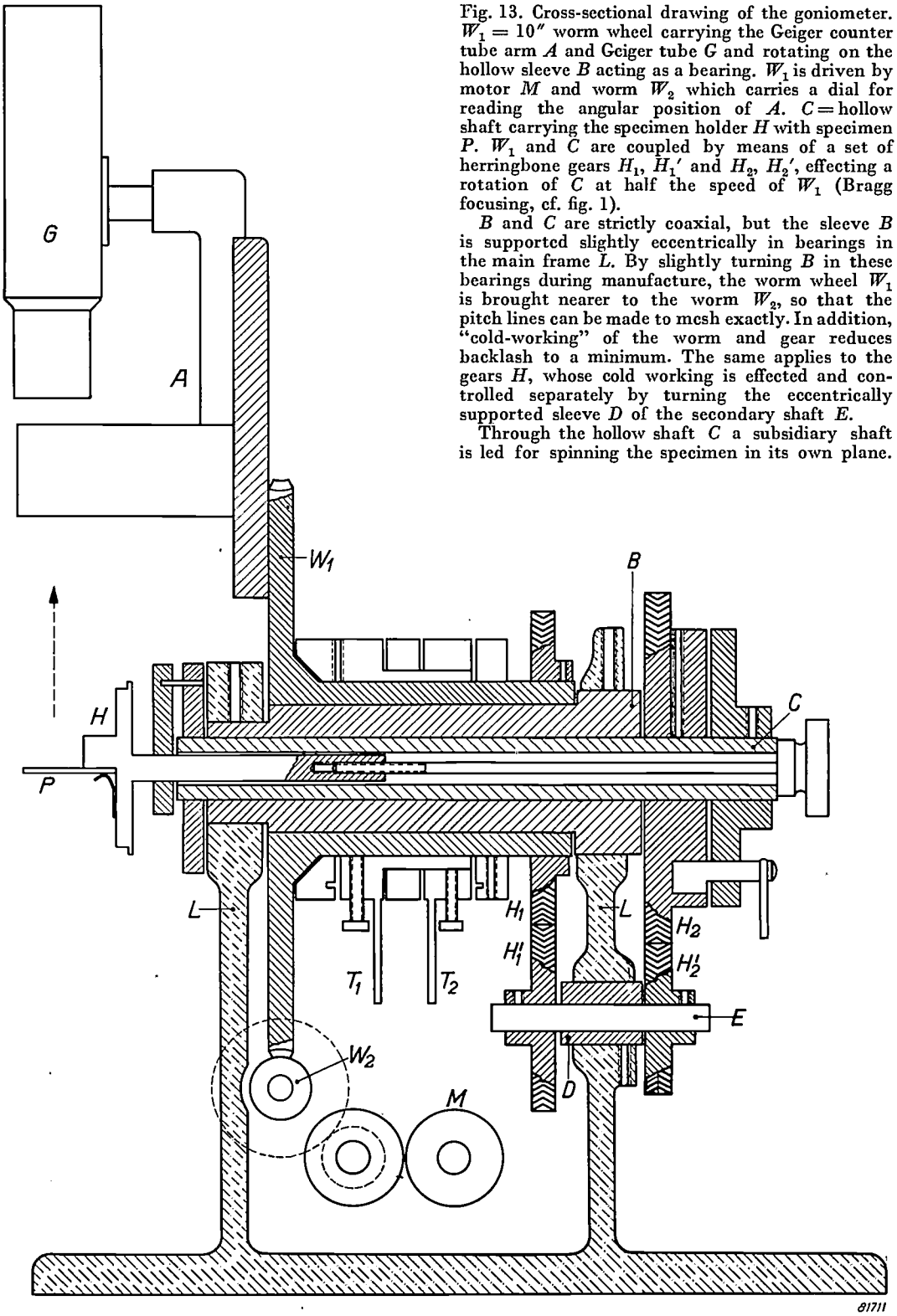


Fig. 13. Cross-sectional drawing of the goniometer. $W_1 = 10''$ worm wheel carrying the Geiger counter tube arm A and Geiger tube G and rotating on the hollow sleeve B acting as a bearing. W_1 is driven by motor M and worm W_2 which carries a dial for reading the angular position of A. C = hollow shaft carrying the specimen holder H with specimen P. W_1 and C are coupled by means of a set of herringbone gears H_1, H_1' and H_2, H_2', effecting a rotation of C at half the speed of W_1 (Bragg focusing, cf. fig. 1).

B and C are strictly coaxial, but the sleeve B is supported slightly eccentrically in bearings in the main frame L. By slightly turning B in these bearings during manufacture, the worm wheel W_1 is brought nearer to the worm W_2, so that the pitch lines can be made to mesh exactly. In addition, "cold-working" of the worm and gear reduces backlash to a minimum. The same applies to the gears H, whose cold working is effected and controlled separately by turning the eccentrically supported sleeve D of the secondary shaft E.

Through the hollow shaft C a subsidiary shaft is led for spinning the specimen in its own plane.

exactly on the goniometer axis. The unavoidable error in the placing of the slit is compensated by repeating the procedure with the specimen holder revolved through 180°. The mean value of the two maximum intensity positions read is the true zero angle position. With the arm in this position, the dial is unscrewed and reset to 0.00°.

When this has been done, the specimen holder in this goniometer position must be adjusted so that the surface of the flat specimen will be exactly parallel to the zero angle direction, in order that this surface on scanning will always be oriented at *exactly* one-half the angle of the Geiger counter tube arm (cf. fig. 1). This so-called 2:1 setting is

performed (with the goniometer arm set at 0.00°) by placing a flat machined piece of metal in the specimen holder (fig. 14*b*), and rotating the holder by means of a micro-adjustment until maximum intensity of the direct X-ray beam is obtained.

As was mentioned above, the reference surface of the specimen holder is manufactured so that it coincides with the goniometer axis to better than 0.01 mm. If diffraction occurs only at the specimen surface, which is approximately the case for a strongly absorbing material, and if the specimen were removed only 0.075 mm from the axis of rotation, the peak positions of the diffraction lines would be shifted 0.045° 2Θ at $2\Theta = 45°$.

Relative line intensity measurements

Since the Geiger counter method of recording X-ray diffraction patterns (unlike the film method) requires point-by-point measurements, the intensity of the primary X-ray beam must be highly stabilized to make relative measurements of the diffracted intensity reliable. The X-ray generator of the basic diffraction unit employed now operates on full wave rectified voltage enabling more reliable regulation than the previous instrument with a self-rectifying X-ray tube. Long term stabilization to better than 0.2% for both voltage and current of the X-ray tube is obtained by use of an electronic voltage regulator and a feedback type current regulator [9]).

Intensity measurements for weak lines are facilitated by the wavelength response of the argon-filled Geiger counter tube, as was discussed in the article quoted above [1]); the background intensity of the diffraction pattern due to the short wavelength continuous radiation is detected with a relatively low efficiency, resulting in a low and even background, comparing very favorably with that of photographically recorded patterns, for equal line peak height. The recording and measuring of weak lines is further improved by the higher intensity of the water-cooled X-ray tube. It is important to note that better accuracy of the intensity measurement is achieved also for *strong* diffraction lines, owing to the full wave rectification. It is well known that for the comparison of high and low line intensities, the non-linear response of the Geiger counter tube, which is caused by the "dead-time" after every recorded count, is an fundamental limitation. By using full wave instead of half wave rectified voltage across the X-ray tube, with the same total number of quanta arriving in a random fashion at the Geiger counter tube the average time-separation of two quanta will be doubled, thus diminishing the influence of quanta arriving in "dead time" periods and hence shifting the non-linearity effect to higher intensities (twice the former limit).

If line intensities in different 2Θ regions have to be measured, the differences of illumination of the 20 mm width of the specimen rotating in the fixed primary beam must be taken into account. With a beam of angular aperture 1° in the scanning plane, the 20 mm specimen width is just covered when the specimen is positioned at 17° 2Θ. As the angle increases, only part of the specimen is illuminated. This does not affect the relative line intensities, provided the sample is of sufficiently uniform reflecting power over its whole surface. For accurate measurements we have found it useful to rotate the specimen in its own plane (at moderate speed, e.g. 77 rev/min) in order to smooth out the statistical fluctuations obtained even with crystallites smaller than 20 μ in the specimen. Such a rotation is performed by means of a subsidiary shaft led through the hollow specimen holder shaft (*C* in fig. 13) of the goniometer and driven by a motor placed at the rear of the instrument. In this way congestion around the specimen is avoided and enough clearance is provided for mounting high or low temperature specimen chambers or other accessories.

In order to make use of the whole specimen surface in the back reflection region, where line intensities are very low, the angular aperture of the primary beam is increased to 4° by inserting a larger aperture-limiting slit. On the other hand, for small Bragg angles, $2\Theta < 17°$, smaller apertures must be used, so that the beam does not exceed the specimen width. (This is desirable even when line intensities are not to be measured, in order to avoid excessive scattering. With very small apertures, down to 5′, it is thus possible to measure diffraction angles corresponding to lattice spacings up to about 90 Å, using CrKα radiation.) The divergence slits, as well as the other slits shown in fig. 7*b*, are designed for maximum reproducibility and convenience by having *fixed* apertures. All slits are constructed of molybdenum, for high X-ray absorption and good mechanical strength.

Scanning of the pattern; counting methods

It will be remembered that due to the random arrival of quanta the accuracy of the line intensity measurement depends on the total number of counts recorded by the Geiger counter tube when traversing

[9]) The Eindhoven version of the diffractometer employs a specially designed basic diffraction unit (type PW 1010) containing a more elaborate stabilization device.

the angular region of the diffraction line [1]). In order to conform to specific accuracy requirements, different counting methods can be adopted. With the simplest method, which is used for most routine analyses, and for recording charts as in fig. 6*b*, the pattern is continuously scanned and the average current intensity produced by the current pulses of the Geiger tube is measured. Automatic scanning is accomplished by driving the goniometer by a fractional horsepower electric motor (*M* in fig. 13). A contact on the main worm shaft of the goniometer with a cam for 0.5° intervals in 2Θ actuates a degree marking pen on the strip chart recorder. The accuracy of the recorded line intensities depends on the scanning speed. Through the use of changeable spur gears, scanning speeds (in either direction) of $^1/_8$, $^1/_4$, $^1/_2$, 1 or 2° per min can be selected. A wider receiving slit is used for scanning at higher speed in order not to lose too much in intensity. Owing to the high X-ray intensity of the tube the accuracy is fairly good even at the highest scanning speed. The chart reproduced in fig. 6*b* was recorded at medium speed ($^1/_4$°/min); the recording took about 5 hours.

The same method can be employed for investigating only small parts of a diffraction pattern. Use is then made of adjustable upper and lower limit stops provided on the goniometer (T_1 and T_2 in fig. 13) for stopping or reversing the scanning motion.

For higher accuracy, manual point-by-point plotting of a chart, using the so-called fixed count method, is sometimes desired. In this case a discontinuous stepwise movement of the goniometer is substituted for the continuous scanning movement. The mechanism causing this stepwise movement is shown in *fig. 15*. It can be adjusted to advance the Geiger counter tube in steps from 0.01° to 0.05°2Θ. This type of scanning motion is required also for the recording *counting rate computer*, a device developed to combine the advantages of fully automatic operation and high accuracy in a wide range of intensities.

For the sake of completeness a fourth method of obtaining data from the goniometer should be mentioned, viz. the integrated area measurement. This method and the other methods mentioned above will be described in detail in a subsequent article. All the circuits necessary for measurements according to these methods are contained in the cabinet shown on the right of fig. 2. The conversion from one method to another can be done simply and quickly by means of tap switch selectors.

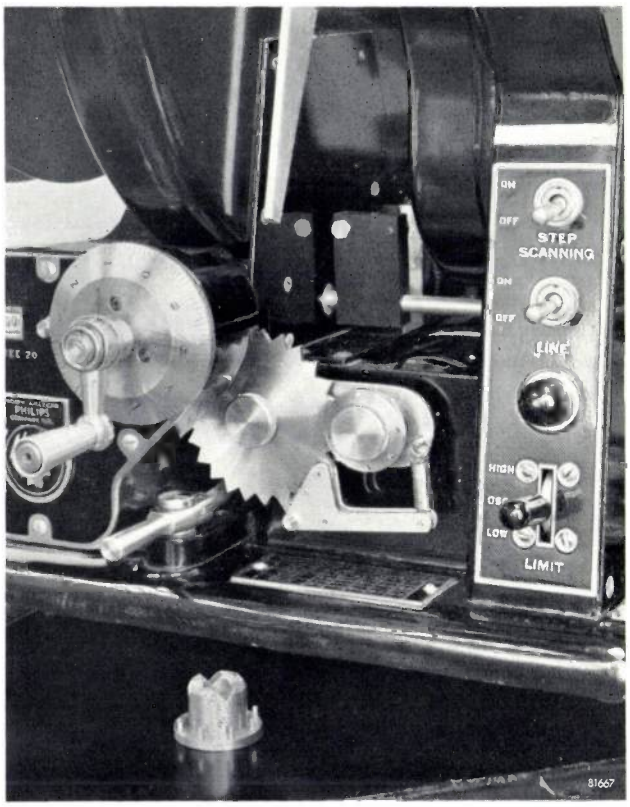


Fig. 15. Mechanism for stepwise movement of the goniometer arm. The goniometer arm rotates through 0.25° for each revolution of the shaft geared directly to the motor. For the stepwise movement this shaft is fitted with a wheel with 25 triangular-shaped teeth which is advanced one tooth at a time by a pin on the driving member on the motor shaft. With one pin on the driving member one revolution of the motor shaft advances the goniometer by 0.01° 2Θ, with two pins 0.02°, etc. The driving member can accomodate up to 5 pins (see driving member in the foreground). Indexing is achieved by a spring loaded roller which falls between the teeth of the wheel.

Summary. The "Norelco" X-ray diffractometer, which has now been commercially available for some years, is a completely re-designed version of the Geiger counter X-ray spectrometer formerly described in this Review. The new version incorporates an improved resolution of diffraction lines (CuKα_1-α_2 separation visible at $2\Theta > 30°$), better accuracy in the measurement of diffraction angles (readings to 0.01° in 2Θ) and of relative line intensities, ansd a higher diffraction angle range (up to $2\Theta = 165°$). These improvements are due to the use of a new X-ray optical system together with a normal basic diffraction unit having a high-powered water-cooled X-ray tube with mica-beryllium windows, and to the design of a high precision mechanical Geiger counter goniometer. The new X-ray optics is based on the use of a line-focus source. Owing to this feature and to a vertical arrangement of the X-ray tube and goniometer circle, it is possible to use two goniometers and two normal photographic powder diffraction cameras simultaneously with the four window X-ray tube of one basic diffraction unit. Scanning speeds varying from $^1/_8$° to 2° per minute can be selected; alternatively an automatic stepwise movement of the goniometer is possible. Apart from automatic recording, either continuously or point-by-point, the electronic circuits developed for the instrument permit the application of special counting methods. The various counting and recording methods will be described in a following article.

A POCKET DOSEMETER, WITH BUILT-IN CHARGER, FOR X-RADIATION AND GAMMA RADIATION

by N. WARMOLTZ and P. P. M. SCHAMPERS. 621.386.82 : 539.16.08

The accumulated dose of γ-radiation and X-radiation to which a person has been exposed is commonly determined by means of a miniature ionization chamber combined with an electroscope charged to some hundreds of volts, which can conveniently be carried in the pocket.

In this type of instrument the radiation causes discharge of the system according to the dosage; this can be read from the electroscope by means of a microscope provided for the purpose; Thanks to new developments in glass technology, however, it is now possible to make these dosemeters completely vacuum-tight and able to withstand higher voltages, which increaese the deflection and so permit the microscope to be dispensed with. Moreover, it is found that this construction is well suited to the incorporation of an electrostatic charger in the same instrument; this does away with the need of batteries and enables the meter to be used under any conditions, even under water.

Ionizing radiations

Effect upon living tissue

Owing to the present large-scale production of artificial radioactive substances, ionizing radiations other than the familiar X-rays are being produced. For example, γ-rays are now being generated on a wide scale and at intensities which at one time would have been considered incredibly high: hence the problems relating to the effect of these radiations upon living organisms are demanding more and more attention.

In view of these developments, the measurement of the dosages to which people are exposed is a subject of considerable importance. Before entering into the details of this subject, however, we will briefly recall the manner in which these radiations affect biological matter and the methods adopted to establish a measure of the irradiation dose based on these biological effects.

Charged particles, e.g. α-rays, β-rays, protons and fission products, cause direct ionization and also excitation and dissociation of molecules. In matter, they proceed along a certain path depending, amongst other things, upon the energy, mass and charge of the incident particle and the nature of the particular material through which it passes. The depths of penetration of these particles with the energies nowadays prevalent are relatively small, ranging from some fractions of a millimetre to a few centimetres in the case of very hard β-radiation.

On the other hand, X-ray and γ-ray quanta (photons) cause very little direct ionization, but the fast electrons released by these quanta have, like β-radiation, a very high ionizing power, and so cause secondary ionization, excitation and dissociation.

For these electromagnetic quanta the probability of absorption is considerably less than for the first-mentioned particles, so that, in general, the rays pass right through the object, though attenuated by scatter and absorption. To illustrate this point, 1 MeV γ-radiation in water has a linear coefficient of absorption of 0.07 cm^{-1}: in other words, this radiation is reduced to about $^1/_3$ of its original intensity in penetrating a barrier of water about 14 cm thick.

Fast neutrons, being uncharged, ionize indirectly by transferring a considerable proportion of their energy to any light-weight nuclei with which they happen to collide, so enabling these nuclei to produce ions in the tissue. Slow neutrons cannot do this, but in certain substances readily produce nuclear reactions whereby ionizing particles are released. The penetrating power of fast neutrons is very high in heavy substances, but decreases with the weight of the atoms constituting the material.

Dosimetry

To study the effect of radiation upon biological matter we require in the first place a measure of the radiation. Let us now consider the manner in which such a measure is established.

The processes initiated by radiation in biological tissue are very complex as may be deduced from the fact they involve such subsidiary effects as the

above-mentioned ionization, excitation and dissociation; in fact, there can be no doubt that our concept of these processes as a whole is still incomplete. However, it has been found in the course of many investigations that the quantity of energy absorbed may be employed as a measure of the effects produced by radiation, for example, the destructive effect upon tissue. Since direct measurement of the energy absorbed by a particular tissue is usually impracticable, it is necessary to find some other means of obtaining this information. An associated quantity suitable for direct measurement is the ionization caused by the radiation in a given volume of air (or other gas).

Since the mean energy to form one ion pair in air is known (32.5 eV), the energy absorbed in the air can then be computed. To ascertain from this the amount of energy absorbed in a particular tissue, it is necessary to take into account the nature of the tissue and that of the radiation, and the geometry of the object and of the beam.

The energy absorbed by a particular tissue can be computed for a wide range of wavelengths of X-rays and γ-rays by multiplying the corresponding value in air by the ratio of the respective X-ray absorption factors of the two absorbers [1]). The unit of dosage employed for X-radiation and γ-radiation is the röntgen: one röntgen of X-radiation or γ-radiation is such that it produces in 0.001293 gram of air (1 cc at 1 atm. and 0 °C) ions of either sign carrying 1 e.s.u. of charge. The amount of energy absorbed by air from one röntgen (r) is 84 ergs per gram.

It is particularly necessary to bear in mind that, in general, the value in röntgens is merely a measure of the X-ray irradiation at a particular point, as evaluated in terms of the ionization caused by such irradiation assuming that there is air at that point. This applies irrespective of the real nature of the matter constituting the point concerned. However, throughout a wide range of wavelengths the röntgen-evaluation is likewise a measure of the effect produced by the radiation. It is found that the amount of energy absorbed per röntgen varies very little as between different biological tissues; for the most important muscle, and other tissues, the energy absorption at nearly all wavelengths is 80 to 100 ergs per gram per röntgen.

However, for relatively soft radiation some tissues absorb considerably less energy (e.g. up to 50 ergs per gram per röntgen in the case of fat) and others considerably more (roughly, up to 500 ergs per gram per röntgen).

Basically, the equipment employed to measure radiation in röntgens comprises an air-filled chamber containing two electrodes between which a voltage is applied, and a sensitive instrument for the measurement, either direct or indirect, of the ionization current [2]).

For β-particles and other radiations which cannot conveniently be measured in röntgens and to which the formal definition of this unit is not strictly applicable, another quantity has been introduced, i.e. the "absorbed dosage", the unit of which is the "rad". The "absorbed dosage" may be defined as the amount of energy transferred from an ionizing radiation to a particular point in a given substance per unit mass by ionizing particles. One rad is equivalent to an absorbed dosage of 100 ergs per gram; X-radiation can also be evaluated in terms of the rad.

Although in general the effect of a particular radiation upon organic tissue depends upon the dosage as measured in röntgens or rads, equal rad-dosages of different radiations do not always produce exactly the same biological effect. Owing to the complexity of the process it is impossible to assign this difference direct to any particular cause, but there is little doubt that it arises from spatial variations in ion density. For example, where an α-particle and a β-particle cause the same overall ionization, the one will produce along its relatively shorter path a much greater concentration of ions than the other. Again, fission products and recoil nuclei of neutrons likewise produce very heavy ion-concentrations as compared with β-particles, or with X-rays and γ-rays. This relative concentration of the ionization sometimes enhances, and sometimes diminishes the biological effect.

Accordingly, another unit has been introduced, viz. the "rem". A dosage in rems equals the same dosage in rads multiplied by a factor representing the relative biological effect appropriate to the particular case. As a rule, tolerance dosages are expressed in terms of the number of rems per week. According to definition, the relative biological effect of X-radiation and γ-radiation is unity; hence the tolerance evaluated in rads equals the tolerance in rems. We have already seen that in certain (usually unimportant) circumstances, the energy absorbed per röntgen by particular tissues may vary quite appreciably above or below 1 rad; however, for purposes of protection against X-radiation and γ-radiation, where the precise maximum tolerance dose is not always known, this possible variation is ignored and 1 röntgen is considered equivalent to one rad.

The measurement of X-radiation and γ-radiation can also be accomplished by means of a suitable Geiger-Müller counter, crystal counter or scintillation counter, and the result expressed in terms of the number of particles so recorded per unit area and per unit time. Provided that the sensitivity of the particular counter and the energy distribution of the radiation are known, the dosage in röntgens can then be computed.

[1]) For a more detailed account of these effects and of the exact method of measurement, see the article by W. J. Oosterkamp in Appl. sci. Res. **B3**, 100, 1953 and **B3**, 477, 1954.

[2]) See the article by J. van Hengel and W. J. Oosterkamp in Philips tech. Rev. **10**, 338 - 346, 1948/49.

Another widely-used method of dosimetry is that based upon the blackening of a photographic plate. Again, the discoloration or fluorescence of certain crystals and glasses has lately been adopted as a measure of large dosages.

Tolerance dose

It has long been evident that exposure to X-radiation or γ-radiation causes varying degrees of damage to living organisms. The effects increase in severity with the dose, ranging from small changes inaccessible to direct observation, and slight and temporary variations in the blood count, to grave, possibly fatal, injuries to essential organs. Similar effects are produced by corpuscular radiations, the least penetrating of which, although affecting only the skin and the subcutaneous tissue, may nevertheless give rise to serious consequences.

The increasing use of radioactive substances and ionizing radiations in science and technology calls for the provision of a survey of tolerance dosages. Accordingly, the International Commission on Radiological Protection has recommended certain maximum limits for the radiation doses which may be considered tolerable in the event of life-long irradiation of the entire body (*Table I*).

Table I. Tolerance dose for irradiation of the entire body.

	X-radiation or γ-radiation	Other ionizing radiation
In the blood-forming organs, sex organs and the eyes . . .	0.3 r or rad per week	0.3 rem per week
In the base layer of the epidermis . . .	0.6 r or rad per week	0.6 rem per week

For irradiation confined to the hands and forearms, the feet and ankles, or the head and neck 1.5 r per week is permissible, provided that in the last case the eyes are protected so that the dose received by them does not exceed 0.3 r per week.

Since α- or β-radiations and protons can usually be prevented from reaching the body by quite simple methods (e.g. gloves), the above tolerances are important first and foremost in work involving X-radiation, γ-radiation and neutrons; on the other hand, substances emitting α-radiation are particularly dangerous when assimilated through the nose or mouth.

All the tolerances are specified in terms of the dosage per week. It is held to be immaterial whether the dose absorbed during this period results from a brief irradiation at a high dosage rate, or from a continuous irradiation at a low one. The effect of the radiation upon living tissue is cumulative, but partial recovery takes place in the course of time.

To complete this survey, *table II* shows the approximate effects of a large short-period dose of γ-radiation under conditions of total body irradiation.

Table II. Effect of a heavy short-period dose of γ-radiation, body totally irradiated.

0- 25 r	: no perceptible injury.
25- 50 r	: changes in the blood count, no serious injury.
50-100 r	: changes in the blood cells, some injury.
100-200 r	: serious injury.
200-400 r	: very grave injury, possibly fatal.

Pocket dosemeters for personal protection

In order to avoid the danger of exceeding the tolerance doses defined above, a person working in the presence of ionizing radiations must acquire, with the aid of one of the usual instruments for measuring dosage rate, an appreciation of the situation in which he works and hence an assurance that in normal circumstances he will not absorb more than the permissible weekly dose. Since he cannot continuously measure the dosage rate prevailing wherever he happens to be throughout the day, and since this rate usually varies at every moment, it is desirable that the worker be provided with a pocket dosemeter from which he can read, at the end of a day or week, the total dose absorbed during that time. This also ensures that any incidental and transient increase in intensity will be measured as well.

Where the particular radiation employed is confined to a narrow beam, special precautions are necessary to ensure that this beam cannot strike any part of the body without also striking the dosemeter.

Dosemeters of the type usually employed include a very thoroughly insulated electrometer system, which is charged from a battery or some other source of about 200 V; the deflection of this meter is read with the aid of a small microscope. The electrometer system is encased in a conductive housing, which also constitutes the ionization chamber containing air at a pressure of 1 atm. The number of ions formed in this chamber per second by the incident radiation corresponds to the dosage rate, and the total number of ions so formed during the particular period of observation governs the discharge of the system. As a rule, the displacement of the electrometer leaves is proportional to this total, and the dosage can be read from the scale of a microscope calibrated in röntgens.

One of the existing types of dosemeter has a sensitivity of 0.5 r full scale deflection; the variation in the reading of this meter in the absence of radiation is only 1 to 2% per 24 hours. Heavier doses are measured by means of dosemeters incorporating well-insulated capacitors, which increase the full scale deflection reading to 100 r. A diagram of such a meter is shown in *fig. 1*.

A simple charging device incorporated in the instrument would enable the latter to be used as a dose-rate meter, from which the time required for a discharge covering a certain number of scale divisions can be determined by counting. Although by no means comparable in quality with the instruments usually employed for this purpose, the above combination would constitute a very simple and inexpensive dose-rate meter. A meter of this type giving a full scale reading of 0.2 r would permit a dosage rate of 72 röntgens per hour to be measured quite easily in a matter of 10 seconds, again, a rate as low as 2 r per hour could be measured with the same instrument in 1 minute by utilising only about 20% of the scale. Higher dosage rates, up to the order of several hundred röntgens per hour, could also be measured with the same instrument. However, to facilitate the task of reading the dosage rate at frequent intervals and under all manner of conditions, it is desirable to find some means of dispensing with the microscope.

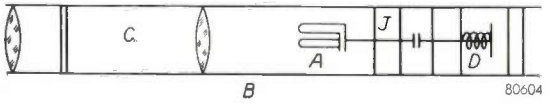


Fig. 1 Simplified diagram of a widely used type of pocket dosemeter. *A* sensitive system, *B* housing, *C* microscope, *D* charging contact and *J* insulator.

With the idea of effecting these improvements C.C. and T. Lauritsen have designed a simple pocket dosemeter with its own charger [3]).

This dosemeter is essentially an electrostatic meter, with the pointer mounted in bearings. It incorporates a charger operating on the principle of friction between two solids, and is approximately the size of a cigarette packet. Although moisture-proof, this instrument is not vacuum-tight, and so contains air at a pressure of 1 atm. It can be used either as a dosemeter, or as a dosage-rate meter.

A new design

A new dosemeter has been designed (*fig. 2*) based on the above considerations but having a vacuum-tight chamber. This allows complete freedom of choice as to the type and pressure of the gas contained in the ionization chamber, and so permits the attainment of a wide variety of sensitivities. Other advantages of this instrument are that it is inexpensive and can conveniently be carried in the pocket.

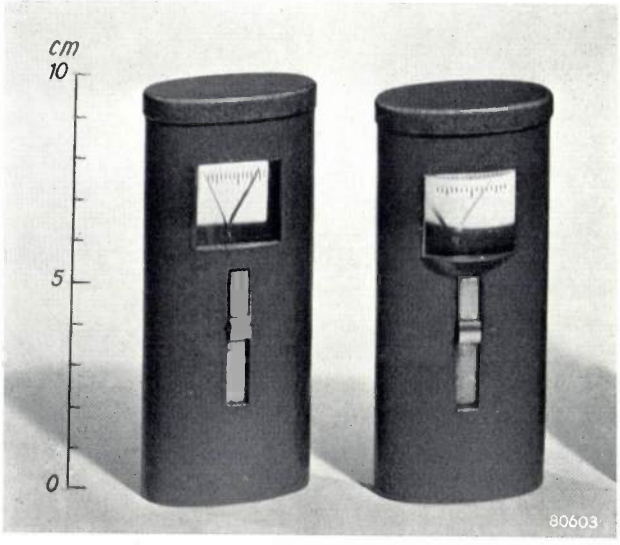


Fig. 2. Pocket dosemeter with built-in charger; the model on the right is fitted with a cylindrical lens window which acts as a magnifier.

The electroscope-system employed comprises two identical, rectangular conductive foils. The material used, besides being easy to trim and mount, is such that it neither sags appreciably under its own weight, nor vibrates unduly when in vacuum. The elasticity of this material must of course be consistent with a reasonable deflection of the system at moderate voltages. One end of the two foils (or leaves) is attached to an efficient insulator in the form of a rod or tube (*fig. 3*). It will be seen that the system as a whole bears a very close resemblance to the old gold-leaf electroscope; in the uncharged condition the leaves are parallel to each other. The electroscope system is housed in an oblong bulb of oval cross-section, made of glass having good conducting properties. This is to prevent the accu-

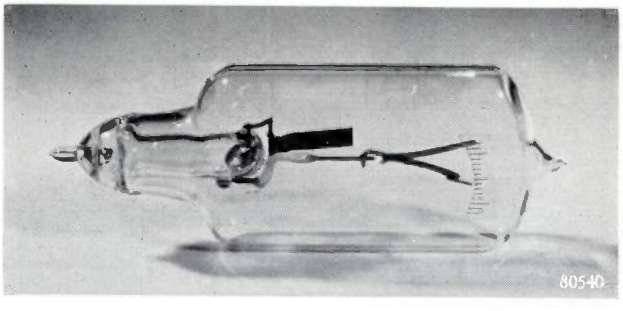


Fig. 3. The electrostatic system of the pocket dosimeter.

[3]) C. C. and T. Lauritsen, Science **112**, 137, 1950.

mulation on the bulb wall of charges which would interfere with the measurement. The sensitivity of this instrument is such that 2.5 kV, gives the full scale deflection of about 10 mm.

The insulator is mounted on the end of the charger incorporated in the instrument. The charger is a glass tube some millimetres in diameter and a few centimetres long containing a drop of mercury. When shaken to and fro in the tube, this mercury acquires a charge, which is transferred direct to a contact wire. Also included in the instrument is a switch, which, when in a certain position, establishes electrical contact between the wire in the charger and the sensitive system. When open, the switch is locked to the wall of the bulb by a small magnet, to eliminate the risk of the system being discharged by accidental contact with the charger whilst the instrument is being carried in the pocket. Charger and switch are sealed vacuum-tight into the bulb, so that the latter can be filled with a suitable gas and used, as already described, as an ionization chamber.

The bulb is enclosed in an aluminium or plastic case, which also contains the operating mechanism for the magnetically operated switch mentioned above. Two apertures, one on each side of the case and exactly opposite each other, are provided for reading the electroscope. The electroscope leaves themselves act as the pointers for this reading. A frosted plate on which a scale may be engraved, is fitted in the rear aperture. The scale can alternatively be inserted at the front, but in some models a large cylindrical lens of plastic is fitted in the front aperture. This facilitates reading generally, but also enables the meter to be read from a distance, for example when placed on a laboratory bench as part of an experiment which must be remotely controlled. In normal use the instrument is fixed in the waistcoat or breast pocket by means of a strong clip provided at the back.

To charge the dosemeter, the magnetic switch is pressed down and, if necessary, the flat back of the instrument tapped lightly on the hand to detach the switch from the wall of the bulb, so that it will drop onto the contact connected to the system. The position of the switch is readily visible. Next, the electrometer is shaken fairly vigorously, lengthwise and at an angle of 45° to the horizontal, with the window end held lower than the switch end. The leaves will then be seen to diverge; when they have reached the limit of deflection, the magnet is pushed towards the window, the meter being held in the same slanting position throughout this process.

Once charged, the instrument may be employed either as a dosemeter or as a dosage-rate meter; the variation in deflection during periods when the meter is not exposed to radiation is less than 1% per 24 hours. Shaking-tests have shown that, provided suitable materials are employed, the mercury-drop charger will continue to function almost indefinitely (i.e. after more than 1 million chargings) without any decrease in the voltage generated. During these tests, the complete instrument was shaken very vigourously to test its robustness; this is a quality usually difficult to produce in chargers operating in the principle of friction between two solids, and is important for an instrument which is to be carried in the pocket.

The sensitivity of an ionization chamber is governed primarily by the volume, nature and pressure of the gas filling, and to a lesser extent by the material constituting the wall of the chamber. A suitable choice of all these variables makes it possible to make radiation meters of this type in three distinct and widely different grades of sensitivity, all three being little dependent on the wavelength of the measured radiation. The most sensitive type, tested with the γ-radiation emitted by cobalt 60 (about 1.2 MeV), gives full scale deflection for a dosage of 0.2 röntgens. The sensitivity is not affected by any decrease in the hardness of the radiation until the energy drops to 300 keV; below this energy, the sensitivity increases until, at 150 keV, it reaches a maximum which is twice as high as the value at 1 MeV.(*fig. 4*). A further decrease in hardness to

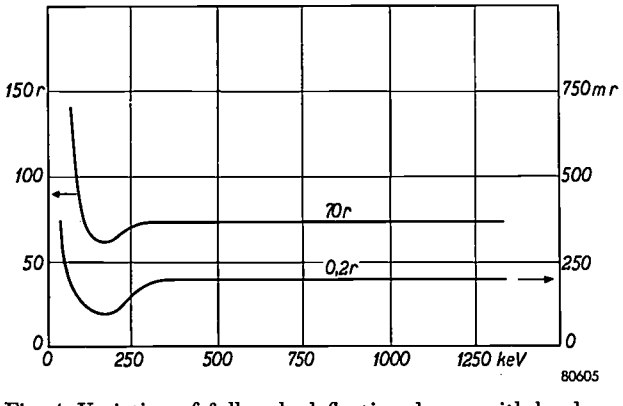


Fig. 4. Variation of full-scale deflection dosage with hardness of radiation, for dosemeters of nominal ranges 0.2 r and 70 r, respectively. (These measurements were carried out in collaboration with Mr. E. Harberink, of the Philips Laboratories in Eindhoven)

75 keV restores the sensitivity to its original level; any further softening of the radiation is accompanied by a decrease in sensitivity owing to absorption by the bulb. However, at 50 keV the sensitivity is still about 60% of the original value.

This relatively greater sensitivity to soft radiation

arises from the fact that it is impossible to employ in the dosemeter materials entirely equivalent to air; hence photo-electric X-ray absorption affects the measurements quite appreciably even at 100 - 200 keV.

The most sensitive model just described is designed primarily with a view to the protection of laboratory workers; a weekly tolerance dosage of 0.3 r is well within its range.

In the second model, full scale deflection corresponds to 70 röntgens; it will be seen from fig. 4 that the sensitivity of this particular instrument is governed only very slightly by the wavelength of the incident radiation; the maximum at 150 keV represents a mere 15% increase in sensitivity. This virtual independence of wavelength is obtained by employing a separate, built-in ionization chamber.

The third model covers dosages up to 250 röntgens at full scale deflection. The sensitivity is higher by a factor of about 2 at the maximum than at 1 MeV. By reason of the low pressure and high field intensity employed, the ion current in this model reaches saturation relatively slowly; this may be an advantage, particularly in the measurement of flash dosages. Model three can be adapted for the measurement of still higher dosages.

All these models are insulated so thoroughly that the discharge per month in the absence of radiation is at most one or two percent.

Other applications

Consider now the electrometer without the charger, as shown in *fig. 5*. As in the instrument already described, the whole of the bulb, other than the

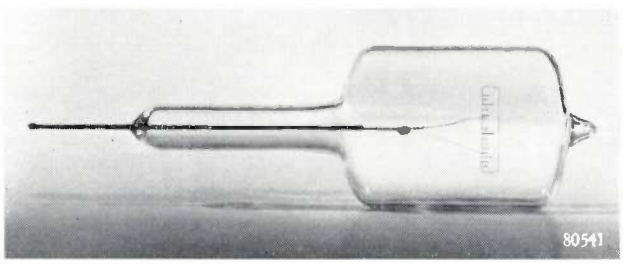


Fig. 5. Electroscope for experimental and educational purposes.

long neck to which the system is sealed, is made of glass having good conducting properties. The neck has good insulating properties, and is coated with a water-repellant material which enables the system to retain its charge for a long time. Being vacuum-tight, the instrument is not affected by

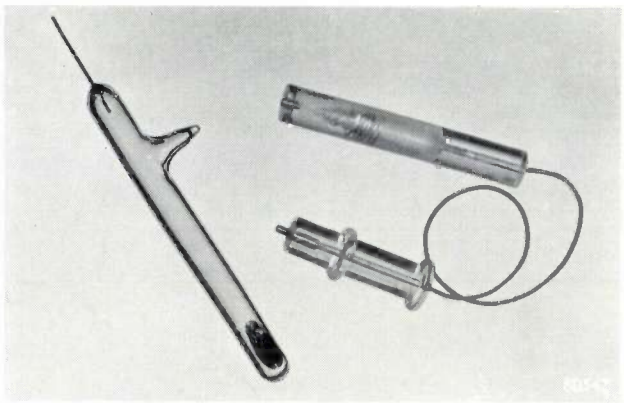


Fig. 6. Two types of mercury friction chargers. The maximum voltage obtainable in the absence of any current leak is about 3 kV.

humidity in the surrounding atmosphere. The sensitivity is 2.5 kV at a full scale deflection of 10 scale divisions. A considerable reduction of the sensitivity can be effected by employing thicker leaves. To determine whether an object is charged, it is only necessary to hold the instrument against or near to the object. The electroscope is easily fixed in an apparatus at any point where a measurement of voltage is required. It is therefore very suitable for use in schools and for demonstration purposes: an enlarged image of the system can be very simply projected on to a screen. The charging unit (*fig. 6*) can also be used individually in demonstrations. The long glass tube containing the drop of mercury and a sealed-in contact is best suited to this purpose; it is capable of generating 3 kV. By shaking the charger vigorously, it is possible to produce a current of some tenths of a μA. The other type of charger, fitted with a flexible connection, is suitable for charging the fountain-pen type of pocket dosimeter.

Summary. Following a brief recapitulation of the effects of ionizing radiations upon biological tissue, the principles of dosimetry, tolerance dosages and the effects of intense irradiation are considered. After descriptions of existing types of pocket dosimeter a more detailed account is given of a new design which incorporates a self-charger and which dispenses with a microscope. This instrument is fully enclosed in a vacuum-tight bulb of conductive glass, which, in turn, is housed in an aluminium case. Full scale deflections corresponding to dosages of 0.2 r, 70 r, 250 r, or more, can be obtained by adopting suitable dimensions for the instrument and by varying the gas pressure. The discharge in the absence of radiation is considerably less than 1% per 24 hours. The charger is a glass tube containing a drop of mercury, which during the charging process, is connected to the sensitive system by means of a magnetic switch. Charging is effected by shaking the instrument several times, with the electrometer system facing downwards. Attention also is drawn to the suitability of the individual electroscope and charger systems for demonstration purposes.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the administration of the Philips Research Laboratory, Eindhoven, Netherlands.

2093: K. S. Knol and J. Volger: Is a superconductor free from noise? (Physica **19**, 46-50, 1950, No. 1/2).

From measurements at 6 Mc/s it is concluded that no extra noise is generated in a ring of NbN when this ring is brought into a superconducting state, even when a persisting current of about 100 A is excited.

2094: G. H. Jonker and J. H. van Santen: Magnetic compounds with perovskite structure, III. Ferromagnetic compounds of cobalt (Physica **19**, 120-130, 1953, No. 1/2).

Polycrystalline mixed crystals of $(La,Sr)CoO_3$ have been prepared. Perovskite structure is found for all compositions. Ferromagnetism is observed for medium Sr concentrations. Curves are given for the saturation magnetization, the paramagnetic Curie temperature, and the effective paramagnetic moment as a function of composition. It is suggested that the ferromagnetism observed is caused essentially by a positive Co^{3+}-Co^{4+} interaction. The sign of the exchange interaction is discussed in connection with the theory of Anderson and Polder, and the theory of Zener.

2095: K. H. Klaassens and J. H. Gisolf: Polymerization in bulk at high pressures (J. Polymer Sci. **10**, 140-150, 1953, No. 2).

An apparatus is described for polymerizing monomers in bulk up to a pressure of 10 000 atmospheres, in which the polymerizing substance can be heated and the temperature of the substance measured. Styrene shows an explosive reaction at 10 000 atmospheres and at about 70 °C. Indene polymerizes slowly when heated at 10 000 atmospheres and shows an explosive reaction at 175 °C. Both compounds give a solid polymer. Croton aldehyde heated at 10 000 atmospheres gives a brittle, high-melting polymer. Coumarone and some chlorinated ethylenes carbonize when heated at 10 000 atmospheres. Butyraldehyde polymerizes to a solid product which rapidly reverts to the monomer.

2096: J. I. de Jong and J. de Jonge: The hydrolysis of methylene diurea (Rec. Trav. chim. Pays-Bas **72**, 202-206, 1953, No. 3).

The hydrolysis of methylene diurea giving urea and monomethylolurea was found to be a monomolecular reaction. The rate of the reaction is directly proportional to the hydrogen ion concentration in the *pH* range measured (3-5) and independent of the buffer concentration. The activation energy appears to be 19.5 kcal/mole. Generally, the reaction of an amidomethylol group with an amide group leading to the formation of a methylene bridge between urea fragments, will be a reversible reaction. The rates of both the forward and the reverse reactions are proportional to the hydrogen ion concentration.

2097: J. I. de Jong and J. de Jonge: Kinetics of the reaction between monomethylolurea and methylene diurea (Rec. Trav. chim. Pays-Bas **72**, 207-212, 1953, No. 3).

The reaction between monomethylolurea and methylene diurea appears to be bimolecular, and the rate constants were found to be directly proportional to the concentration of the hydrogen ions. An influence of the buffer concentration was not observed. The activation energy was found to be 15 kcal/mole.

2098: J. I. de Jong and J. de Jonge: The reaction of methylene diurea with formaldehyde (Rec. Trav. chim. Pays-Bas **72**, 213-217, 1953, No. 3).

The reaction of methylene diurea with formaldehyde shows a close resemblance to the previously studied reaction of urea and formaldehyde (see these abstracts, No. **2046**) i.e. the reaction proved to be bimolecular and the rates were found to be directly proportional to the concentration of hydrogen ions; an influence of the buffer concentration on the rate was also observed. Obviously the reaction is subject to general acid and/or base catalysis. The activation energy appeared to be about 15 kcal/mole. The values for the rate constants were almost the same as were found for the reaction between urea and formaldehyde.

VOL. 16 No. 5-6, pp. 141-180 NOVEMBER-DECEMBER 1954

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

FERROXDURE II AND III, ANISOTROPIC PERMANENT MAGNET MATERIALS

by A. L. STUIJTS, G. W. RATHENAU *) and G. H. WEBER. 621.318.124:538.246.2

*The $(BH)_{max}$ product, i.e. the maximum value of the product of the induction B and the magnetizing force H for points on the demagnetization curve, is a measure of the quality of any permanent magnet material. In Ferroxdure **), a ceramic permanent magnet material already discussed in this Review, values of $(BH)_{max}$ up to about 7200 J/m³ [0.9 × 10⁶ gauss oersted] are attained. This article describes the development of an anisotropic type of Ferroxdure, which has a $(BH)_{max}$ product comparable to that of modern magnet steels, viz. up to about 28 000 J/m³ [3.5 × 10⁶ gauss oersted].*

Characteristic properties of Ferroxdure

Ferroxdure **) is the name given to a new group of ceramic permanent magnet materials, some properties of which have been fully described in an earlier issue of this Review [1]). The magnetically essential phase of Ferroxdure I (which is now in quantity production) is the compound $BaFe_{12}O_{19}$, whose crystal structure is hexagonal, with close-packed oxygen ions. One of the oxygen sites in this structure is occupied by a Ba^{2+} ion; the Fe^{3+} ions occupy five non-equivalent lattice sites in the interstices between the oxygen ions.

The saturation magnetization of the compound $BaFe_{12}O_{19}$ is low compared with that of iron; extrapolation to absolute zero gives $J_0 = 0.66$ Wb/m² [$4\pi J_0 = 6600$ gauss] [2]) as against 2.18 Wb/m² [21 800 gauss] for iron. Again, at room temperature the value J_s for this compound is 0.475 Wb/m² [$4\pi J_s = 4750$ gauss], whereas the corresponding value for iron is 2.15 Wb/m² [21 500 gauss]. This relatively low saturation value and also the variation of the paramagnetic susceptibility as a function of temperature, suggest that the source of magnetization in the compound is what may be described as non-compensated antiferromagnetism, or ferrimagnetism [3]). Although the magnetic moment of each Fe^{3+} ion in this ionic compound is more than twice the moment per atom of metallic iron, only one third of the total moment of these ions is really effective, since the residual elementary moments are oriented antiparallel to each other and therefore cancel. Moreover, the magnetization is all the lower in relation to that of iron owing to the fact that the structure of the compound is greatly "diluted" with oxygen ions.

The strong coercive force exhibited by Ferroxdure is attributable to a pronounced magnetic anisotropy, in this case mainly crystal anisotropy, as will now be explained. In every Ferroxdure crystal there exists one direction of easy magnetization (preferred direction), parallel to the hexagonal axis, and only a strong magnetic field can turn the magnetization away from this preferred direction. It is to this distinct crystal anisotropy, combined with a relatively low saturation magnetization, that Ferroxdure owes its magnetic hardness. As a result of the low saturation magnetization, the couple

$$M = J_s H \sin \varepsilon, \quad \ldots\ldots \quad (1)$$

which a magnetic field at an angle ε to the preferred

*) Now professor at the University of Amsterdam; formerly with the Philips Laboratories at Eindhoven.

**) Known in U.S.A., U.K., and some other countries under the trade name "Magnadur".

[1]) J. J. Went, G. W. Rathenau, E. W. Gorter and G. W. van Oosterhout, Ferroxdure, a class of new permanent magnet materials, Philips tech. Rev. **13**, 194-208, 1951/52.

[2]) Rationalized Giorgi-units are employed throughout this article (see Philips tech. Rev. **10**, 55-60, 1948/49 and **13**, 194, 1951/52. For convenience, the formulae and values as expressed in unrationalized cgs-units (Gauss units) are also quoted, in square brackets.

[3]) See article referred to in note [1]).

direction can exert to rotate the magnetization away from this direction is relatively weak, whereas the couple

$$M = K \sin 2\Theta \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

tending to turn the magnetization back to the preferred direction is strong (K is the anisotropy constant and Θ the angle which the direction of of magnetization in a field H makes with the preferred direction).

Given this process of rotation alone, the coercive force, as computed for a material in which the preferred directions of the constituent crystals are distributed at random, is:

$$_JH_c = 0.96\, K/J_s\,.$$

For Ferroxdure at room temperature, $K = 3.1 \times 10^5$ J/m^3 [3.1×10^6 erg/cm^3] and $J_s = 0.475$ Wb/m^2 [$4\pi J_s = 4750$ gauss]: hence a value of $\mu_0 {}_JH_c = 0.79$ Wb/m^2 [$_JH_c = 7900$ oersted] is calculated. In practice, however, a value of about 0.3 Wb/m^2 [3000 oersted] is obtained in the case of Ferroxdure I, which suggests that Bloch wall movements, which occur far more readily, are also involved. Now, no stable Bloch walls can be formed in particles of diameter less than a certain critical value. In the case here considered, the critical diameter is about 1μ, which, again by reason of the distinct crystal anisotropy and relatively low saturation magnetization of the material, is considerably larger than the critical diameter for iron and cobalt crystals. Accordingly, an effective sintering process, producing a compact material whose constituent particles are smaller than the critical diameter, is the first essential. In practice, however, the sintered product will invariably contain a number of larger particles, which tend to diminish the coercive force by fostering the formation and movement of Bloch walls.

Permanent magnets are often evaluated in terms of the $(BH)_{max}$ product, that is, the maximum value of the product of the induction B and the field H for points on the demagnetization curve. The volume of material required to maintain a given magnetizing force in a particular air-gap, broadly speaking, is inversely proportional to the $(BH)_{max}$ product. Hence there is every reason to increase the value of this product as far as possible. The formula for an "ideally permanent" magnet ($_JH_c = \infty$, $\mu_0 B H_c = B_r$) is (see *fig. 1*):

$$\left.\begin{aligned}(BH)_{max} &= \frac{1}{\mu_0}\left(\frac{B_r}{2}\right)^2 \\ \text{or,}\quad \left[(BH)_{max} &= \left(\frac{B_r}{2}\right)^2\right]\end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (3)$$

In the case of Ferroxdure I, $\mu_0 {}_JH_c \gg B_r/2$ [$_JH_c \gg B_r/2$], which allows the above formula to hold reasonably well, as will be seen from the fact that the theoretical maximum is:

$$(BH)_{max} \approx \frac{1}{\mu_0}\left(\frac{0.21}{2}\right)^2 \approx 8800 \text{ J/m}^3$$

$$\text{or,}\quad \left[(BH)_{max} \approx \left(\frac{2100}{2}\right)^2 \approx 1.1 \times 10^6 \text{ gauss Oe}\right],$$

as compared with 7200 J/m^3 [0.9×10^6 gauss oersted], which is the value established by experiment.

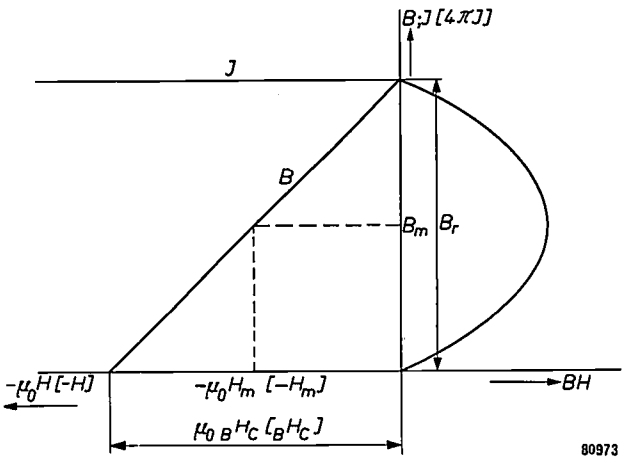


Fig. 1. Demagnetization curve for a material of which $_JH_c = \infty$. In the case here considered,

$$(BH)_{max} = \frac{1}{\mu_0}\left(\frac{B_r}{2}\right)^2, \quad \text{or} \quad \left[(BH)_{max} = \left(\frac{B_r}{2}\right)^2\right].$$

Whereas a low remanence limits the value of $(BH)_{max}$ in Ferroxdure I, the fact that $\mu_0 {}_JH_c \ll B_r/2$ [$_JH_c \ll B_r/2$] is mainly responsible for the limitation of this product in *metallic* magnets.

Ferroxdure I is very suitable for use in cases where a high coercive force is required, that is, where the material is exposed to very strong demagnetizing fields, as in focusing magnets for television tubes [4]), magnets for bicycle dynamos, yoke magnets and magnetic clutches. By virtue of its high resistivity, this material can be employed for producing a biasing field in high-frequency transformer cores of Ferroxcube [5]). Other advantages of Ferroxdure are its simple composition and low density, and also the fact that it is made of relatively inexpensive and readily obtainable materials. Moreover, the magnetization loss caused by demagnetizing fields is limited by the low permeability of this material.

The use of Ferroxdure I to sustain a strong magnetizing force in a narrow air-gap, as in a loudspeaker circuit, is confined mainly to those cases which

[4]) J. A. Verhoef, Philips tech. Rev. **15**, 214-220, 1953/54 (No. 7).

[5]) W. Six, Philips tech. Rev. **13**, 301-311, 1951/52.

favour the use of a very flat magnet. However, a complete change in the situation, particularly in this respect, has taken place as a result of a new development [6]), which has led to the production of Ferroxdure materials showing an appreciably higher induction on the demagnetization curve, and having $(BH)_{max}$ values comparable with those of modern anisotropic magnet steels.

Anisotropic Ferroxdure

In the isotropic material, the preferred directions of magnetization of the constituent particles are oriented at random. It is found that the remanence B_r of material so constituted is equal to half the saturation magnetization, i.e. about 0.21 Wb/m² [2100 gauss]. Given a material in which the preferred directions of the individual particles are aligned parallel to one another, however, the remanence in this direction will be the same as in a single crystal, viz.:

$$B_r = J_s, \qquad [B_r = 4\pi J_s].$$

Accordingly, we have as a maximum:

$$(BH)_{max} \approx \frac{1}{\mu_0}\left(\frac{0.42}{2}\right)^2 \approx 35\,000 \text{ J/m}^3$$

$$\left[(BH)_{max} \approx \left(\frac{4200}{2}\right)^2 \approx 4.4\times 10^6 \text{ gauss oersted}\right].$$

To procure this optimum value, it is necessary to ensure that in all circumstances: $\mu_0 J H_c > B_r/2$ $[J H_c > B_r/2]$. The fulfilment of this condition becomes more and more difficult as the remanence B_r increases. Moreover, the coercivity is invariably weaker in anisotropic than in isotropic material, owing to the fact that the changes in the direction of magnetization are caused by the formation of Bloch walls. The field H^* required to form a Bloch wall in a particle is a minimum when the preferred direction of magnetization of the particular particle coincides with the direction of the applied field H; also, the movement of Bloch walls across grain boundaries takes place far more readily in anisotropic materials, which in this respect resemble single crystals.

Accordingly, the problem is to procure an anisotropic orientation of the grains in Ferroxdure, at the same time avoiding any appreciable weakening of the coercive force.

Such anisotropy can be induced in several ways. For example, one approach to the problem is to take advantage of the non-spherical shape of the Ferroxdure single crystals. These crystals exhibit very strong growth along the basal plane, but relatively slight growth in the hexagonal direction coinciding with the preferred direction of magnetization. By packing the powder in a steel tube, welding the ends of this tube and passing it through rolls at a high temperature, a preferred direction of magnetization at right-angles to the plane of rolling is obtained. Similarly, it is possible to produce a small amount of anisotropy merely by compressing the Ferroxdure powder under high pressure.

However, a far superior method is to employ the magnetocrystalline anisotropy of Ferroxdure to the best advantage. When an external magnetic field H is applied, (*fig. 2*), the preferred axes of magnetization of any particles which are free to rotate will move into alignment with this field. This produces in the material a texture corresponding approximately to the magnetic state of a single crystal, that is, one preferred direction, parallel to the direction of the field.

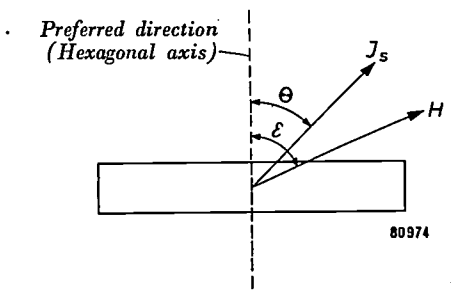


Fig. 2. The state of equilibrium of the magnetization J_s in a uniaxial magnetic particle when an external field H is applied at an angle ε to the preferred direction.

The aligning couple exerted by the magnetization J_s on the crystal is:

$$M = K \sin 2\Theta. \quad \ldots\ldots \quad (2)$$

The absolute value of this couple reaches a maximum equal to K when $\Theta = 45°$ or 135°, and is zero when $\Theta = 0°$, 90° or 180°.

It will be seen, then, that the highest gain in $(BH)_{max}$ is attained when those particles whose hexagonal axes make an angle ε of about 90° with the field direction are also brought into alignment. If at $\varepsilon = 90°$ the magnetizing force exceeds $2\,K/J_s$, the magnetization will rotate in the direction of H ($\Theta = 90°$), the orientation of the crystal axis remaining unchanged; the mechanical couple is then zero. The rotation of the magnetization is instantaneous; hence the application of the field does not usually produce a simultaneous rotation of the particle. Accordingly, the latter will also be

[6]) G. W. Rathenau, J. Smit and A. L. Stuijts. Z. Phys. **133**, 250, 1952. A. L. Stuijts, G. W. Rathenau and E. W. Gorter, J. appl. Phys. **23**, 1282, 1952.

able to remain out of alignment when the magnetizing force is increased, However, if H be small enough, e.g. $H = \sqrt{2}K/J_s$, a sudden application of the field will produce the state corresponding to $\Theta = 45°$, i.e. that with the maximum couple: hence the particle will also be brought into alignment with the field. This suggests that the value of an external magnetic field suddenly applied should be approximately equal to $\mu_0 H = \mu_0 \sqrt{2}\, K/J_s \approx 1.1$ Wb/m^2 [$H \approx 11000$ oersted].

Method of preparation

To apply in practice the process considered here, it is necessary to fulfil a number of conditions.

Firstly, the particles subjected to the field treatment must be magnetic: the basic material is isotropic Ferroxdure. Secondly, the basic material must consist of separate particles, each having only one crystal orientation; this condition can be fulfilled by grinding the sintered material until it contains no aggregates comprising more than one crystal.

A magnetic particle introduced into a magnetic field will turn to the particular direction most favourable from a magnetic point of view, if able to overcome the local forces opposing this orientation. To procure optimum efficiency, then, it is necessary to minimize the tendency of one particle to impede the orientation of another. Since the flatness of Ferroxdure crystals is unfavourable in this respect, too close a packing of the powder subjected to the magnetic field treatment has to be avoided.

When once the particles in the magnetic field are properly aligned, they must be stabilized without any disturbance of the texture thus obtained. By mixing the powder with a little melted paraffin wax and introducing this sludge into the magnetic field, stabilization is obtained when the wax solidifies. However, this method is of little practical value owing to the fact that the induction B is proportional to the density of the magnetic material, which would be low. In the case of ceramic materials, a product combining stability with adequate density is normally obtained by press-forming the powder at high pressure and then sintering it. Experience has shown that pressing is an excellent method of stabilizing the particles, provided that it takes place in a magnetic field; although the forces exerted during this process far exceed the aligning force of the magnetic field, they do not seriously disturb the texture of the material. As regards sintering, the condition that undue particle growth must be avoided in order to maintain a high coercive force is even more important than in the case of Ferroxdure I.

Finally, the high temperature employed in the sintering process, which is well above the Curie point, must not spoil the magnetic texture of the material. This is a point to which we shall shortly return.

The powder is press-formed as an aqueous sludge. A diagram of the apparatus employed is seen in *fig. 3*. The die is made of non-magnetic material,

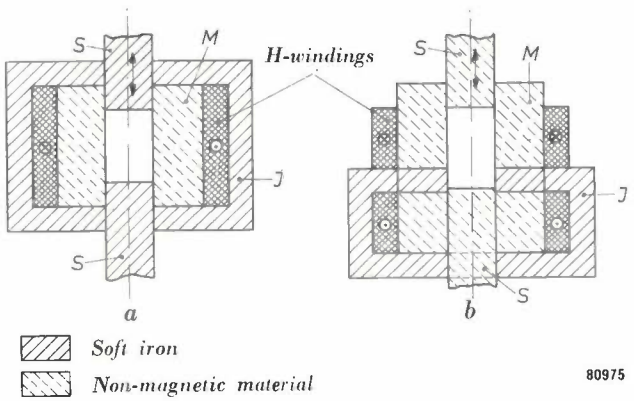


Fig. 3. Diagram of the dies M used to compress the powder in a magnetic field. S = plunger. J = iron yoke.
a) Direction of compression // to direction of field.
b) Direction of compression ⊥ to direction of field.

and iron pole pieces are provided to intensify the field within the powder. The surplus liquid is ejected along the plunger. Experience has shown that the relative directions of the compressing force and the magnetic field are immaterial.

After pressing, the powder still contains a large amount of water, and must be dried thoroughly before sintering. The technique of press-forming such an aqueous sludge differs considerably from that employed for dry powder and, in fact, the adaption of the process to large-scale production involved the solving of many technological problems.

The effect of sintering

It will be seen, then, that the texture of the compressed powder is not ideal; the particles obstruct one another during alignment, compression affects the texture, and the aligning couple decreases with angle ε (see fig. 2).

It is also necessary that the texture be preserved during sintering. However, the notable feature of this process is that it not only preserves, but actually enhances, the texture. Consider *table I*, which shows the remanence values of a particular material as measured parallel and at right-angles to the preferred direction, after sintering at different temperatures. In the case of a fully anisotropic material

Table I. Remanences of anisotropic Ferroxdure as measured parallel and at right-angles to the direction of the field in which the crystals are oriented, after sintering at different temperatures.

Sintering temp.	Density		Remanence $B_{r//}$ // to original field-direction		Remanence $B_{r\perp}$ ⊥ to original field-direction		$B_{r\perp}/B_{r//}$	Coercivity	
°C	kg/m³	g/cm³	Wb/m²	gauss	Wb/m²	gauss		$\mu_0 {}_JH_c$ Wb/m²	${}_JH_c$ oersted
1250	4060	4.06	0.270	2700	0.121	1210	0.45	0.321	3210
1275	4360	4.36	0.292	2920	0.125	1250	0.43	0.285	2850
1300	4660	4.66	0.326	3260	0.117	1170	0.36	0.247	2470
1320	4840	4.84	0.362	3620	0.072	720	0.20	0.117	1170
1340	4880	4.88	0.385	3850	0.025	250	0.07	<0.020	<200

$B_{r\perp}/B_{r//} = 0$, whereas in that of the isotropic material $B_{r\perp}/B_{r//} = 1$.

From the magnetic point of view, the texture changes from an imperfect one to that of a single crystal after the Ferroxdure has been sintered at a high temperature. *Fig. 4* shows part of the hysteresis loops of such a material, again measured parallel and at right-angles to the preferred direction. *Figures 5a* and *5b* are photo-micrographs showing the surfaces of a cube of the same material. Note the distinct texture produced by sintering at a high temperature. In the upper photograph the direction of the field applied during compression is at right-angles to the plane of the paper; in the lower photograph it is parallel.

This improvement of textural alignment during sintering could be explained by supposing the small particles in the material to be eliminated in favour of the larger ones by the action of interfacial tensions at the crystal boundaries during the sintering process; this explanation, if correct, implies that the smaller particles are not properly aligned.

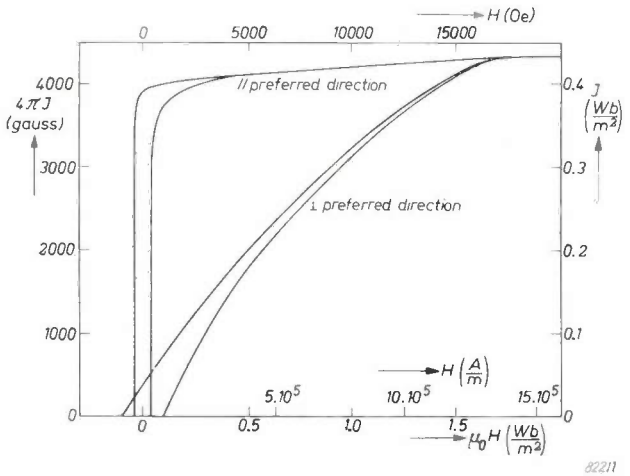


Fig. 4. Hysteresis loops (upper half only) of an isotropic Ferroxdure sample sintered at a high temperature, as measured // and ⊥ to the preferred direction. The magnetic texture closely resembles that of a single crystal.

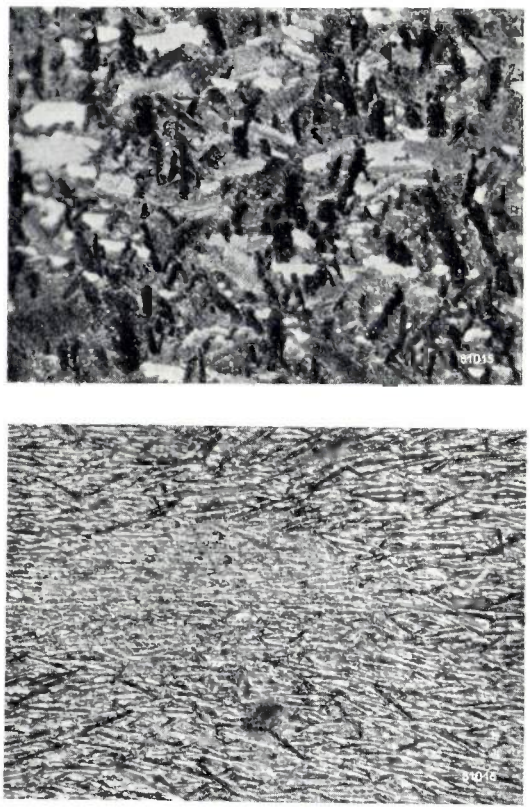


Fig. 5. Photo-micrograph of Ferroxdure with crystals aligned. In (*a*) the hexagonal basal planes are in the plane of the photograph, and in (*b*) they are at right angles to this plane. Magnification 10×.

However, experience has shown that the most significant improvement of the texture takes place when the density of the material is already about 90% of that of an ideal single crystal; it will be seen from the table that the coercive force then decreases very appreciably.

Accordingly, the following explanation is in our opinion more feasible, although, no doubt, the above-mentioned mechanism is also involved.

In the case of metals, the beneficial effect of grain growth on a preferred orientation has been closely studied. This effect can be observed direct with the aid of an emission electron-microscope [7]). It is seen that appreciable grain growth takes place only when the difference in orientation between adjacent crystals is such that the boundary energy between these crystals is high. Given a matrix of identically, or almost identically, oriented crystals surrounding one crystal differently aligned, this odd crystal will usually be eliminated, whereas the boundaries between the similarly aligned crystals will be preserved. Again, it is found that the odd crystal can grow only if it is large in relation to the crystals of the matrix.

[7]) G. W. Rathenau and G. Baas, Physica **17**, 117, 1951.

A calculation analogous to that suggested by C.S. Smith for polygonal crystals [8]) may be employed for the flat crystals of Ferroxdure. Consider *fig. 6*, which shows a cross-section of a thin, flat crystal, of thickness δ and diameter D, making an angle Θ with the neighbouring mutually aligned crystals, of thickness d. If σ_{dt} be the boundary energy per unit area between the crystals meeting at angle Θ, and σ_{tt} that between the aligned crystals, the odd crystal will be able to grow provided that this growth reduces the overall boundary energy, that is if

$$2\,D\,\sigma_{dt} < \frac{D\sin\Theta}{d}\,\frac{\delta}{\sin\Theta}\,\sigma_{tt}.$$

It has been assumed in this calculation that angle Θ is not small, and that $D \gg \delta$.

From the above formula, then:

$$\frac{\delta}{d} > \frac{2\sigma_{dt}}{\sigma_{tt}}.$$

Now the boundary energy σ_{dt} increases very sharply with the deviation from the general orientation; in most cases, then, the incorrectly aligned crystal will be eliminated.

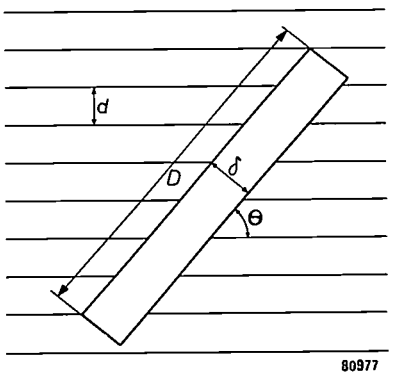


Fig. 6. Particle whose orientation deviates from that of the adjacent, mutually aligned crystals. As a general rule, such an oddly oriented crystal will be eliminated from the system.

Ferroxdure II and III

It will be seen from the above that any improvement in the anisotropy, that is, any increase in the remanence, is invariably associated with grain growth and therefore incompatible with a high coercivity (see table I). Accordingly, the theoretical maximum $(BH)_{max}$ value has so far proved unattainable in practice, owing to the fact that when the remanence is sufficiently high, $\mu_0 JH_c$ [JH_c] falls below the value corresponding to $B_r/2$. Nevertheless, the $(BH)_{max}$ values obtained by the method described above are considerably higher than those of isotropic material. It is possible to produce anisotropic materials with demagnetization curves specially suited to particular ranges of applications. At present, two types of anisotropic Ferroxdure magnet material are in production (see *fig. 7*).

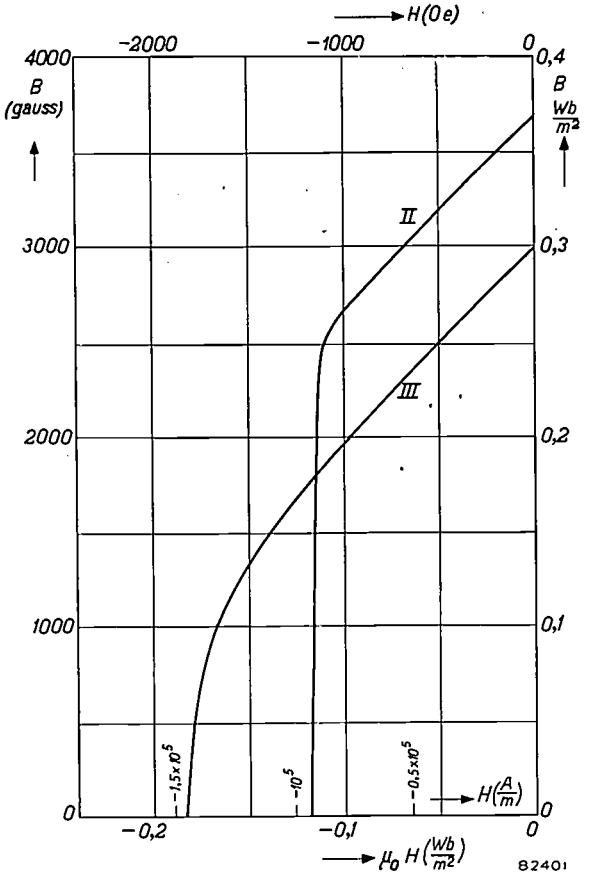


Fig. 7. Demagnetization curves of the anisotropic permanent magnet materials Ferroxdure II and Ferroxdure III.

One of them, Ferroxdure II, is intended for applications involving a fixed air-gap, in which high induction is the most important requirement, e.g. for loudspeakers. Its magnetic properties are as follows:

$B_r = 0.35 - 0.40$ Wb/m² [3500 - 4000 gauss]

$\mu_0 JH_c = 0.11 - 0.15$ Wb/m² [$JH_c = 1100 - 1500$ oersted]

$(BH)_{max} = 19\,000 - 24\,000$ J/m³ [$2.4 - 3.0 \times 10^6$ gauss oersted]

The other, Ferroxdure III, is intended for use in a strong demagnetizing field, or in circumstances where the working point may vary, as in magnetic clutches, dynamo magnets and yoke magnets. The magnetic properties of this material are:

$B_r = 0.28 - 0.32$ Wb/m² [2800 - 3200 gauss]

$\mu_0 JH_c = 0.19 - 0.25$ Wb/m² [$JH_c = 1900 - 2500$ oersted]

$(BH)_{max} = 16\,000 - 18\,500$ J/m³ [$2.0 - 2.3 \times 10^6$ gauss oersted].

[8]) C. S. Smith, Trans. Am. Inst. Mining Met. Engrs. **175**, 15, 1948.

It can be seen that in the case of Ferroxdure II, $\mu_0 {}_JH_c < B_r/2$ [${}_JH_c < B_r/2$], whereas for Ferroxdure III, $\mu_0 {}_JH_c > B_r/2$ [${}_JH_c > B_r/2$).

For the average properties of Ferroxdure types I, II and III at room temperature, see *table II*.

Table II. Average values of certain physical properties of Ferroxdure I, Ferroxdure II and Ferroxdure III at room temperature.

Magnetic moment per unit mass		8.75×10^{-5} Wb m/kg [70 gauss cm³/g]			
Density		4800-4900 kg/m³ [4.8-4.9 g/cm³]			
Density of ideal single crystal		5300 kg/m³ [5.3 g/cm³]			
Electrical resistivity		$> 10^6$ Ωm [10^8 Ωcm]			
Temperature coefficient of induction		—0.2 % per °C			
		Ferroxdure			
		I	II	III	
Remanent induction		0.205	0.37	0.30	Wb/m²
		2050	3700	3000	gauss
Coercivity	$\mu_0 {}_JH_c$	0.26	0.12	0.21	Wb/m²
	${}_JH_c$	2600	1200	2100	oersted
Coercivity	$\mu_0 {}_BH_c$	0.15	0.12	0.20	Wb/m²
	${}_BH_c$	1500	1200	2000	oersted
$(BH)_{max}$		7200	22 000	16 750	J/m³
		$0.9\cdot10^6$	$2.8\cdot10^6$	$2.1\cdot10^6$	gauss oersted
Working point	B	0.11	0.26	0.17	Wb/m²
		1100	2600	1700	gauss
	$\mu_0 H$	0.08	0.10	0.125	Wb/m²
	H	800	1000	1250	oersted
Curie point		450 °C			

A high temperature coefficient of induction for points on the demagnetization curve is still an inconvenient feature of the anisotropic materials Ferroxdure II and Ferroxdure III.

A magnetizing force equivalent to $\mu_0 H = 1.40$ Wb/m² [$H = 14\,000$] oersted is required to produce effective saturation in Ferroxdure I, whereas Ferroxdure II and Ferroxdure III can be saturated by a much weaker field.

In small-scale production, magnetic properties appreciably higher than those specified above can be attained. For example, a $(BH)_{max}$ of about 26 000 J/m³ [3.3×10^6 gauss oersted] was obtained on each type of demagnetization curve, Ferroxdure II and Ferroxdure III.

The best result obtained so far is a $(BH)_{max}$ of 28 000 J/m³ [3.5×10^6 gauss oersted] [9]; it will be seen from *fig. 8*, which shows the demagnetisation curve of the material concerned, that this is Ferroxdure II.

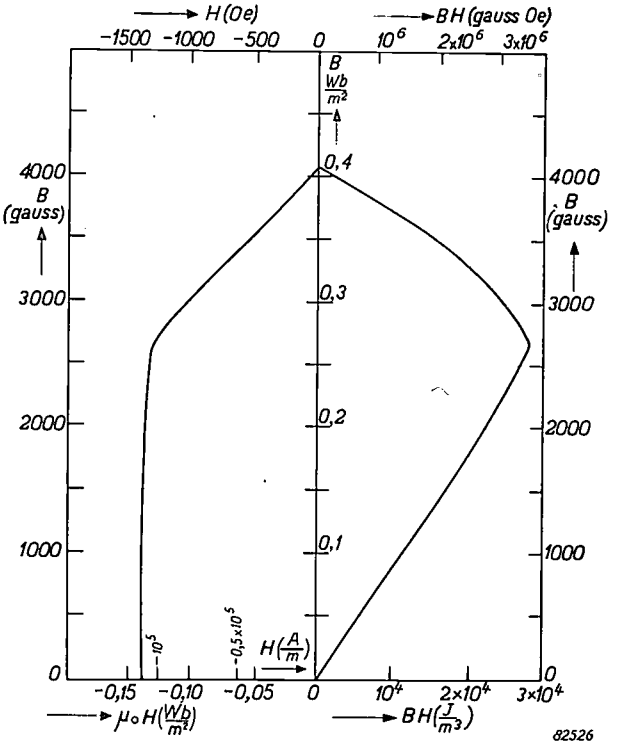
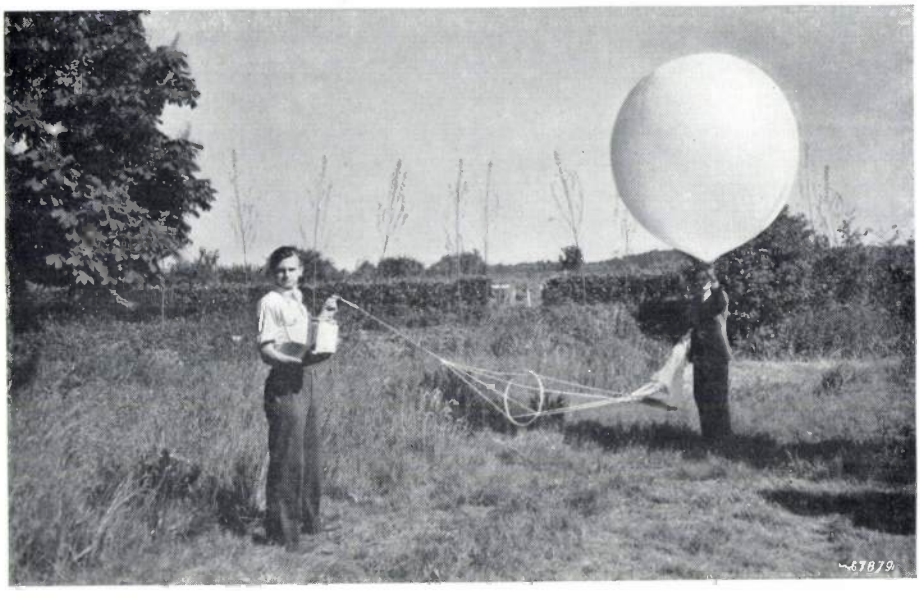


Fig. 8. Demagnetization curve of a particular sample of Ferroxdure II the $(BH)_{max}$ value of which is 28 000 J/m³ [3.5×10^6 gauss oersted].

Summary. Owing to the low remanence of Ferroxdure I, the $(BH)_{max}$ value of this material is likewise low. The preferred directions of the particles constituting Ferroxdure I, an isotropic material, are distributed at random; hence the remanence of this material is no more than half its saturation value. However, by virtue of the distinct magnetic anistropy of the constituent crystals, the remanence of such materials can be increased considerably by press forming the powder in a magnetic field before sintering it. Before being so pressed, the Ferroxdure powder is ground until the diameter of the individual grains is less than the critical value for the formation of Bloch walls. Pressing in a magnetic field aligns the hexagonal axes of the grains more or less parallel to the direction of the lines of force, thus producing in this direction $(BH)_{max}$ values 3 to 4 times as high as that of the isotropic material. It is found that grain growth taking place during the sintering process enhances the magnetic anisotropy of the material considerably, although the decrease in coercivity caused by this growth limits the possible $(BH)_{max}$ value to some extent. Two anisotropic materials, Ferroxdure II and Ferroxdure III, each having its own particular range of applications, are described.

[9]) F. G. Brockman and W. G. Steneck, Philips tech. Rev. **16**, 79-87, 1954/55 (No. 3).



A RADIO SONDE FOR METEOROLOGICAL OBSERVATIONS

by A. HAUER *) and M. van TOL.

621.396.91

Radio sondes, as nowadays commonly employed by meteorological services, must satisfy fairly stringent requirements as regards accuracy, and must at the same time be as light and inexpensive as possible. The sonde developed by Philips in collaboration with the Royal Dutch Meteorological Institute and described in this article embodies several new features of importance in the fulfilment of these requirements.

Others, apart from the writers, who assisted in the development of this sonde were Messrs. R. J. Ritsma and H. J. A. Vesseur of the Royal Dutch Meteorological Institute, Mr. J. L. M. Reijnders of the Philips Measuring Equipment Development Dept., and Messrs. H. van Suchtelen and D. J. H. Admiraal of the Philips Research Laboratories in Eindhoven.

The radio sonde

The study of physical conditions in the upper-air, i.e. in that part of the atmosphere not within direct range of measuring instruments on the ground, is an essential branch of modern meteorology. Knowledge so acquired is particularly valuable in the preparation of weather forecasts.

The principal quantities to be measured are the temperature and the humidity as a function of altitude, but accurate pressure measurements are also required as a means of determining the altitude. Observations from aircraft enable us to record these data at altitudes up to about 10 km, but are very expensive and can be carried out only in the vicinity of an aerodrome; hence another method, i.e. balloon-borne measuring apparatus is preferred. Balloons have already been used for many years by meteorologists as a means of measuring wind velocity, but other information concerning the physical condition of the atmosphere can also be obtained from an ascent, by attaching recording equipment for pressure, temperature and humidity below the balloon. The latter, filled with hydrogen, climbs at a constant rate of about 5 m/sec., and bursts on reaching an altitude of 15-25 km. A parachute then opens to lower the recording equipment (sonde) slowly to the ground.

It is necessary to search for, and recover the sonde when once it has reached the ground, a task which may often present some difficulty, particularly in sparsely populated areas. Other serious disadvantages of this method are that the results of the observations cannot be ascertained until, and are

*) Royal Dutch Meteorological Institute, De Bilt, Holland.

never known unless, the sonde is found; hence a period of hours, or even of days may well elapse between the recording and the actual reading of the data.

However, in about the year 1930, a method was evolved whereby the results of the measurements can be ascertained whilst still being gathered, that is, whilst the sonde is still airborne.

This method is that of the "radio sonde", now employed by most meteorological services. In principle, a radio sonde comprises two individual units. One of these, the measuring unit, contains instruments sensitive to temperature, humidity and pressure, and a circuit whose function is to pass the information to the other unit, that is, the transmitter. The latter is a small short-wave transmitter; the signal proceeding from this transmitter and received by equipment on the ground contains the results of the measurements in code form.

The radio sonde as a whole must satisfy one or two fundamental requirements, the most important of which is that it shall be sufficiently accurate (the accuracy considered ideal by meteorologists has not yet been attained by any radio sonde). It is likewise very important that the sonde be as light as possible, so that it will reach the highest possible altitude; this severely restricts circuit design, not least owing to the consequent limitation of battery capacity (the batteries make up a considerable proportion of the total weight). Hence the available power supply is very small.

To be suitable for regular, and at the same time economical use, the radio sonde must be inexpensive. In the Netherlands, the probability that it will not be recovered after an ascent is about 50 %, and even if recovered it may be seriously damaged (despite the parachute). In view of the latter possibility it is of course desirable that the radio sonde be mechanically robust.

In collaboration with the Royal Dutch Meteorological Institute at De Bilt, Philips have developed a radio sonde which embodies a number of unconventional features. A description will now be given, dealing with the items in the following order; the measuring unit, modulation of the transmitter by the measuring elements, the transmitter itself, and finally the results of several trial ascents by the new sonde.

Temperature, pressure and humidity measurements

To satisfy fully all the requirements of the meteorologists, a radio sonde would have to be capable of effecting extremely accurate measurements. In practice, however, the accuracy is limited to what is technically and economically practicable. Accordingly, the following requirements were taken as a starting point for the project here considered. Firstly, temperature measurements are to be accurate to within 0.5 °C, that is, in effect, to within 0.5 %, since the temperature in the particular regions of the atmosphere where the measurements are carried out varies on an average between 20° and —70 °C. Secondly, a similar accuracy is required in the measurement of pressure. The latter varies between 1050 and 70 millibars and the requirement is that it be measured accurately to within about 5 millibars, i.e. to within 0.5%. Thirdly, possible errors in the measurement of the relative humidity, which is expressed as a percentage of the saturation value, must not exceed 5—10 %.

The accuracy of such measurements depends firstly upon the particular measuring element in the sonde, secondly upon the manner in which the data are transmitted to the receiver on the ground, and thirdly upon the measuring instruments on the ground. In this article we shall consider only the first two factors.

Most radio sondes include several moving parts, the purpose of which is to convey the readings of the instruments measuring temperature, pressure and humidity mechanically to the transmitter, e.g. through a variable capacitance, resistance or self-inductance. Many also include a switching mechanism to connect the output of each measuring element in turn to the transmitter. This mechanism may be a rotary switch turned by a small windmill (which, in turn, is driven by the downward flow of air relative to the ascending sonde), a clock, or in some cases a small electric motor. Experience has shown that the reliability of sondes is often impaired by such mechanical parts.

In the Philips radio sonde the idea of switching was abandoned; measurements of the three quantities are transmitted simultaneously as three audio-frequency signals of variable frequency on a single carrier. This method necessitates the use of a separate audio-frequency oscillator for each of the quantities measured; altogether, then, three such oscillators are required, whereas if the signals were transmitted consecutively one oscillator would be sufficient. However, the switching mechanism has now been dispensed with, and by using suitable sub-miniature valves for the oscillators, the use of three oscillators has not increased either the total weight or the price of the radio sonde compared with earlier types. The oscillators will be considered presently. As we shall now explain, the number of moving parts in the sonde has been reduced to a minimum.

The measurement of temperature

The temperature measurement can be accomplished without the aid of any moving parts whatever by means of a temperature-sensitive element capable of acting direct upon the oscillator frequency, e.g. a resistor sensitive to temperature. This resistor must be made of a material having a temperature coefficient of resistance high enough to ensure the desired response. Such a material is to be found in the N.T.C. resistor, or thermistor [1]), that is, a semiconductor having a high negative temperature coefficient of resistance. It is found that a single N.T.C. resistor is capable of changing the frequency of an *R-C* oscillator by more than a factor of 2 for a decrease in temperature from 20° to —70 °C. For this purpose a straight-wire resistor was used whose resistance varies from about 30 kΩ at room temperature to about 500 kΩ at —70 °C. The change in the oscillator frequency per 0.5 °C (the desired accuracy) is then about 0.5 %, which can easily be measured.

The thermometer (see *fig. 1*) is a straight-wire N.T.C. resistor about 10 mm long, which is introduced direct into the frequency-controlling network of an audio-frequency oscillator (*R-C* oscillator). One advantage of this resistor over the bimetallic strip thermometer often used in other sondes is that it is very much quicker in response: it was found during trial ascents that the N.T.C. resistor responds quite noticeably to temperature fluctuations not detected by a bimetallic thermometer. Again, the "radiation error", which may be very appreciable in the case of a bimetallic strip (up to 10 °C), is very much smaller when a thin wire thermometer is employed. Hence the thermometer wire can be mounted quite freely about 10 cm outside the housing of the sonde, without any radiation screen or other cover.

The measurement of pressure

Metal aneroid barometers have been used as pressure-meters in almost all the radio sondes produced hitherto. However, the lag in response of such a barometer is very liable to introduce systematic error into the pressure measurements; moreover, the instrument itself is difficult to manufacture. For these reasons, and also to dispense with moving parts, another method based on the principle of the hypsometer was adopted, at the suggestion of the Meteorological Institute.

A hypsometer is a vessel, open to the atmosphere, containing a liquid maintained by some means at boiling-point; the boiling-point of the liquid depends upon the pressure of the surrounding atmosphere, that is, it decreases with this pressure. If the precise relationship between pressure and boiling-point is known, then, the one can be determined by measuring the other (e.g. by means of an N.T.C. resistor immersed in the liquid). The hypsometer is very reliable as a pressure-meter provided that the liquid used in it is sufficiently pure, since for pure liquids the relationship between boiling-point and pressure is very well established.

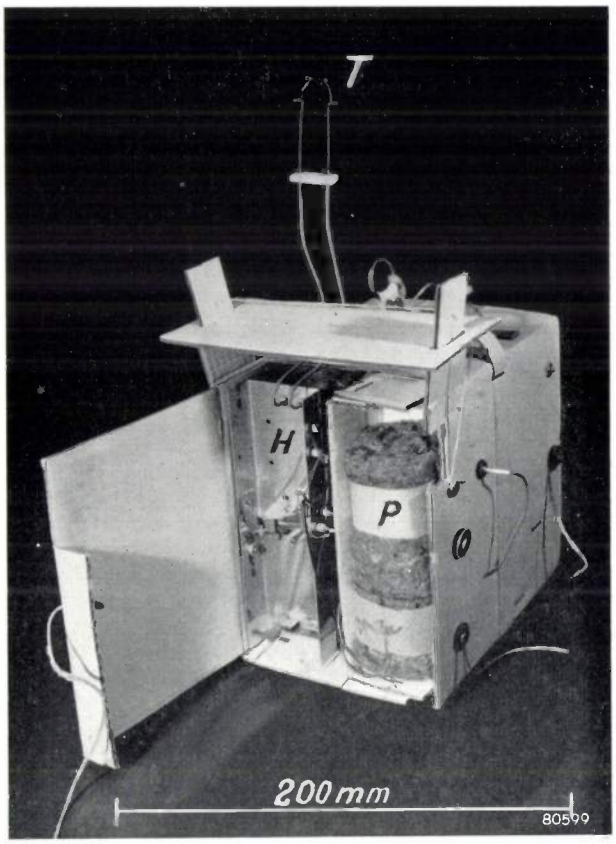


Fig. 1. The Philips radio sonde. The entire instrument is housed in a cardboard case, 165 × 120 × 155 mm. The temperature-measuring element *T* is a straight-wire N.T.C. resistor 10 mm long, mounted outside the sonde; the pressure and humidity measuring-elements, *P* and *H* respectively, are partly visible at the open side of the sonde. Altogether, the sonde weighs 500 grams; it is suspended from a hydrogen-filled balloon by a line about 10 m long, 5.5 m of which is covered with conducting sleeving to act as an aerial.

To maintain the liquid at boiling-point, at which some of it evaporates, a certain amount of energy is required. This would of course be a big disadvantage if batteries were the only source of energy available, since these would increase the weight of the sonde. However, if we have a liquid whose boiling-point will remain below the temperature of

[1]) E. J. W. Verwey, P. W. Haaijman and F.C. Romeyn, Semi-conductors with large negative temperature coefficient of resistance, Philips tech. Rev. 9, 239-248, 1947/1948.

the surrounding atmosphere throughout the ascent, the energy to keep the liquid boiling will be supplied by the atmosphere itself. This is in fact practicable since although the temperature of the atmosphere drops as the radio sonde ascends, the accompanying decrease in pressure causes a similar drop in the boiling-point of the liquid. Curves *1* and *2* in *fig. 2* show the relationship between decreasing pressure and decreasing temperature in the atmosphere (these curves are based on the summer and winter averages, respectively, of observations carried out at various times and places). Accordingly, we must choose a hypsometer-liquid whose pressure/boiling-point curve is to the left of curve *2* in fig. 2. At the same time the two curves should not be too far apart or the liquid will evaporate too quickly and it will be necessary either to insulate the hypsometer heavily, or to fill it with a very large amount of liquid. Freon, the vapour pressure of which is indicated by curve *3* in fig. 2, is a liquid which satisfies these conditions very well indeed.

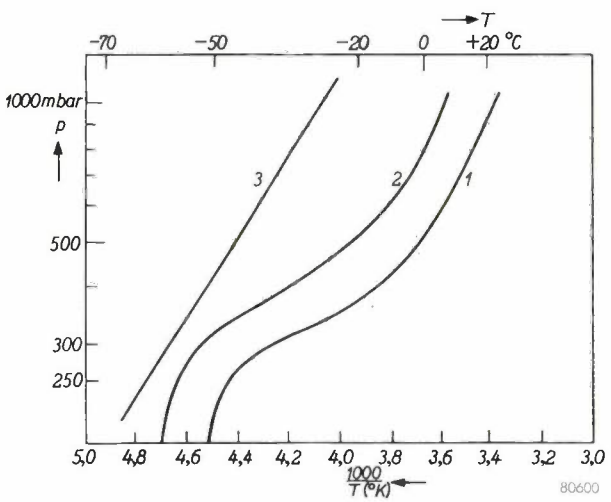


Fig. 2. Variation of temperature with pressure in the "average" atmosphere during summer (curve *1*) and winter (curve *2*) Line *3* represents the vapour-pressure curve of freon. The pressure of the atmosphere p is plotted on a logarithmic scale against the reciprocal of the absolute temperature T, so that curve *3* is then nearly a straight line.

There is a limited region of the atmosphere whose average temperature (in winter) is very close to the boiling-point of freon, so that on cold days the temperature of the air in this region may drop below the boiling-point of the hypsometer liquid. Apparently, then, freon fails to satisfy the condition governing spontaneous boiling, but despite this the liquid is prevented from "going off the boil" by a second source of heat within itself. This may be explained as follows. During the ascent of the sonde the entire mass of liquid, the temperature of which cannot of course be any higher then is consistent with the continually decreasing boiling-point, must dissipate more and more of its heat content. This heat helps to promote evaporation and the power thus made available will be all the greater, the more rapid the decline of the boiling-point, that is, the higher the rate of ascent of the sonde. Provided that the rate of heat transmission through the insulation of the boiling-vessel to the relatively cold surroundings is less than the rate at which energy from the heat of the liquid becomes available, a certain surplus of heat will always be available to keep the liquid boiling.

It is found that a relatively small quantity of freon (10 cm³), insulated with 1 cm thickness of cotton wool, is sufficient to supply the hypsometer during an entire ascent. The vessel used to contain the liquid is an ordinary radio-valve bulb; the N.T.C. resistor is secured to two lead-in pins through the bulb in such a way as to be completely immersed in the liquid (*fig. 3*). The resistor, like the air-temperature measuring element, is part of the frequency-controlling network of an *R-C* oscillator. The outer air has access to the liquid through a narrow tube (exhaust stem) at the top of the bulb; a completely open vessel would be unsuitable since, apart from the chance of spilling, it would allow the freon to evaporate too quickly and so perhaps assume a temperature below the boiling-point. On the other hand, access to the surrounding atmosphere should not be too severely restricted, in view of the possibility of the liquid overheating.

Tests have shown that a pressure variation of

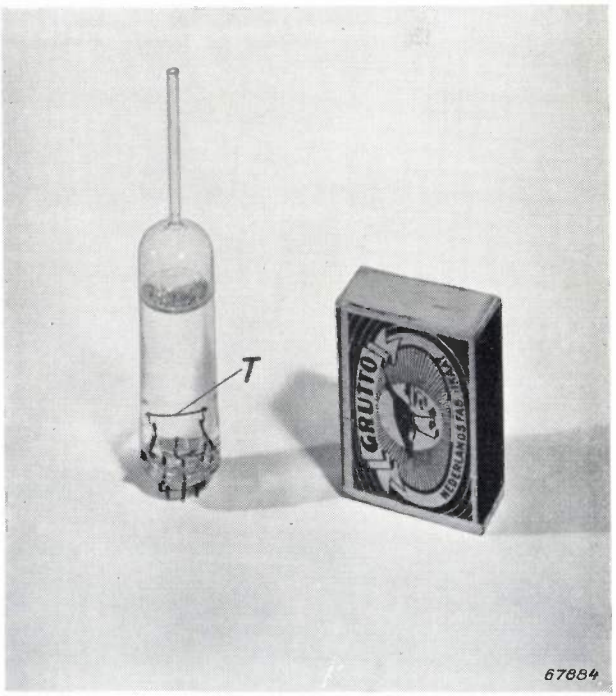


Fig. 3. Hypsometer. This is made from a radio-valve bulb containing freon, the boiling-point of which is measured by means of an N.T.C. resistor T secured to two lead-in pins in the bulb and completely immersed in the liquid; the latter has access to the outer air through a tube (the exhaust stem) at the top of the bulb.

5 millibars corresponds to a variation of the boiling-point such as to produce a relative variation of 0.2 % or more in the oscillator frequency.

The measurement of humidity

In the Philips radio sonde, as indeed in several others, a piece of gold-beater's skin is used as a means of measuring the relative humidity of the atmosphere. This skin is an animal tissue, the chief constituent of which is keratin, and like human hair, expands and contracts with changes in the humidity of the surrounding atmosphere. Unfortunately this material is affected by hysteresis when it passes from a humid to a dry state, and then back to the humid state. For this reason the electrical resistance of a solution of lithium chloride is employed instead of gold-beater's skin as a measure of humidity in sondes of one American make. This is a very sensitive method, but one which requires alternating current to make it fully effective and moreover, involves the use of rather complex measuring equipment; hence gold-beater's skin is preferred in the radio sonde described here, despite the above-mentioned defect.

The variation in the length of this skin affects the width of the air-gap in the magnetic circuit of a coil (*fig. 4*) and is thus converted into a variation of the self-inductance of this coil. The coil is part of an audio-frequency *L-C* oscillator, the frequency of which varies by approximately a factor of 2 for a variation in relative humidity from 10 to 100 %.

The audio-frequency oscillators

As we have already seen, mechanical switching is dispensed with in the Philips radio sonde by virtue of the fact that the three audio-frequency signals corresponding to the three quantities to be measured are modulated on one carrier wave and transmitted simultaneously. These signals reach the ground station receiver continuously, and the response time of the recording equipment connected to this receiver does not have to be very small since only gradual variations take place in any of the quantities measured.

Accordingly, there is an audio-frequency oscillator corresponding to each measuring element in the sonde. Since the sensitive elements in the temperature and pressure measuring devices are resistors, the obvious course was taken in adopting *R-C* circuits for the two associated oscillators. As already explained, the oscillator coupled to the hygrometer is of the *L-C* type. The frequencies of the three oscillators are adequately spaced; hence the three signals are readily separated in the receiver.

The frequency bands are as follows:

1— 3 kc/s for pressure;
4— 8 kc/s for temperature;
12—25 kc/s for relative humidity.

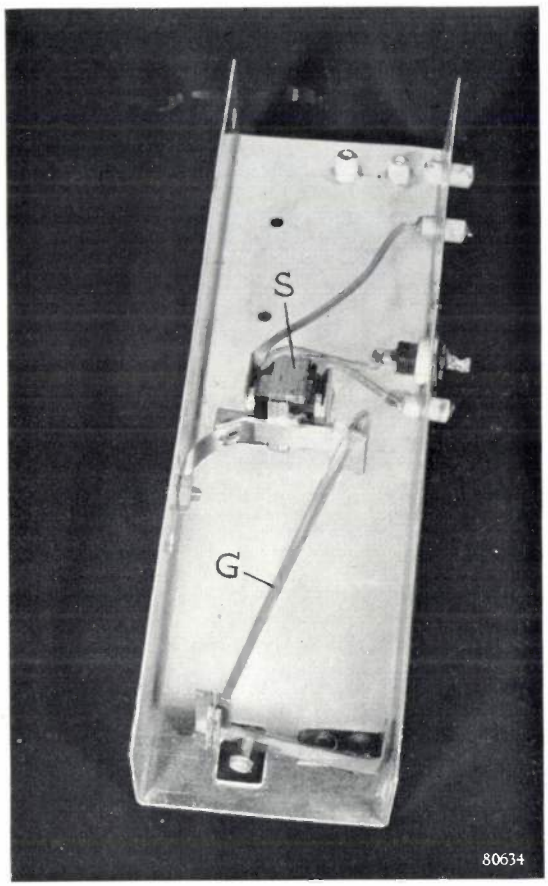


Fig. 4. Hygrometer. The variation in length of a piece of gold-beater's skin *G* is converted into a variation in self-inductance by virtue of the fact that the skin controls the width of the air-gap in the magnetic circuit of coil *S*.

The *R-C* oscillators must satisfy certain requirements:

1) The input power must be as low as possible (with a view to the weight of the batteries).
2) The number of components should be as small as possible (with a view to the price and total weight of the sonde).
3) All the components of the different networks (apart from the N.T.C. resistors, of course) should remain as far as possible unaffected by temperature and pressure variations.
4) The circuit should be insensitive to battery-voltage variations.

To satisfy requirement 1), sub-miniature valves

are employed. These are very small valves designed to operate with batteries; the type used in the oscillators is the DL 67, which has a diameter of 7.5 mm and a length of 35 mm. Valves of the same type are used in hearing-aids, etc., and were originally developed for military purposes, e.g. for proximity fuses. *Fig.* 5 shows the DL 67, together with a valve of the type DL 41 also employed in the sonde. The former are wired direct into the circuit.

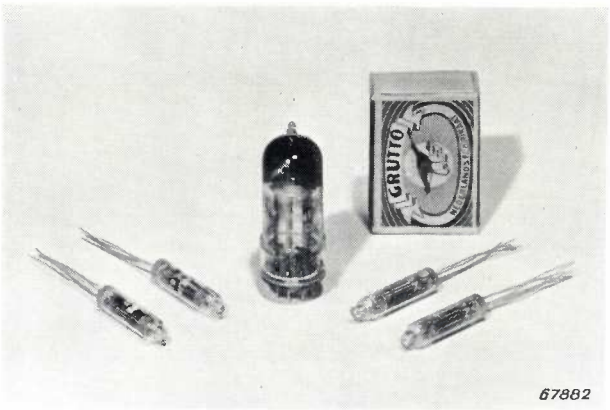


Fig. 5. The five valves used in the radio sonde. The four small valves are battery-operated sub-miniatures type DL 67 (length 35 mm, diameter 7.5 mm), three of which are employed in the three oscillators and the fourth as a reactance valve. Centre: the battery-operated valve type DL 41 (dimensions 20×55 mm) used as the transmitting valve of the radio sonde.

In as far as they relate to the *L-C* oscillator, the other requirements can be satisfied merely with the aid of standard components (of the smallest possible size), but the *R-C* oscillators present a more difficult problem. The circuit diagram of an *R-C* oscillator is shown in *fig. 6*, from which it will be seen that a triple *R-C* network is employed as feedback between anode and control grid (this is the minimum number of components required to produce the phase-inversion necessary for oscillation). Small carbon resistors and ceramic or miniature mica capacitors are used in the networks.

Temperature variations affect the circuit mainly through the carbon resistors (and the batteries), the capacitances being virtually independent of the temperature. However, the effect of such variations can be minimized by insulating the circuit as thoroughly as possible from the surrounding air, so as to prevent the transfer of heat.

There is a direct connection between the above and the fourth requirement, which may be explained in the following manner. During the ascent of the sonde there is a slight decrease in the battery voltages, partly owing to the low capacity of the batteries, and partly to the fact that these voltages also depend upon the temperature. To fully appreciate the deleterious effect of this decrease, we must re-examine the diagram shown in fig. 6.

If the circuit is to oscillate, the attenuation caused by the network of capacitors and resistors must be less than the amplification effected by the valve; at the same time, an unlimited increase in the amplitude of the oscillations must be avoided. Normally, then, such an oscillator circuit will contain either an incandescent lamp or an N.T.C. resistor, so positioned that the positive feedback of alternating current in the circuit decreases with the amplitude of the oscillations. However, amplitude limitation by this method cannot be employed in our circuit owing to the fact that the A.C. power available is very low. Hence we make use of a certain "natural" restriction of amplitude arising from the fact that the grid of the valve is conducting during a small part of each cycle. As a result of this, the control grid acquires a negative bias (the grid capacitor is charged), which reduces the mutual conductance of the valve. As the amplitude of the oscillations increases, the voltage on the control grid becomes more and more negative, until the mutual conductance is so reduced that the amplification by the valve is only just sufficient to compensate for the attenuation by the network. The amplitude of the oscillations then remains constant.

However, during the time that the grid is conducting, a new conducting link is formed between the grid and the cathode (effective for A.C. as well as for D.C.), so that in effect an extra resistance R_s is introduced in parallel with the terminating impedance R of the network (*fig.* 7); hence a slight change takes place in the frequency of the oscillator, which is governed by the capacitance and the resistances of the network.

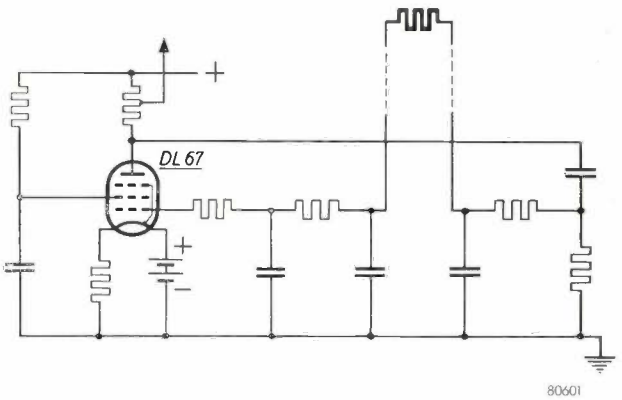


Fig. 6. Circuit diagram of *R-C* oscillator, two of which are employed in the radio sonde. The measuring resistor is drawn in bold lines.

Although insignificant in itself, this frequency shift is governed by the battery voltage, since the amount of the extra resistance depends upon this voltage. The reason for this as follows. Following a reduction of the battery voltage, and thus a reduction of the mutual conductance of the valve, a relatively smaller negative grid voltage is sufficient to bring about the further decrease in mutual conductance then required. Moreover, the grid current is also somewhat reduced during this process

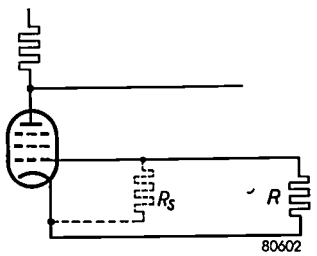


Fig. 7. Equivalent circuit of the grid circuit of the oscillator valve in an *R-C* oscillator, showing the apparent resistance R_s resulting from the flow of grid current. R is the total terminating impedance of the filter.

When the grid A.C. voltage is high, the grid current is, to a first approximation, proportional to it, so that R_s is then constant. This proportionality does not exist however at the low grid voltages which occur in the present circuit. In the case here considered, then, R_s depends upon the grid current, that is, upon the battery voltage.

To minimize this effect two measures were adopted. Firstly, a high-value resistor was included in the screen grid circuit to moderate the effect of battery-voltage variations upon the screen grid voltage, that is, upon the mutual conductance of the valve. Secondly, the control grid was given an initial negative bias such as to reduce the grid current as far as possible without causing too great a decrease in amplification. Since the heater voltage of the DL 67 is only 1.25 volts whereas the available battery voltage is 2.3 V, the required bias is readily extracted from the difference between the two by including a resistor in the cathode circuit.

As it happened, the battery effect persisted despite these two measures to prevent it. However, it was found that the frequency shift is also governed by the anode resistance, and since, amongst other things, the decrease in the anode-battery voltage causes a frequency shift opposed to that resulting from the decrease in the voltage of the heater battery, the entire effect can be practically eliminated by choosing a suitable anode resistance. Experiments have shown that in this way the frequency shift caused in each oscillator by a decrease of about 10% in the two battery voltages can be limited to less than 0.2%.

In practice the decrease in voltage during an ascent is invariably less than 10% [2]) and, as will be seen from the accuracy requirements already defined for the pressure, temperature and humidity, the frequency shift is then sufficiently small.

The transmitter

It was at first intended to employ 1 metre as the wavelength of the high-frequency carrier to be modulated by the three audio-frequency signals, The position of the sonde could then be determined at regular intervals during the ascent by means of a radio-theodolite in order to measure wind-velocities, as was, in fact, originally the principal object of such balloon ascents. However, since the bearings so obtained become rather inaccurate when once the sonde has travelled any considerable distance from the ground-station (>50 km), and also owing to the difficulty of designing a simple, one-valve high-frequency oscillator suitable for modulation at such a short wavelength, the idea of position-finding was abandoned for the time being and a wavelength of 11 metres (28 Mc/s) was adopted for the carrier wave (this wavelength is employed in several other radio sondes). The battery valve type *DL* 41 (see fig. 5) is used for the transmitter, since it supplies sufficient power for our purpose (about 100 mW).

Fig. 8 shows the circuit diagram of the high-frequency oscillator; the aerial is a conducting sleeve one half-wavelength in length round the wire on which the sonde is suspended from the balloon. In principle, either amplitude modulation or frequency modulation could be employed; the latter invariably necessitates the use of an intermediate stage, whereas, in theory at least, the former enables the modulating voltage to be applied direct to the control grid of the transmitter valve. However, in the case here considered the output power of the oscillators is insufficient for direct application, and since for this reason alone it is essential to include an intermediate stage, frequency modulation is preferred, by reason of the fact that the signal-to-noise ratio is very much more favourable. Accordingly, this method of modulation was adopted.

The three signals of the "pick-up" oscillators, separated by intervals between the frequency bands, are mixed in a potentiometer circuit (see fig. 8)

[2]) In fact, the batteries are accumulators, specially designed for use in radio sondes, the voltage of which depends only very slightly upon temperature.

and are applied to the grid of a "reactance" valve (again a DL 67), the principle of which has already been described in this Review [3]); the effect of such a valve is the same as that which would be produced

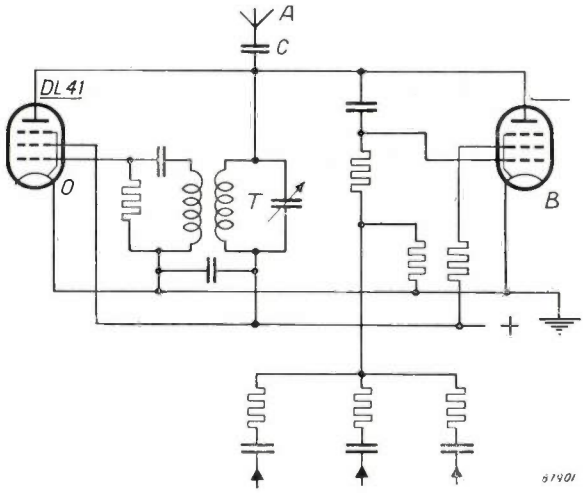


Fig. 8. Circuit diagram of the high-frequency oscillator containing the reactance valve. The three audio-frequency signals are mixed in a potentiometer and applied to the grid of the reactance valve *B*. The frequency of the H.F. oscillator *O* is governed by the voltage on this grid. The oscillator is connected across a capacitance *C* to the $\lambda/2$ aerial *A*. The frequency of the oscillator can be adjusted by means of trimmer *T* just before the ascent of the sonde.

by a capacitance or self-inductance (in this case a capacitance) in parallel with the high-frequency oscillator circuit, its value being governed by the negative voltage on the grid of the valve. This voltage, which varies at audio frequency with the frequencies of the three oscillator-signals, thus modulates the frequency of the oscillatory circuit.

One factor affecting the accuracy of the meteorological data as received on the ground is the frequency sweep of the transmitter; the greater this sweep, the stronger the audio-frequency signal in the receiver. In this case, a maximum frequency sweep of 25 kc/s was adopted as being the most consistent with the accuracy requirements imposed and the possibilities of the circuit.

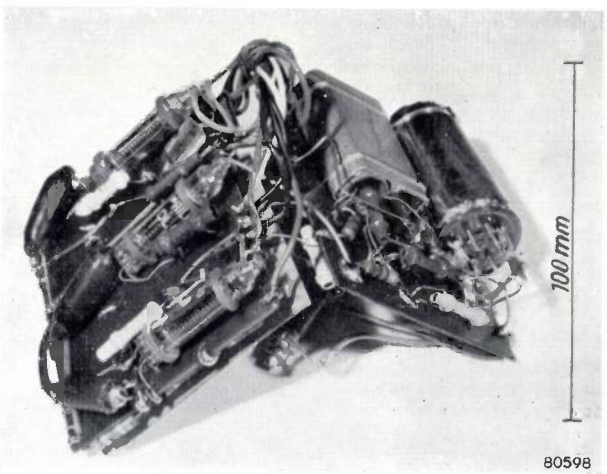
Constructional details of the sonde; features of the ground-station equipment

The high-frequency section is completely screened from the remainder of the sonde, to eliminate possible interference; the carrier-wave frequency can be changed slightly just before an ascent by means of a small trimming capacitor provided for the purpose.

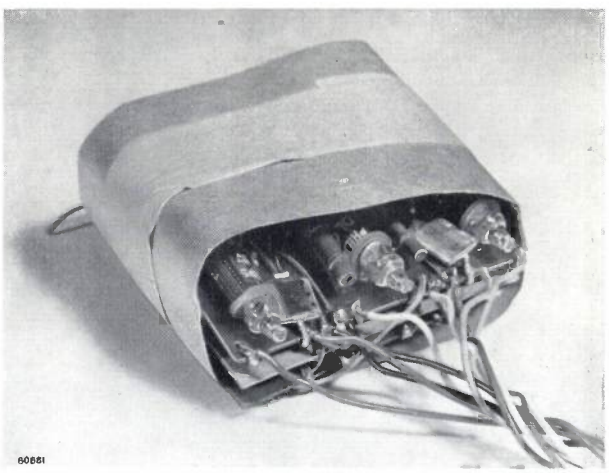
The transmitter and the oscillators are together accommodated in a robust cardboard cover; this greatly increases the probability that they will be recovered intact. *Fig. 9* shows the interior structure of the sonde.

A brief description of the ground-station receiving equipment will suffice. In principle, this may comprise a frequency-modulated receiver with three audio-frequency bandpass filters, one for each of the quantities to be measured, and three frequency meters to determine the frequencies of the signals passed by the filters. Preferably, the meters used to measure the frequencies should be of the recording

[3]) For exemple, see Th. J. Weyers, Frequency modulation, Philips tech. Rev. 8, 42-50, 1946 (particularly page 47).



a



b

Fig. 9. *a*) Electronic section of the radio sonde. The three oscillators are mounted on the left-hand panel and the transmitter on the right-hand panel. *b*) In the assembled sonde the two panels lie back to back and are enclosed in a cardboard cover. The measuring elements and batteries are connected to the panels by flexible leads.

type. *Fig. 10* shows the receiving equipment at present installed in the Royal Meteorological Institute at De Bilt.

Fig. 10. Photograph of the ground-station receiving equipment as installed in the Royal Dutch Meteorological Institute. The F.M. receiver is seen at the centre, and the recording meter on the right. At the extreme left is the calibrating equipment which supplies a constant comparison-frequency. Calibration is preferably effected with the aid of a cathode-ray oscillograph.

Testing the radio sonde

To test the accuracy of the radio sonde described in this article, the Dutch Meteorological Institute carried out several "twin-ascent" trials, that is, they sent up two such sondes suspended from one balloon, and compared the temperature and pressure data transmitted by these instruments. It was found as a result of these trials that the pressure measurements of the two sondes agreed to within 5 millibars up to an altitude of 5000 metres, and that the differences at altitudes up to 10000 metres did not exceed 10 millibars. The temperatures so measured agreed to within 0.5 °C.

Although the number of observations carried out during these trials is too small to justify any definite conclusions, the results show that the pressure measurement is at least as accurate, and the temperature measurement about twice as accurate, as those quoted in the literature for other sondes. Moreover, it is found that the results produced by the hypsometer at the temperatures occurring within its practical range of operation are entirely independent of the temperature: hence the pressure measurements of the sonde do not require correction to compensate for possible temperature fluctuations.

Again, it appeared from the twin-ascent trials that the new sonde is very convenient to handle, and that it is not easily damaged. Another important feature of the sonde revealed by these trials is its relative lightness; the combined weight of two of the new sondes is still less than that of one of the type hitherto employed.

A number of the new instruments were recovered after ascents, and there were very few instances of damage to the electronic equipment contained in them. In fact, one of the advantages of the new sonde is that the temperature and pressure calibrations are not affected by damage other than to the measuring resistors, since they are governed solely by the electrical properties of the sonde. This was verified by re-calibration of the sondes recovered.

Another point worth mentioning in connection with the calibration of the new sonde is that this can be accomplished without exposing the instrument as a whole to the effects of cold at the low calibrating temperatures. It is sufficient to plot the resistance v. temperature curves of the resistors, and the frequency curves of the oscillators as a function of the particular values of the variable measuring elements.

Summary. The instrument described as a "radio sonde" is used as a means of observing physical conditions in the atmosphere; in it, elements for measuring temperature, pressure and humidity are combined with a short-wave transmitter emitting a signal which contains the information gathered by the three elements. All this equipment is attached to a balloon, which is then sent aloft. Such a sonde must satisfy a number of stringent requirements, the most important of which are that it shall be accurate, light and inexpensive. The sonde described in this article differs from earlier types in that it transmits the three items of information simultaneously and so requires no switching mechanism. Each measuring element controls a small oscillator, the frequency of which is influenced by the particular variable to be measured. Pressure-variations are transmitted in the frequency band between 1 and 3 kc/s, temperature-variations in the band between 4 and 8 c/s, and variations in humidity in the band between 12 and 25 kc/s. To minimize the number of moving parts required in the sonde an N.T.C. resistor, included in the appropriate R-C oscillator circuit, is employed as a thermometer, and the pressure is ascertained from the boiling-point of freon, which is likewise measured by an N.T.C. resistor. The relative humidity is established with the aid of a piece of gold-beater's skin, variations in the length of which affect the self-inductance of a coil forming part of an *L-C* oscillator. The description of the R-C oscillators includes an account of how the frequencies of these oscillators are rendered independent of the battery voltage, which decreases to some extent during the ascent of the balloon. Sub-miniature valves are used in the oscillators; altogether, the sonde contains four of these valves. Frequency modulation is employed, the wavelength being 11 metres and the maximum frequency sweep 25 kc/s. The total output power of the transmitter valve (the battery-operated DL 41) is about 100 mW. Particulars of the mechanical design of the sonde and the principles of the ground-station receiving equipment are given in brief. Finally, the results of a number of practical tests on the new radio sonde are described.

SEASONING OF SODIUM LAMPS

The photograph shows a group of sodium lamps during the "seasoning" process — a stage in the manufacture during which the sodium is distributed in droplets along the discharge-tube, in order to ensure a uniform brightness in the finished lamp.

A PROFESSIONAL CINE PROJECTOR FOR 16 mm FILM

by J. J. KOTTE.

778.55.068.25

Originally 16 mm film was introduced to offer film-making facilities to the amateur at a reasonable cost. The advent of 8 mm film, however, was an even more important step in bringing amateur film making within the reach of the general public, and has now largely supplanted 16 mm film for amateur use. Nevertheless, 16 mm film survived, and acquired a range of applications peculiar to itself, including, for example, instructional and educational films. It is nowadays employed mainly by professionals, whose standards, as regards sound and picture quality as well as general reliability of equipment, are far more exacting than those of the amateur.

Most of the latest 16 mm projectors are simply more elaborate, and improved forms of the original equipment designed for amateurs. However, in the Philips projector, which has now been in production for some time, every link with the early amateur equipment has been severed; nevertheless, this projector is so simple to operate that it can be used even by persons not specially trained for the purpose.

For some time now, Philips have been producing in addition to standard 35 mm film projectors, a conveniently transportable 16 mm sub-standard film projector, type EL 5000, specially designed for use by professionals. *Fig. 1* is a photograph of this projector. The need for such a projector arises from the fact that 16 mm film has now largely passed from amateur, into professional hands [1]).

In view of this fact, special consideration has been given to such features as high luminous flux, to permit of projection on to a relatively large screen (up to 4 × 3 m), and robust construction, to enable the projector to be employed continuously for several hours per day, as well as to the overall quality of picture and sound. The total luminous flux falling upon the screen from the projector, using a 750 watt lamp and in the absence of film, is 500 lumens, which compares very favourably with the values [2]) usually rated as "acceptable" (200 lumens) and "very good" (275 lumens).

The means employed to secure this very high light flux will now be described, and in the course of this description reference will be made to the optical system, the shutter and the intermittent mechanism. Being in some respects novel in design, the last of these three items will be dis-discussed more fully than the others. Next, several details of the design, some of which are shared by Philips standard 35 mm projectors, will be considered, and, finally, a concise account of the sound system will be given.

[1]) Many aspects of 16 mm film, including industrial applications. are discussed in detail in "Sixteen mm Sound Motion Pictures" by W. H. Offenhauser, published by Interscience Publishers, New York, 1949.

[2]) See page 352 and 453 of the book by Offenhauser referred to above.

The luminous flux

The theoretical maximum of luminous flux Φ_{th} that can be projected on to the screen is governed entirely by the area (S), the luminance (B_s) of the film gate, and the angle α subtended by the effective diameter of the objective at the film gate (*fig. 2*).

This flux is given by [3]):

$$\Phi_{th} = \pi B_s \cdot S \sin^2 \tfrac{1}{2}\alpha.$$

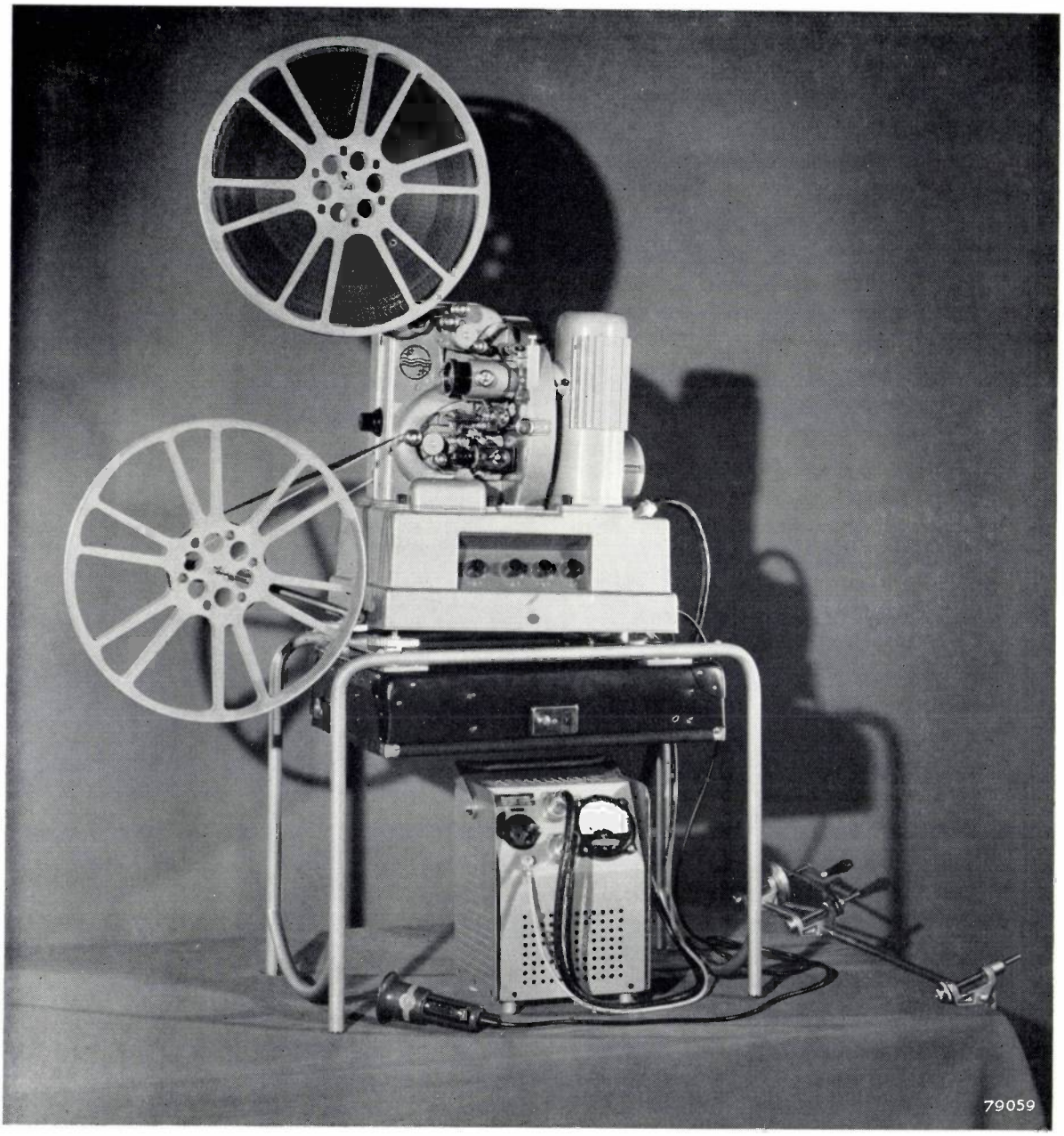
According to a well-known optical principle, the luminance along a beam of light proceeding through an optical system devoid of absorption and reflection losses is constant. However, since losses caused by absorption and reflection occur in every optical system, B_s is invariably less than the luminance B of the light source. Similar losses occur in the objective. The shutter gives rise to another unavoidable loss, which will be examined more closely later on. Taking all these losses into account by means of a factor k, we find that the total luminous flux emitted by a projector is:

$$\Phi = k \cdot \pi B S \sin^2 \tfrac{1}{2}\alpha \quad . \quad . \quad . \quad . \quad . \quad (1)$$

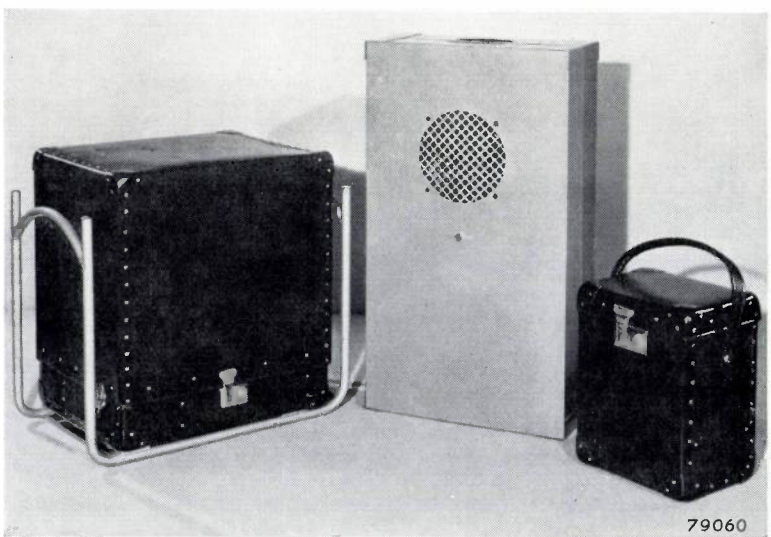
Since the area of the film gate is governed by the size of the film (being 0.75 cm² for 16 mm film), maximum light yield can be obtained only by increasing B, $\sin^2 \frac{1}{2}\alpha$ and k to the fullest extent.

Formula (1) is not strictly valid unless the particular lighting system (in this case comprising the lamp and a condenser) is so designed that all rays completely fill the objective, irrespective of the point at which they pass through the film gate (*fig. 3a*). In principle, any projector lighting-equipment can be so designed, but in practice, using a light source of

[3]) See G. Heller, A film projection installation with water-cooled mercury lamps, Philips tech. Rev. 4, 2-8, 1939.



a



b

Fig. 1. *a*) The 16 mm sound film projector EL 5000 ready for use. It is seen mounted on a standard support which brings the top of the upper film spool approximately to head level. The transformer for converting the mains voltage to 110 Volts will be seen beneath the projector. On the left of this transformer is a hand microphone, and at the extreme right a manual re-winder.
b) The equipment packed for transport in three cases, that on the left containing the projector, that in the centre the two loudspeakers, associated cables, spools and re-winder, and that on the right the transformer.

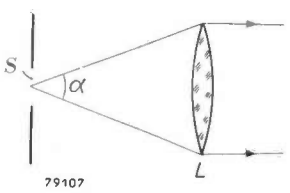


Fig. 2. The film gate (S) is uniformly illuminated by the lighting system (not shown). Since the projection screen is very remote from the projector the film gate is positioned for all practical purposes in the focal plane of the objective (L), here represented, for convenience, by a single lens.

moderate size and a reasonably inexpensive condenser, it is necessary to effect a compromise which satisfies the above condition only in respect of points in the central zone of the film gate. Rays passing through points outside this zone fail to fill the objective completely (fig. 3*b*) and so produce illumination which fades gradually towards the edges of the projection screen. Another reason for this fading is that rays arriving at a large angle to this axis are partly obscured by the edge of the objective (vignetting).

It is usual in cinematography to quote as the luminous flux the particular value corresponding to the illumination measured at the centre, assuming uniform illumination over the whole screen; this is in view of the fact that the measurement of the real luminous flux (which is always smaller) is a rather complex process.

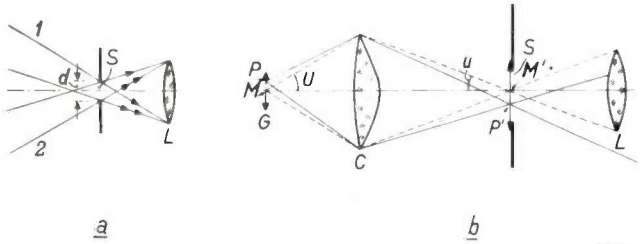


Fig. 3. *a*) To make full use of the objective it would be necessary to illuminate the film gate in such a way that even the light rays proceeding from the corners of this gate completely fill the objective. If d be the diagonal of the film gate, light rays *1* and *2* must still be present.
b) If the condenser C is employed to form an image of the light source A in the film gate, and the angle subtended at the centre M' (the image of M) of the film gate (S) by the condenser is the same as the angle subtended there by the objective (L), the beam passing through point M' will completely fill the objective. On the other hand, a beam passing through another point, P' (image of P), chosen at random in the film gate, will not fill the objective completely.

The light source

The most suitable light source for a transportable projector is the incandescent lamp, since this is relatively small and light in weight, and can be connected to the mains without any complex ballast. For various practical reasons, it was decided to employ a 750 Watt, 110 Volt lamp, the rated life of which is 25 hours. (The filament temperature, and hence the luminance level, may be increased, provided that a shorter lamp-life is considered acceptable). Quantity B in formula (1) is the average luminance of the area covered by the spiral filament). To procure the highest possible luminance B, a bi-plane filament is employed [4]) as seen in *fig. 4*, with a concave mirror mounted behind it. This mirror is actually formed on the inside of the lamp envelope, part of which has the form of a sphere. The cylindrical shape of the usual incandescent lamps employed in sub-standard film projection (fig. 4) necessitates the use of an external mirror, which, to protect it from damage, is usually silvered at the back. Hence the light reflected from such a mirror to the screen must pass six more air-glass interfaces than it would if similarly reflected from an internal mirror. Since a reflection loss of about 4% occurs at each boundary plane, the internal mirror is obviously more efficient. Other advantages of the internal mirror are that it is factory-adjusted, does not tarnish readily, and is renewed automatically whenever a lamp is replaced. Another, incidental advantage of the spherical lamp shape is that, indirectly, it improves the cooling; this may be explained as follows. The lamp is enclosed in a vertical housing (see, for example, fig. 1), through which cooling air is blown. The spherical part of the bulb narrows the air-gap, and so increases the velocity of the air, precisely opposite the source of heat (the incandescent element).

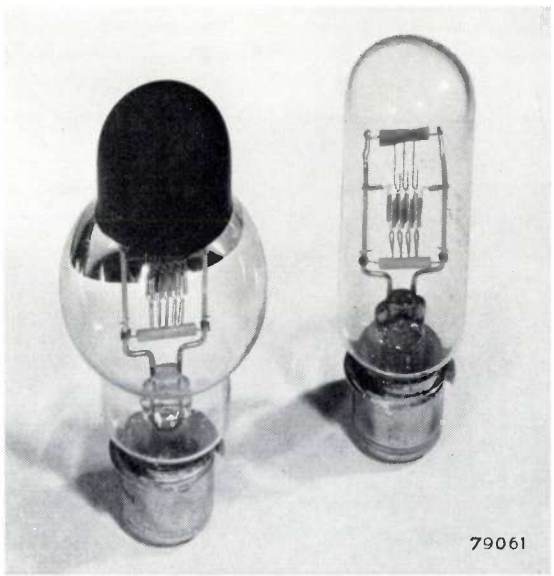


Fig. 4. The lamp employed in the 16 mm projector EL 5000 (left) and a conventional lamp for a 16 mm projector (right). The area covered by the filament of the left-hand lamp is rectangular, (same shape as the film gate), and the bulb of this lamp has a spherical form, part of which is internally silvered. Most of the material vaporised from the filament is deposited in the cylindrical top of this bulb, which is blackened to prevent light from escaping upwards.

[4]) For particulars of the design and properties of incandescent lamps for film projection, see Th. J. J. A. Manders, Philips tech. Rev. **8**, 72-81, 1946.

One feature of the incandescent lamp employed in this projector is that the filament is rectangular (2 × 5 spirals), instead of square (2 × 4 spirals) as in the lamps usually fitted in projectors (see fig 4). To explain why the new shape is preferred, it is assumed for the moment that the condenser forms an image of the filament in the film gate (fig. 3*b*). (In reality, the image is formed at a point beyond the film gate, so as to prevent the structure of the filament from appearing on the screen.) Some of the light emitted by a square filament invariably falls wide of the film gate; therefore by making the contour of the filament indentical with that of the film gate, without increasing the total filament area, the amount of magnification involved in the formation of the image in the gate is reduced, so that angle u is increased. This enables the marginal rays to fill the objective more completely.

The objective

Objectives of relative aperture (diameter of lens divided by focal length) 1:1.6 are employed in almost all modern 16 mm projectors. The relative aperture of the objective in the EL 5000 projector however, is 1:1.3. The relative aperture $1/m$ and the angular aperture a already referred to are related by the equation [5]):

$$1/m = 2 \sin \tfrac{1}{2}a$$

According to equation (1), then, increasing $1/m$ from 1:1.6 to 1:1.3 is equivalent, under otherwise identical conditions, to increasing the luminous flux by a factor of $(1.6/1.3)^2 \approx 1.5$.

As in most modern equipment of this kind, all the refracting surfaces of the objective are coated (bloomed) with a substance which reduces reflection [6]).

Shutter losses

One of the principal sources of light loss is the shutter. This is a rotating disc with two sectors, namely the frame sector and the flicker sector. The frame sector intercepts the light during the period in which the film is moved forward to bring the next frame, or picture, into the film gate. Unfortunately, the remainder of the total period available per frame cannot be employed exclusively for projection, since during it the beam must be cut off once more to prevent flicker (doubling the frequency). This second interception is effected by the flicker sector [7]). The duration of the two interceptions must be the same, or, in other words, the frame and flicker sectors must be exactly equal in width, since the slightest disparity between them will give rise to a perceptible and annoying flicker in the picture. This is due to the fact that if the two sectors are not properly matched, the fundamental frequency of exposure will still equal the frame frequency, instead of being twice the latter, as is the case when the sectors are identical.

Accordingly, if the frame-shift period is equal to p % of the total period per frame, a light loss of $2p$ % occurs; hence the importance of keeping the moving period as short as possible.

The loss occurring in the shutter is all the greater owing to the so-called "covering" angle (*fig. 5*). In theory, the beam should be completely cut off at the exact moment when the film starts to move, and should not be re-exposed until the film is stationary. Accordingly, the frame sector and also the flicker sector should be an angle β (covering angle) wider than the angle γ corresponding to the frame-shift period. Fortunately, however, practical experience has shown that the true situation is more favourable than the above argument suggests. Since the speed of the film is low at the start and end of each moving period. it is possible to "steal" part of the theoretical shutter width. To do so we determine by experiment exactly how narrow the shutter sectors can be made without permitting any part of the film movement to show on the screen. If the particular intermittent device employed is a Maltese cross mechanism with a four-slot cross (see below), as it almost invariably is in the case of 35 mm film, the moving period is governed entirely by this mechanism and gives $p = 25$ %. It is then possible to "steal" almost as much as is lost owing to the covering angle; hence the total shutter loss is about 50 %.

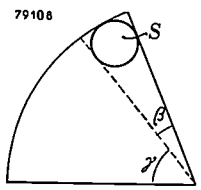


Fig. 5. The covering angle associated with the shutter sectors. The film must not start to move until the light beam (section S) is completely cut off. Hence the sector must be an angle β wider than the angle γ corresponding to the frame-shift period.

As we shall see later on, one attractive feature of the intermittent mechanism employed in the present projector is that it gives a free choice of the frame shift period. This is limited only by the amount of acceleration which the film and the mechanism will stand. We adopted $p = 16\frac{1}{2}$ %, which corresponds to a theoretical value of the frame angle (γ) of 60°. However, the covering angle β in our projector is so small as to enable us to "steal" more from the width of the shutter sectors than is added by this angle. Hence each sector is required to span only 50°, and the shutter loss is thus limited to 28%.

[5]) It should be borne in mind that the well-known paraxial formulae are not applicable, owing to the fact that the angles made by the light rays with the optical axis are not small.

[6]) The objective is manufactured by the N.V. Optische Industrie "De Oude Delft", of Delft, Holland.

[7]) See J. Haantjes and F. W. de Vrijer, Flicker in television pictures, Philips tech. Rev. **13**, 55-60, 1951/52.

The covering angle is minimized firstly by positioning the shutter where the beam diameter is smallest that is, immediately behind the film gate, and secondly by making the distance between the optical axis and the shutter axis relatively long. The shutter is therefore large, and has a high moment of inertia; hence it also performs the function of a flywheel which is required to ensure that the central driving spindle rotates smoothly. Moreover, blades fitted to this large shutter convert it into a fan (*fig. 6*),

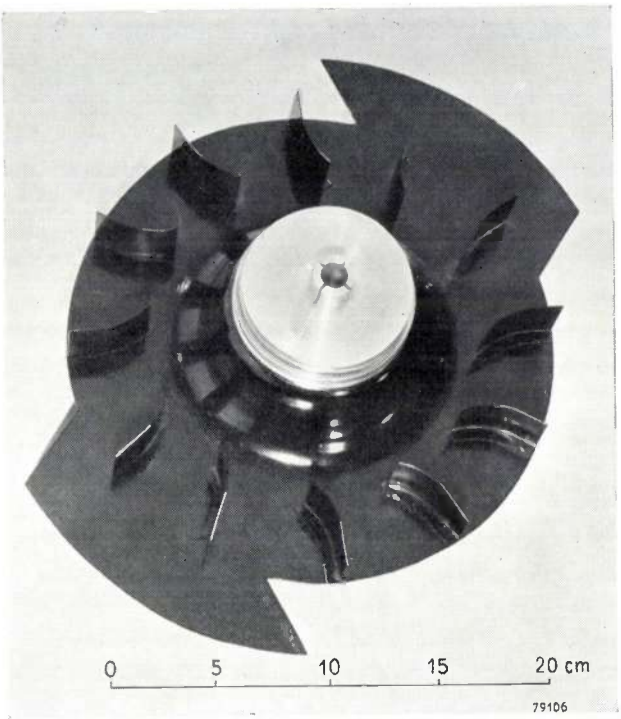


Fig. 6. The two-sector shutter. The frame sector and the flicker sector each cover 50°. The sectors are so shaped that the interception of the beam takes place parallel to the (long) side of the film gate, and therefore at a slightly higher speed than would be possible if sectors with truly radial edges were employed. Note also the large overall diameter, which enables the shutter to be used as a flywheel, and also the vanes which enable it to be used as a fan.

which despite a relatively low speed of rotation (1440 r.p.m.), provides sufficient cooling air for the projector lamp, the film mask (border of the film gate) and the film itself. This method of cooling is superior to the small fan driven at 5000-6000 r.p.m by a separate motor, which is used in many projectors, in that, apart from the eliminiation of the extra motor, it is virtually noiseless (no tendency to "whine").

The intermittent mechanism

Principle

The intermittent mechanism, which enables us to procure the very short moving period already referred to, differs in many respects from the types usually employed in film projectors (a Maltese cross mechanism or a cam-driven claw mechanism). The principles of these conventional mechanisms are explained in *fig. 7*. The principle adopted for the projector that we are now considering may be understood from the drawing and photographs of *fig. 8*.

The intermittent sprocket, that is, the sprocket which feeds the film forward, is mounted on one end of a spindle. At the other end of this spindle is a disc, which has projecting pins spaced evenly all round its periphery. These pins fit into the grooves of a cam mounted on the main spindle, and rotating with the spindle at a constant speed, which, in terms of revolutions per second, is equivalent to the number of frames per second. It will be seen from the shape of the grooves, shown in fig. 8, that the frame-shift period corresponds to the angle δ through which the grooves move in advancing the pins by one position. With each individual pin-movement, the intermittent sprocket turns sufficiently to feed the film forward exactly one frame-length.

The associated accelerations are governed by by the shape of the grooves over that part of the cam corresponding to the angle δ. A suitable choice of this shape has enabled us to procure an angle δ of only 60° without imposing undue strain either on the film or on the mechanism, and thus to limit the light loss in the shutter, as explained in the preceding section, to only 28%. The acceleration of the film is then such as to preclude any further narrowing of angle δ (that is, any further shortening of the frame shift period).

Another valuable feature of this mechanism is that it gives freedom of choice as to the number of teeth on the intermittent sprocket. If the number of pins protruding from the disc be n, the angle through which the intermittent sprocket rotates with each full revolution of the cam is $360\,°/n$. Taken at the periphery of the intermittent sprocket, this angle must correspond to the frame pitch (length of film per frame); hence the circumference of the sprocket should be n times the frame pitch. Similarly, since 16 mm film contains only one perforation per frame, the sprocket should have n teeth. In the case here considered, $n = 12$. Accordingly, we have on the sprocket 12 teeth, which engage with the film perforations about six at a time; hence the teeth, as well as the sprocket holes in the film are subject to only a small amount of wear. Moreover, experience has shown that films whose perforations are so damaged that a claw mechanism will not properly engage with them, can be shown on this

projector without any trouble at all, by virtue of the mechanism described here.

Defects in spacing

Mechanisms involving grooved cams operating on the above principle are widely used in automatic machines (e.g. bottle-filling machines), and have been attempted in cinema projectors. The novelty of the present intermittent mechanism lies in its structural design, which turns entirely upon the fact that the mechanism must be exceptionally accurate in every detail if it is to produce a stable picture on the screen. Unlike the claw mechanism, which returns to its starting point after feeding forward only one frame, the grooved cam mechanism does so only after advancing the film a distance of twelve frames. Inaccuracies in the distribution of the pins on the disc and of the teeth on the intermittent sprocket and also eccentricity of this sprocket are to some extent unavoidable. Though slight, these defects tend to affect the register of the intermediate frames so these do not all appear in exactly the same position; hence a kind of regular quivering, or dancing, of the picture results, which is all the more noticeable precisely because of its regularity. When once this effect is noticed, it continues to engage the attention of the observer and is therefore most irritating. To avoid it, the above-mentioned inaccuracies must be be kept to a minimum. For example, experience has shown that the Maltese cross mechanism employed in conjunction with 35 mm film, produces the familiar dancing pictures at a frequency of 6 per second (24 frames per second and a four-pole Maltese cross) if the wobble of the intermittent sprocket exceeds 1/100 mm. For 16 mm film, which requires a greater magnification, the wobble tolerance is smaller by a factor of 2.5, viz. about 4 μ. This fine tolerance can be attained in the manufacture of the intermittent sprocket. The teeth of the sprocket can also be cut to a sufficiently accurate *spacing*. Conventional methods, however, proved unequal to the task of accurately spacing the *pins* on the disc S_2 (fig. 8) of the new intermittent mechanism.

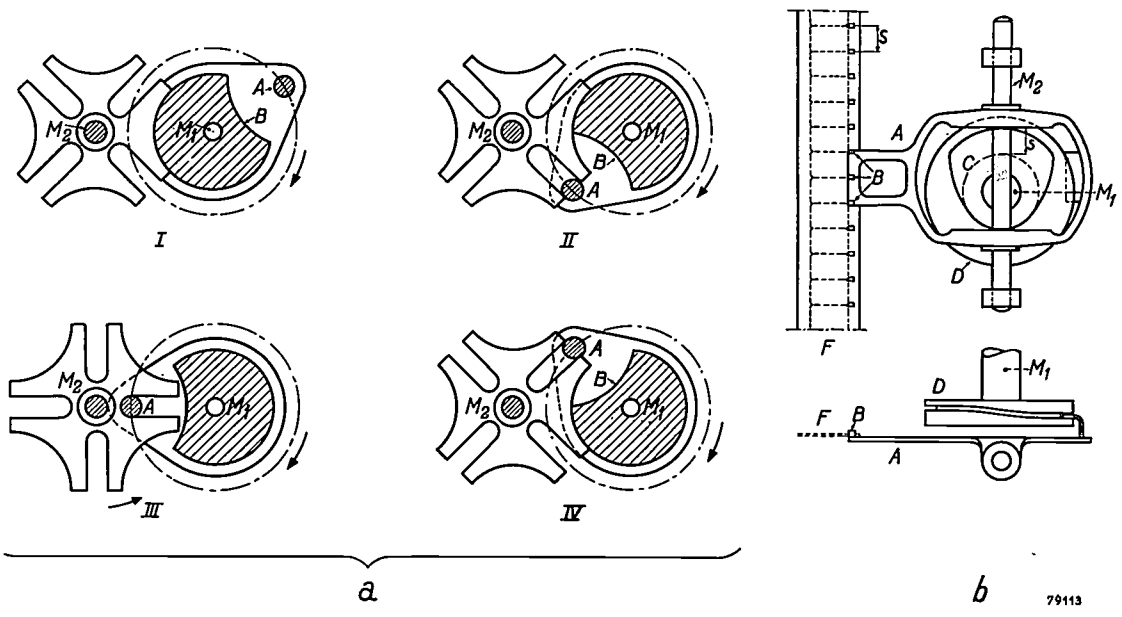


Fig. 7. Principle of the Maltese cross movement and the claw mechanism for intermittent film movement.
a) The Maltese cross movement in four consecutive positions. The end of spindle M_1, which rotates at a uniform rate, carries a cam with pin A and stop disc B. On spindle M_2 are the Maltese cross (here containing four slots), and the intermittent sprocket (not shown) which engages with the film perforations. In position I the Maltese cross is stationary; in position II the striker pin is just entering the cross, and the latter is starting to move; in position III the cross has been rotated through 45° and is now moving at its maximum velocity; in position IV the cross has completed a 90° rotation and is again stationary.
b) The claw mechanism. The arm (A) has a number of claws (B) (in this case three), which engage with the perforations of the film (F). A rotating cam (C) moves the arm up and down, the stroke of this movement being equal to the frame pitch. Cam C and the grooved cam (D) are mounted on the same spindle (M_1). The grooved cam turns the arm to and fro through a narrow angle about spindle M_2, so that the claws engage the film during the downward stroke, but miss it during the upward stroke.

A method was then adopted which enables correct pin-spacing to be effected automatically during the assembly of the disc, provided that each part is individually accurate. The pin-disc assembly seen in fig. 8 comprises a flat disc surrounding which 12 long and 12 short pins are arranged alternately, in intimate contact with each other and the outside of the disc. The pins are hollow, to reduce the mass to make the width of the pin grooves about 10 μ *less* than the diameter of the pins and so preclude all possibility of play between the two. Continuous lubrication is ensured by housing the cam and the pin disc in an oil bath (fig. 8).

The shutter is mounted direct on the main spindle, which in this machine is at right-angles to the intermittent sprocket spindle. The main spindle

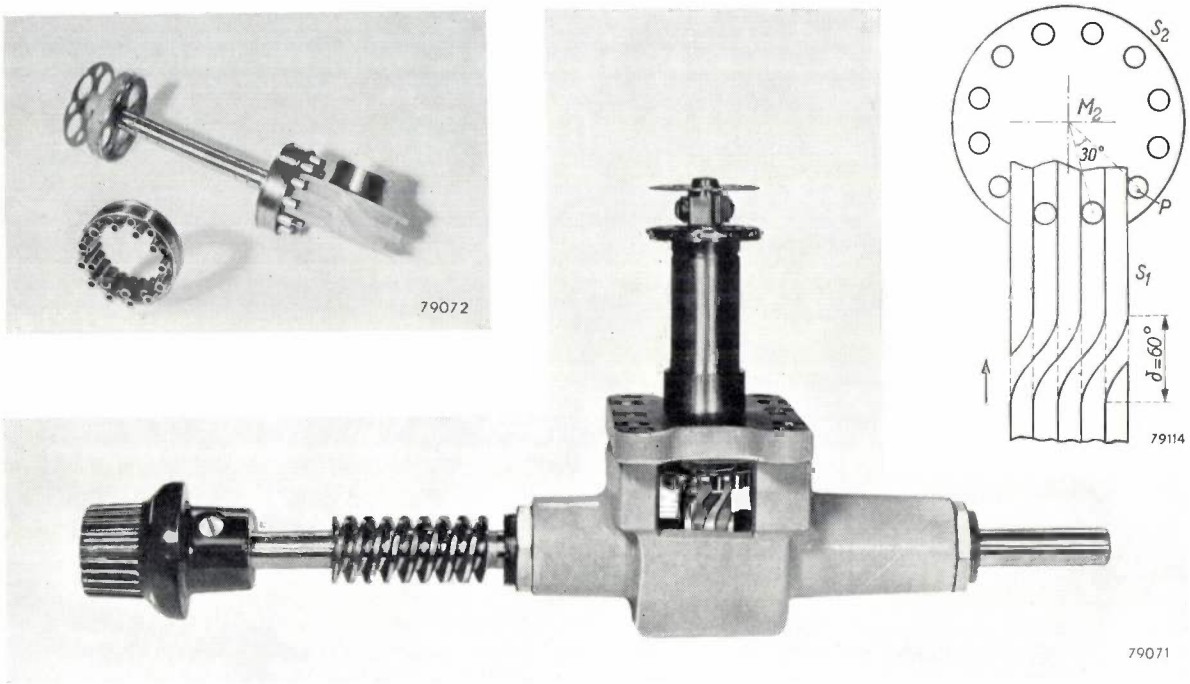


Fig. 8. The intermittent mechanism of the EL 5000. Top left: The most essential parts, viz. a spindle, with at one end the intermittent sprocket, and at the other a disc with projecting pins on its periphery. Meshing with these pins is the grooved actuating cam. The periphery of the disc with the pins positioned around it is included in the photograph to show how correct spacing of the projecting pins is achieved with the aid of shorter pins which do not project beyond the edge of the ring.
Top right: Diagram showing the action of the mechanism and the shape of the grooves of the cam (S_1). On the sprocket spindle (M_2) is the disc (S_2) with twelve projecting pins (P), which advance 1 position per revolution of S_1.
Below: Complete assembly. The pin disc, together with the grooved cam mounted on the main spindle, is completely enclosed in a housing (here cut open to show the interior) which is partly filled with oil. On the right-hand end of the main spindle will be mounted the shutter and a double belt pulley (fig. 6), which enables the mechanism to be driven at two different speeds. The worm on the main spindle drives the take-off sprocket and the take-up sprocket (see fig. 9 and fig. 12*a*). On the left is the inching knob used to move the film along by hand.

of the assembly and are held in position by a clamping ring. The merit of this arrangement is that it enables very slight (unavoidable) variations in respect of the correct pin diameter to be compensated by deformation distributed uniformly amongst the pins. At the same time, the free, projecting ends of the long pins are not perceptibly affected by this deformation. Measurements showed that the pin-spacing is thus maintained accurate to within 1-2 μ.

Another feature of this mechanism is that the grooved cam is made of nylon, which is very durable and resilient; so much so, in fact, that it is possible carries a worm, the under-side of which engages direct with the bottom, or take-*up*, sprocket. The top part of this worm drives a large intermediate wheel, which actuates the top, or take-*off*, sprocket (*fig. 9*). In fact, the top and bottom sprockets both rotate at a constant rate, the one feeding, and the other ejecting the film (for a general view of the film lacing path, see fig. 12*a*). A double belt pulley enables the intermittent mechanism to be operated at two different speeds (18 frames per second for silent film and 24 frames per second for sound film).

Concluding this general description, we may

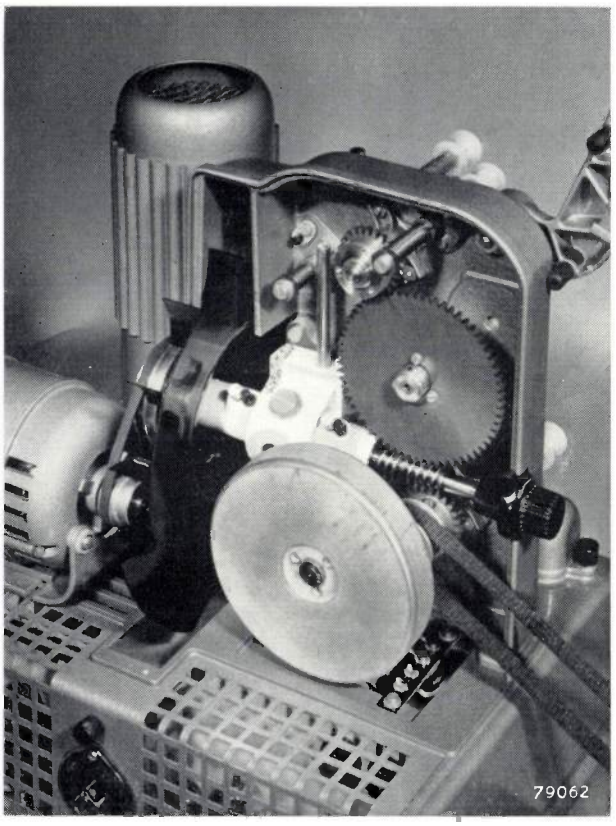


Fig. 9. Drive of the projector EL 5000. The main spindle, driven by the motor, passes through the housing of the intermittent mechanism, the top part of the worm on this spindle actuates a large intermediate wheel to drive the take-off sprocket, and the bottom part of this worm engages direct with the take-up sprocket (see fig. 12 *a*). The flywheel on the sound drum spindle can be seen in the foreground.

summarise as follows those features of the intermittent mechanism which are most relevant to its task: these are, a free choice of the frame shift period (which can hence be made very short) and of the exact way in which the film is accelerated, as well as of the number of teeth on the intermittent-sprocket; robustness and simplicity of design combined with accuracy, and, last but not least, efficient lubrication.

Comparison with the claw mechanism and the Maltese cross mechanism

To complete the description so far given it is necessary to add some remarks on the subject of the claw mechanism and the Maltese cross movement (fig. 7).

As regards the former a few words will be sufficient. Although in principle the claw mechanism is very accurate (see "Defects in spacing"), so much so, in fact, that almost all modern cinematograph cameras are equipped with it, none of the existing forms of this mechanism are suitable for use in a professional projector, for the following reasons. Firstly, these mechanisms invariably operate with very few claws, usually, in fact, only three, which between them must absorb all the wear arising from contact with the film. From this point of view alone, then, the life of our mechanism (12 teeth) will be four times as long. Secondly, the claws engage with the film perforations only three at a time, as compared with six at a time in our mechanism; hence the claws soon damage the perforations. Moreover, they fail to impart the correct movement to a film which has been so damaged. Thirdly, the claws must be light in order to perform the rapid reciprocating movement required of them, and this lightness is not compatible with robust construction. Finally, the lubrication of a claw mechanism presents a very difficult problem.

With regard to the Maltese cross mechanism, it can of course be adapted to 16 mm film, but the mechanism is then deprived of much of its attractive simplicity. To explain why the ordinary form of Maltese cross mechanism is never employed for 16 mm film, we shall now describe it more fully.

A diagram showing the principle of this movement (here supposed to be of the four-pole type) will be seen in fig. 7. One condition which every Maltese cross movement must fulfil in order to operate smoothly is that the striking pin shall be travelling in the direction of the centre line of the slots in the Maltese cross at the precise moment of entering or leaving one of these slots (which are stationary at such a moment). In the diagram shown in *fig. 10*, the striking pin is describing a circle about M_1; M_2 is the spindle of the Maltese cross, A_1 the position of the striking pin on entering, and A_2 the position of this pin on leaving a slot in the cross. It will be seen that angles α_1 and α_2 of the quadrilateral $M_1A_1M_2A_2$ are right angles; hence $\beta_1 + \beta_2 = 180°$. In the case of a cross with m slots, $\beta_2 = 360\,°/m$. The period in which the pin is engaged with the cross, i.e. the moving period, is proportional to β_1; to procure a short moving period, then, we must make β_1 small and consequently β_2 large. A three-slot Maltese cross furnishes the smallest angle β_1, viz. 60°, which then corresponds to a moving period as short as that of the new mechanism employed in the present projector. However, such a cross gives no freedom of choice as to the precise way in which the film is accelerated; in fact this is so violent that the forces involved soon cause damage to the film and to the mechanism. Accordingly, a Maltese cross should contain at least four slots, the moving period then being 25 % of the total period per frame, and the light loss in the shutter about 50 % of the total flux. In the case of the four-slot cross, the circumference of the intermittent sprocket, when mounted direct on the spindle of the cross, must be equivalent to four frames of

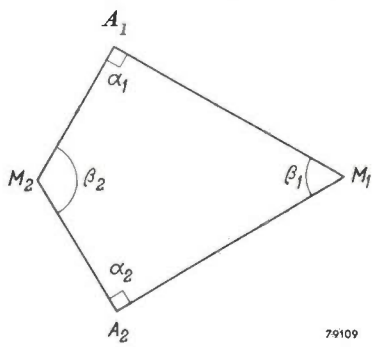


Fig. 10. Diagram showing method of determining the moving period of an m-pole Maltese cross mechanism. The striker pin rotates about M_1 at a constant rate. M_2 is the centre of rotation of the Maltese cross. Points A_1 and A_2 are the respective positions of the striker pin on entering and on leaving the cross; hence $\beta_2 = 360\,°/m$. Since the principle condition of smooth operation is that $\alpha_1 = \alpha_2 = 90°$, angle β_1 (and, therefore, the frame shift period) increases with m.

the film. For 16 mm film, then, the sprocket required would be so small that it would force the film into a very tight loop, which would tend to unstick any splices in the film. A still more serious disadvantage of such a sprocket is that it can accomodate only four teeth, which can engage with the film only two at a time; hence the film perforations would soon be damaged and the teeth soon worn.

Freedom of choice as to the number of teeth on the intermittent sprocket can be obtained by mounting the sprocket on a separate spindle geared-down to the spindle of the cross. However, apart from the extra spindle, this involves an extra gear, which, in view of the spacing defects already referred to, must be very accurately machined. The moving period can be shortened by varying the speed of rotation of the striking pin in pulses, so that it rotates quickly when actuating the cross, but slowly when clear of it. On the other hand, the practical application of this method involves one (or two) extra spindles. It will be seen, then, that a Maltese cross mechanism of approximately the same efficiency as our grooved disc mechanism would be very complex.

Details of the film threading

Automatic tensioning of the film

The bottom, or take-up, sprocket feeds the film to the take-up spool: this is driven, as usual, by a drive which slips when necessary to adapt the speed of rotation to the varying diameter of the film reel. It is desirable that the tension of the film between the sprocket and the take-up spool be maintained virtually constant in the region of 160 grams, since at this tension the film is wound just taut enough. The film tension, being equal to the frictional driving torque divided by the radius of the film reel, will be constant if one of these factors is made proportional to the other. To fulfil this condition a very simple attachment is employed. This is an arm secured to the projector frame by a hinge and carrying at its free end the spindle for the take-up spool (*fig. 11*). The free end of this arm is supported solely by an endless belt of braided cotton, which passes round a pulley at one end of the spool spindle and constitutes a slipping drive for the take-up spool. The friction between belt and pulley, and hence the driving torque, increases with the weight of the film reel. Now, by adopting a suitable pulley diameter and positioning the driving belt and the spool arm at suitable angles to the vertical, it is possible to make the driving torque proportional to the radius of the film reel, and their ratio equal to the desired film tension. In the projectors now in production, the film tension changes only from an initial 170 grams for an empty spool to a final 150 grams for a full spool (600 m of film). The variable friction clutch to be found in most projectors has been abolished in this system, and with it has gone the possibility of incorrect adjustment. This is an even greater advantage from the point of view of sub-standard film than from that of ordinary 35 mm stock, the former being by far the more fragile of the two owing to its relatively small size and less durable (though non-inflammable) quality Moreover, it is probable that projectors for sub-standard film will sometimes be operated by untrained, or inexperienced persons.

In the case of the feed spool, the problem of film tension was solved in a similar manner. The torque set up by frictional forces in the bearing of the spindle carrying this spool decreases with the weight, that is the radius, of the film reel, and so maintains the film tension virtually constant. A frictional torque consistent with the desired film tension is procured by employing a bearing of the appropriate diameter.

Threading the film

In view of the possibility already referred to, that the projector will be operated by persons not specially trained for the purpose, the process of threading the film has been made as simple as possible. The mere turning of a handle clears the film path of all obstructions (see *fig. 12*), and positions the pad rollers so that the film can be threaded in taut. When once the film is so threaded, the handle should be turned back to its original position, and the projector is then ready to operate. Loops of the exact length required to permit of the necessary intermittent movement through the film gate are formed automatically above the gate and beyond the intermittent sprocket. The length of the bottom loop (beyond the intermittent sprocket) is especially

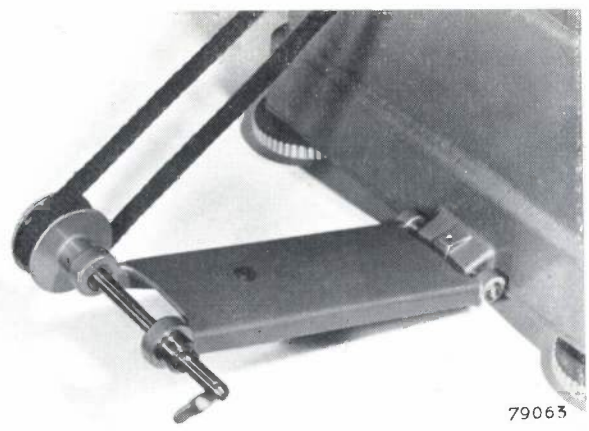
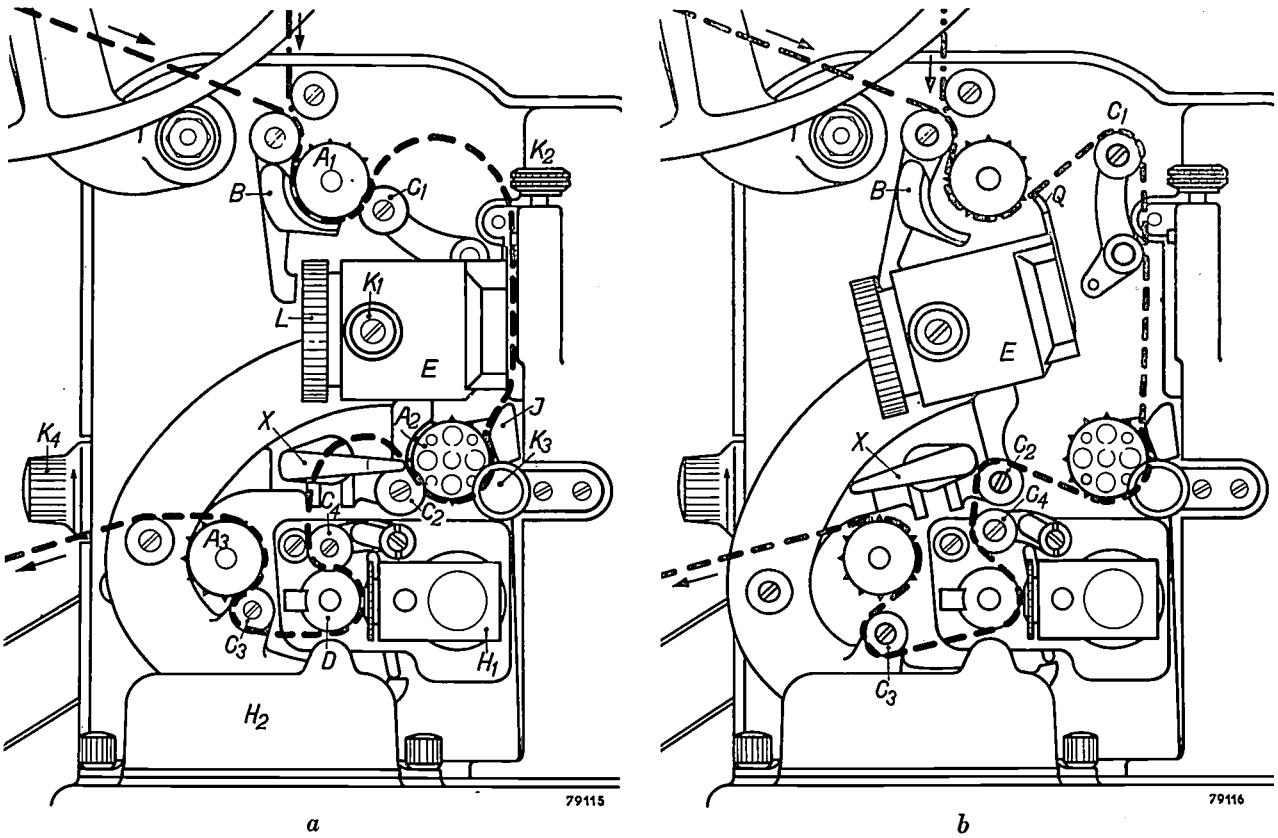


Fig. 11. Drive and mounting of the take-up spool. One end of the spool arm is hinged to the projector, and the other end, carrying the spool spindle, is supported solely by the belt of braided cotton which drives the spindle. This belt slips whenever necessary to adapt the speed of rotation of the spool to the diameter of the film reel: the film tension remains almost constant at approximately 160 grams.

critical, since it governs the distance between the particular frame exposed in the film gate at a given moment and the part of the sound track that will be scanned at the same moment. In the case of 16 mm film, the distance between the two should always be 26 frames, this being the standard space between a frame and the associated sound. Proper length adjustment of the bottom loop is all the more important in the case of 16 mm film because here the number of frames per unit length is a factor of 2.5 more than in 35 mm film. Hence an error of one or two cm in the length of this loop causes far more asynchronism in a 16 mm, than in a 35 mm film.

The film gate

The film is held stationary in the film gate between two straight guides (the runner plates), one in front and one behind. One of them, which is spring-loaded, may be described as the spring guide, and the other as the fixed guide (fig. 12 *b*). Facilities for adjustment of the spring pressure by the operator are provided for the following reason. Films, especially new ones, deposit dirt on the film guide. This dirt usually accumulates rather rapidly and may so increase the spring pressure that the latter causes the film to break. By means of the adjustment provided, the operator can reduce the pressure



a *b*

Fig. 12. *a*) General view of the film lacing path. The take-off sprocket A_1 draws the film from the feed spool (at this point the film may follow either the broken line or the chain-dotted line, depending on how it has been wound). Guide shoe B and pad roller C_1 prevent the teeth from slipping out of the film perforations. The intermittent sprocket A_2 draws the film intermittently through the film gate, between the fixed guide and the spring guide. (The fixed guide Q will be seen in (*b*)). Next, the film travels over the spring-loaded pressure roller C_4 and round the sound drum D, and thence to the take-up sprocket A_3 with pad roller C_3, which feed it to the take-up spool. E is the objective mount, L the objective itself and K_1 the objective locking screw. K_2 is the knob for varying the spring pressure on the film in the gate, and J the framing shoe with adjusting knob K_3. H_1 is a cover to protect the exciter lamp and optical system of the sound-head, and H_2 similarly protects the photo-electric cell. K_4 is the inching knob and X the handle to clear the projector for film threading.
b) The projector cleared for film threading, with film threaded in. The handle X has been turned, so that the objective holder E, together with the fixed guide Q attached to it are tilted back. At the same time, the shoe B near the take-off sprocket and pressure roller C_4 on the sound drum, are lifted and so positioned that the film can be threaded *taut* along the proper path. When handle X is turned back to its original position, loops of the correct lengths will be formed above the film guides and beyond the intermittent sprocket (see (*a*)). Sound and picture are thus synchronized automatically.

whilst the film is running, and so need not interrupt a programme in order to remove the dirt.

The fixed guide, unlike those in ordinary 35 mm projectors, is attached to the objective holder: this ensures that the film will remain in focus even if handle X is not turned quite as far as it should be after the insertion of the film, and the objective is therefore slightly out of alignment (fig. 12*b*).

The framing adjustment

The system must be flexible enough to enable individual frames to be accurately registered in the film gate as and when required (framing). Framing is necessary to compensate for film shrinkage (new films are invariably longer than old ones) and for any frames positioned either too high or too low in relation to the sprocket holes, as may happen, for example, in the process of copying an original film. Framing adjustment can be obtained by making the length of the film path between the gate and the intermittent sprocket variable. Variability of the film path is procured by taking the film from film gate to intermittent sprocket over an adjustable shoe (J in fig. 12*a*); moving this shoe to the left lengthens the film path. One advantage of this system is that it keeps the picture centred on the screen during framing, unlike, for example, the widely used method of making the film gate itself, or more precisely the film mask mounted in this gate, adjustable. The last-mentioned method would here necessitate vertical re-adjustment of the projector after framing, which, for professional equipment, is inadmissible. We were able to adopt the more convenient solution by virtue of the fact that the framing travel required for 16 mm film is only about 1 mm. Far more range of adjustment is required for 35 mm film. In this case, the pitch of the sprocket holes is 1/4 of the frame pitch, therefore a badly positioned splice may cause the film to jump $\frac{1}{4}$, $\frac{1}{2}$, or even $\frac{3}{4}$ of a frame; hence the 35 mm framing system must be capable of correcting such variations. However, similar picture displacements do not occur in the case of 16 mm film, where the sprocket hole pitch is equal to the frame pitch.

The film path in the sound head

How to maintain the film speed constant within the sound scanning system has remained a problem ever since the advent of sound film. Vibrations are produced in the film owing to the discontinuous movement of the intermittent sprocket and also to the fact that the teeth of the take-up sprocket do not draw the film perfectly smoothly through the sound-head. These vibrations must be prevented from reaching the point at which the sound track is scanned. A conventional solution to this problem, viz. a rotating sound drum, is used. This is a drum driven by the film and coupled to a flywheel and therefore rotating at a very uniform rate. The sound track is scanned at a point on the drum [8]). The loops described by the film in passing round the different rollers act as a spring, and this spring, in conjunction with the combined mass of the flywheel and the sound drum, forms a mechanical filter through which the vibrations cannot pass. The lighter the spring and the greater the moment of inertia of the flywheel, the more effective is this filter.

A light spring is obtained when the loops in the film are loose, that is, when the tension of the film is low. Now, low film tension between the sound drum and the take-up sprocket can be procured by running the spindles of the drum and the associated pressure roller in ball-bearings lubricated with thin oil instead of grease, which permits of very free rotation.

The film is often threaded in an S-bend, to make it fit closely round the sound drum (*fig. 13a*).

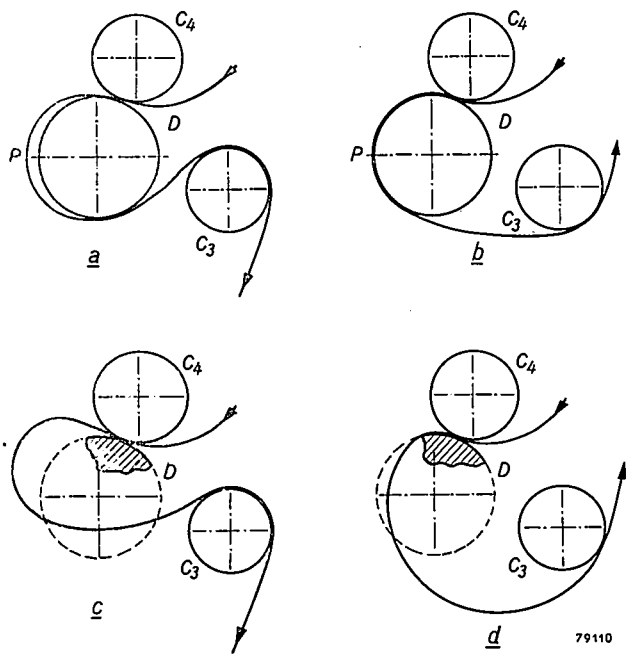


Fig. 13. The film lacing path in the sound-head. D is the sound drum, C_4 the pressure roller, C_3 the guide roller and P the sound scanning point.
a) Threading the film round the sound roller and the guide roller in an S-bend, in conjunction with low film tension, tends to cause the film to stand away from the sound drum precisely at the scanning point P ("breathing").
b) With the film threaded in a U-bend, it will remain in close contact with the sound drum at point P despite low film tension.
c) The loop which the stationary film would describe in case (*a*), if the whole of the sound drum other than the portion hatched in the diagram were removed.
d) The same, in respect of case (*b*).

[8]) See J. J. C Hardenberg, The transport of sound film in apparatus for recording and reproduction, Philips tech. Rev. 5, 74-81, 1940.

Logical though this may seem, however, it is found that in fact, owing to the above-mentioned low film tension, the film then becomes detached from the drum precisely at the scanning point, and goes into almost imperceptible radial oscillation ("breathing"); its consequent failure to remain continuously in proper focus relative to the optical system gives rise to perceptible sound distortion. On the other hand, the film remains in close contact with the drum if taken round the other side of the guide roller to form a *U*-bend instead of an *S*-bend (fig. 13*b*). This apparently strange behaviour of the film is nevertheless to some extent understandable bearing in mind what shape the film would assume if the the whole of the sound drum other than the small portion actually in contact with the pressure roller were removed (fig. 13*c* and *d*).

The sound installation

Details of the scanning system

The light source employed for sound scanning is an illuminated slit, a sharp image of which can be focussed on the sound track of the film by turning the large, knurled ring shown in *fig. 14*. Moreover, the slit can also be adjusted so that it is at right-angles to the sound track (azimuth adjustment). The two adjustments are quite independent of each other.

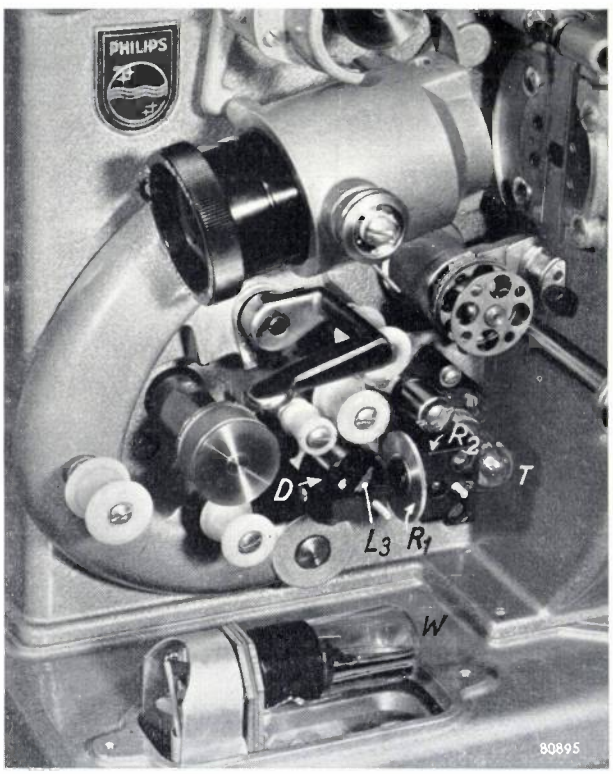


Fig. 14. The sound-head of projector EL 5000. *T* is the exciter lamp. Knurled ring R_1 is used to focus the slit-image on the sound track, and R_2 to adjust this image at right-angles to the track (azimuth adjustment). The plano-convex lens L_3, which projects the light down into the photo-electric cell *W*, is just visible in the photograph, inside the sound drum *D*, This photograph shows quite clearly that the guide rollers are not made of steel; they are made of nylon.

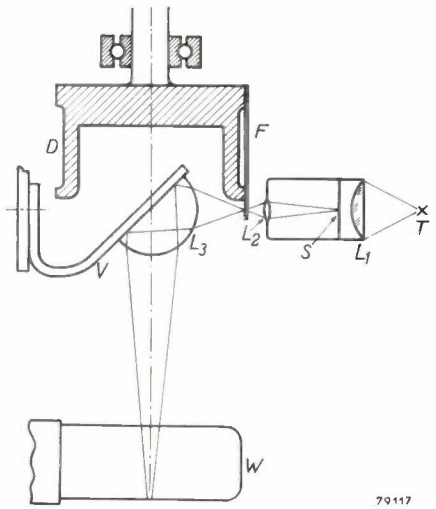


Fig. 15. Diagram showing the path of light-rays during scanning. The sound track of the film *F* projects beyond the edge of the sound drum *D*. *T* is the exciter lamp, *S* the slit, L_1 the condenser lens, L_2 a cylindrical lens and L_3 the plano-convex lens, silvered on the plane side; *W* is the photo-electric cell and *V* a support for L_3. In reality, the sound drum is mounted with its axis at right angles to the position shown, i.e. perpendicular to the plane of the diagram.

The lamp is exited by an oscillator supplying H.F. current (90 kc/s), and therefore produces no audible note. Since the output of the oscillator must in any case be extracted through an output transformer, there was, with the available power of 10 watts, a free choice as regards the current and voltage for the exciter lamp. In the case here considered, a low voltage and a heavy current were adopted, viz. 2.5 volts and 3 amps., which differ considerably from those employed in most sound-heads. Accordingly, the filament is a short, tough spiral of thick wire, producing no microphony.

Another point worth mentioning is that the exciter lamp may be switched over to a 50 c/s supply if a fault develops in the oscillator. The resultant hum is relatively slight, owing to the high thermal capacity of the thick filament.

The light emitted from the slit is projected downwards by a plano-convex lens, silvered on the plane side, into a photoelectric cell (*fig. 15*) secured to the mounting plate of the amplifier. The removal of two srews, one on either side of the apparatus, enables the whole projector to be lifted from the amplifier, leaving the photo-electric cell still in

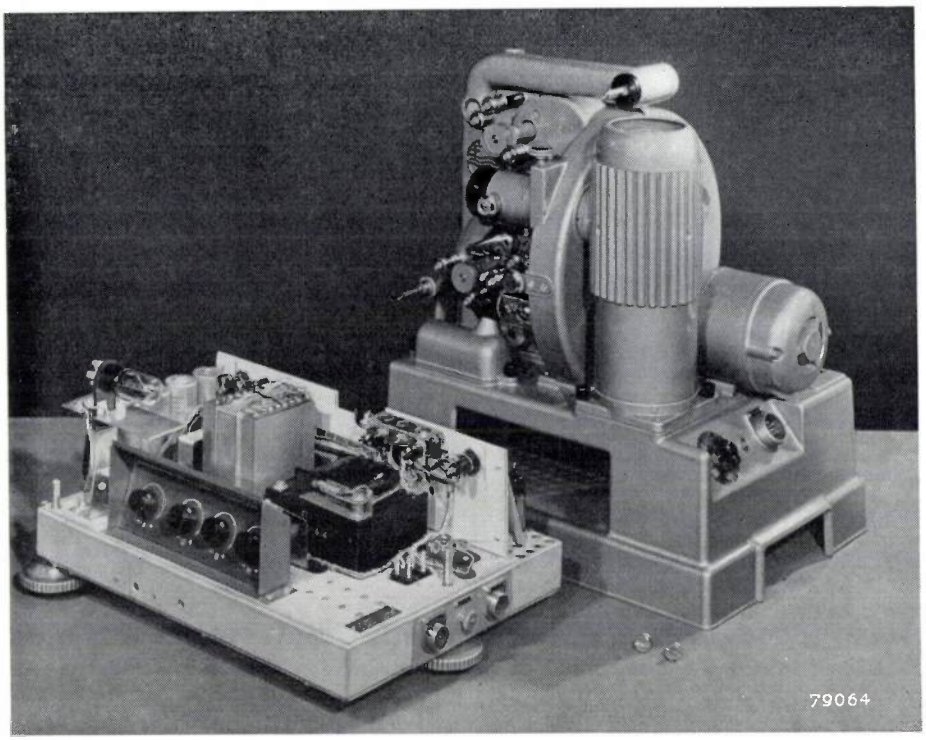


Fig. 16. The projector is completely detachable from the amplifier. Note the photo-electric cell at the top left-hand corner of the amplifier.

position (*fig. 16*). In this way, neither a photo-electric cell cable, or the associated plugs, which are apt to cause interference (crackle), are required.

The projector is so designed that, from the rear, the film is at the left-hand side and the sound track accordingly on the outer edge of the film; hence the sound scanning beam does not fall foul of the spindle of the sound drum (see fig. 15) and the plano-convex lens can be mounted inside this drum. Moreover, the lens is easily removed, for cleaning or inspecting the optical system.

The sprocket holes being along the inner edge of the film the latter can be inserted without being passed full-width across the sprocket teeth and so exposed to possible damage by these teeth.

Features of the amplifier

The sound amplifier gives an total output of 15 watts, with 3% distortion; this output is divided between two loudspeakers, together powerful enough for a hall seating 750 people. The high and low note response can be varied separately. A socket for a microphone (or gramophone) is provided in the amplifier, so that a commentary or explanatory talk can be given to accompany the pictures on the screen (see fig. 1). The volume of the sound proceeding from the film itself and of the spoken, or "live" commentary can be varied independently.

The design of the amplifier, like that of the remainder of the equipment described here, is very largely the outcome of a desire to make the projector as reliable as possible, in view of the fact that it is intended for professional use. In order to achieve this, the number of valves in the amplifier is quite considerable (i.e. 12, including the high-frequency supply unit for the sound track exciter lamp). This has enabled us so to arrange matters that in the event of failure of one or more of the components during the programme, the amplifier can still function, though with some deterioration in quality and (or) volume of the sound. In fact, the amplifier will operate with only six valves.

Only four different types of valve are employed; hence four valves in all constitute a complete reserve set. Each of them is a standard type of radio valve and therefore readily obtainable.

Further particulars

The projector is driven by a split phase motor, amply dimensioned and therefore unresponsive to variations either in voltage or in load. In this type of motor there is no commutator to cause sound distortion owing to sparking at the brushes. The motor is mounted on rubber to prevent the transmission of vibrations to the projector.

Suspended in the lamp housing is a vane, which cuts off the light passing to the film when the latter is at rest. When the projector is running, that is,

when the film is moving, however, the draught of cooling air set up by the shutter-fan lifts the vane, thus admitting the light to the film. This vane is provided, not so much against the actual risk of fire, but to prevent overheating and damage to the film, should the lamp remain switched on with the film stationary.

A separate re-winder is provided (see fig. 1), which dispenses with the inconvenience of taking the projector out of service in order to re-wind a film.

Practical experience has shown that the projector described here fully satisfies the requirements imposed on it during design. The quality of the sound and picture depends, of course, also on the quality of the available films and regrettably few sub-standard films at present on the market are of really high quality as regards both picture and sound. However, the present trend of development is towards a considerable improvement in this respect.

Summary. For some time past, the Philips factories at Eindhoven have been producing a 16 mm sub-standard film projector (type EL 5000) designed for professional use. Two important features of this projector are its high light output (500 lumens) and its robust construction. A high light yield, that is, sufficient for the projection of 4 × 3 metre pictures, is obtained firstly by employing an objective of large relative aperture (1:1.3), secondly by carefully matching the lighting system with this objective, and thirdly by minimizing shutter losses. The shutter losses are reduced by means of an unconventional intermittent mechanism, which ensures a very short frame shift period. Another advantage of this mechanism is that an intermittent sprocket having a large number of teeth can be employed. This mechanism is described in detail. Shutter losses are further reduced (to 28%) by means of a special, large-diameter shutter. The size of this shutter also enables it to be used as a fan and simultaneously as a flywheel. Other details of the design discussed in this article are: automatic control of the film tension at the feed and take-up spools; automatic looping of the film, and automatic synchronization of picture and sound during the threading of the film; framing adjustment without de-centring of the picture on the screen; very low film tension before and after the sound-head, whereby a constant film speed is maintained at the scanning point; and, lastly, the sound scanning and amplification system itself.

VISUAL ACUITY IN CONNECTION WITH TELEVISION

by G. J. FORTUIN.

621.843.6

In 1951, Philips' Medical Department tested 228 persons of ages varying from 7 to 64 years in order to collect data concerning visual acuity and its relation to contrast and brightness. In this article the results of this research are discussed in connection with a study of the observation of television images.

The information obtained when looking at a two-dimensional, flat object, e.g. a photograph or a television picture, is determined on the one hand by the sharpness and contrast of the image itself, and, on the other, by the properties of the observer's eye and the conditions of observation (brightness, distance between the eye and the object).

The property of the eye that is decisive in this respect, viz. the visual acuity, can be expressed quantitatively by the angle D_0 subtended at the eye by the smallest perceptible detail. This angle is frequently assumed to be approximately one minute of a degree, but it is highly misleading to attribute a fixed value to D_0. As will be seen, D_0 depends not only on the luminance, but also to a very considerable extent on the contrast. Furthermore, there may be individual differences, which may be accentuated by insufficient correction of refraction errors (wrong spectacles, or no spectacles at all). Moreover the *age* of the observer greatly affects the value of D_0.

In television the smallest detail in the vertical direction is determined by the scanning width: in the case of a 625-line, 30 × 40 cm picture this width is approximately 0.5 mm. The smallest detail in the horizontal direction is determined by the highest modulation frequency, and it can be assumed that it is of the same order of magnitude. If d be the size of the smallest detail, then the maximum distance at which it can be discerned, is $a = d/D_0$. With $d \approx 0.5$ mm and $D_0 = 1'$, it follows that $a = 1.8$ m. This, of course, is only an approximation, for a is actually a dependent on all the above-mentioned factors which affect the value of D_0.

It is evident that as a rule the observer wants to watch the screen from a distance at which it is just possible for him to discern the smallest details with smallest contrasts. As the contrast between the lines of the image and the spaces in between is greater than the smallest contrast the observer wants to see, the scanning lines themselves will also be clearly distinguished. Experience has shown that viewers do not object to this.

It is important in this connection to consider more closely the factors influencing the value of D_0.

In 1951 Philips' Medical Department carried out extensive research into the relationship between visual acuity, contrast and brightness in connection with the age of the observer [1]), and the results obtained in this work are applicable to the problem under discussion. The 228 persons tested (of ages varying from 7 to 64 years) comprised school children, applicants for jobs, and in the higher age categories, employees of the Philips works. To permit conclusions of practical value to be drawn from the tests, all these persons were subjected to the examination in their normal, every-day circumstances, i.e. no attempts were made to improve their eyesight, for example, by means of spectacles. The original object of the investigation was to examine the influence of age on vision during very delicate manufacturing operations (assembling of radio tubes) and to what extent this can be compensated by raising the level of illumination .The tests partly confirmed earlier results about the relationship between visual acuity, contrast and brightness [2]), which could now be combined into a single empirical formula, also containing a factor concerning the influence of the observers' age on his visual acuity.

Testing procedure

Since an object can be fairly easily recognized according to its shape (this is illustrated, e.g. by the varied shapes of printed letters), it is desirable in an investigation into visual acuity to use some kind of standardized object. In the present case a so-called Landolt ring was used for this purpose (*fig. 1*).

The rings are on a white panel (*fig. 2*) which is divided into 16 × 11 squares, in each of which a black or grey paper Landolt ring is pasted. There

[1]) G. J. Fortuin, Visual power and visibility, Diss. Groningen, 1951; see also Philips Res. Rep. 6, 251-287, 347-371, 1951.

[2]) See e.g. P. J. Bouma, Visual acuity and speed of vision in road lighting, Philips tech. Rev. 1, 215-219, 1936.

are eight different ring positions (opening at top, bottom, left, right, or at 45° in between). In each vertical column the outside diameter of the rings diminishes from top to bottom according to a geometrical progression from 50 to 2.2. In each horizontal row the contrast diminishes from right

Fig. 1. Landolt ring. The aperture and the width of the ring are equal to one fifth of the outside diameter.

to left according to an arithmetical progression from 0.94 to 0. The contrast is determined by the difference in luminance $L_a - L_r$ between background and ring divided by the background luminance L_a; alternatively if background and ring receive equal amounts of light, the contrast is determined by the difference in reflective power $\varrho_a - \varrho_r$ between background and ring divided by the reflective power ϱ_a of the background:

$$C = \frac{L_a - L_r}{L_a} = \frac{\varrho_a - \varrho_r}{\varrho_a}.$$

The visual acuity in this Landolt ring test is determined by the reciprocal value of the angle (expressed in minutes of arc) subtended by the aperture of the ring, in the case of the smallest ring whose position the observer is able to ascertain. If this angle be D_0, then $1/D_0$ represents the visual acuity. The test supervisor indicates, by means of an arrow-head (see fig. 2), one ring at a time. *Fig. 3* shows how the person under test responds by placing a rotatable ring into a position corresponding with that of the ring indicated on the panel. A correct answer is signalled electrically (by a white

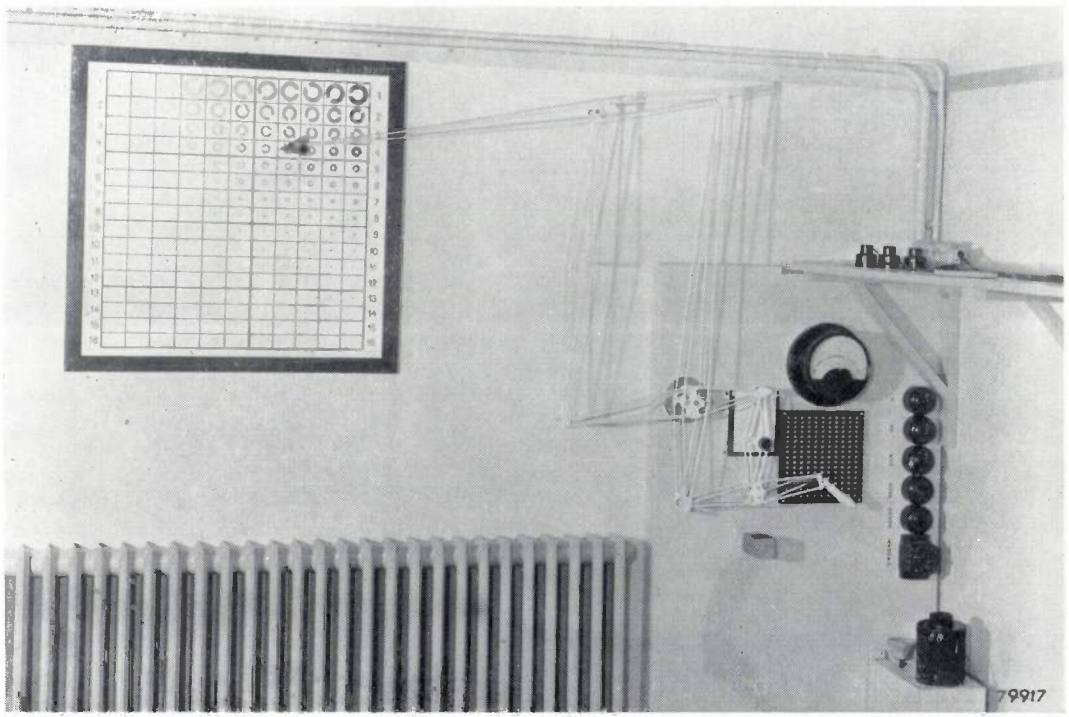


Fig. 2. Equipment for examining visual power. The panel on the left is divided into 11 × 16 squares, each of which contains a Landolt ring of known size and contrast value, but set in an arbitrary orientation. Each time the arrow points at one of the rings; the person under test, who stands at a distance of 5 m, has to try to distinguish its orientation. On the right, the semi-automatic recording device.

disc appearing in front of an aperture, visible only to the investigator).

Fig. 2 also shows how this signal is produced. The movement of the arrow-head is reproduced on a smaller scale by means of a parallelogram linkage to a pin moving across an insulating board with holes and to a needle moving across a sheet of paper

in a holder. Both the insulating board and the sheet of paper are small-scale reproductions of the ring panel. Each hole contains a metal bush which is connected to one of the eight contacts under the rotatable ring manipulated by the person under test. As soon as the pin is pressed into the hole, a circuit is completed provided the position of the rotatable ring corresponds with that of the paper ring indicated by the arrow-head.

When the answer is correct, the examiner presses the needle in, thus punching a hole in the paper and recording the result (*fig. 4*).

The final result is then ten figures [3]) for each test person at various luminance values L_0 (four in our case) of the field of vision, viz. the numbers of the last ring in the various columns for which a correct answer was given. The value of D_0 was taken as the geometric mean of the angles relating to this last ring and the next smaller one. For each value of D_0 there is an accompanying value of the contrast C_0. Together L_0, D_0 and C_0 are the factors determining what is termed the "threshold value". In the investigation the various threshold values per test person were studied and compared individually.

Fig. 3. The person under test answers the question by placing the rotatable ring in the same orientation as the one indicated.

[3]) In the extreme left-hand column the contrast was zero (paper of rings same colour as background). This column was not taken into consideration in the tests.

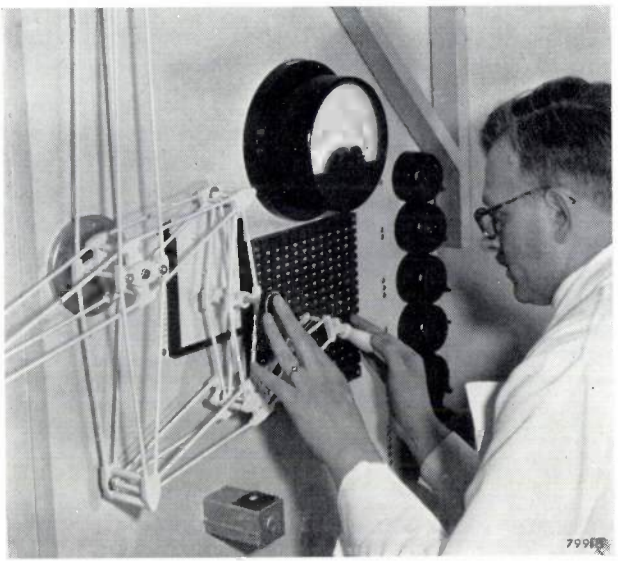


Fig. 4. The examiner at the instrument panel. Top: the voltmeter for checking the voltage of the lamps. With his left hand the examiner presses the needle down. Below: the electric signalling device which shows if the answer is correct.

Test results

The average result of all persons tested can be represented by a three-dimensional diagram, where D_0 is plotted as a function of C_0 and L_0 (*fig. 5*). It will be seen that D_0 diminishes (and consequently the visual acuity increases) at a given luminance as the contrast increases, and at a given contrast value as the luminance increases.

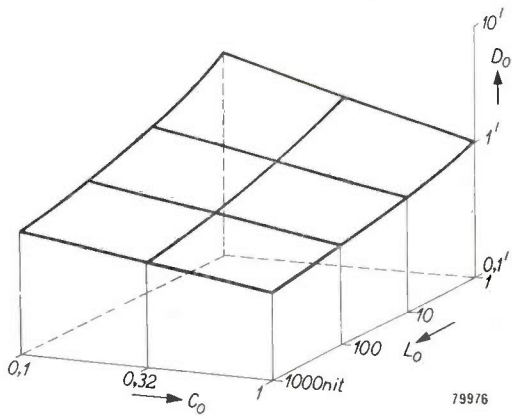


Fig. 5. The average result for all persons, represented in a three-dimensional diagram. It shows the relationship between the contrast value C_0 of the test object, the luminance L_0, and the smallest detail perceived U_0.

The surface in fig. 5 can be expressed by the empirical formula:

$$\log D_0 = 2.17 \frac{1.57 - \log C_0}{\log L_0 + 3.95} - 0.79;$$

where $\log D_0$ represents the average value of $\log D_0$

(for given C_0 and L_0) for all tested persons [4]). For each individual the same formula applies, viz.;

$$\log D_0 = 2.17 \frac{1.57 - \log C_0}{\log L_0 + 3.95} - \log G, \quad . \quad . \quad (1)$$

in which the value of G varies according to the individual tested.

G is a direct measure of the visual power. If in the case of two persons the value of G for the one is n times as large as that for the other, the former will, under equal conditions, be able to discern details that are n times as small [5]). Physically, 0.1 G is the reciprocal value of the angle at which the observer is just able to see the opening in the Landolt ring at given combinations of C_0 and L_0, i.e. when the term of formula (1) containing C_0 and L_0 is equal to 1 (e.g. $C_0 = 1$ and $L_0 = 0.3$ nits = 0.9 lux on white). Formula (1) may also be interpreted in the sense that the product of G and D_0 is a general function of contrast and luminance.

In order to find out to what extent formula (1) applies to a given person, one may use this formula to calculate log G for each of the 40 threshold situations examined (4 luminosities × 10 contrast values). If the formula applies strictly, then the same value should be found in all 40 cases, or at least (owing to accidental measuring errors) a number of values grouped around an average. This appears to be the case with a great number of persons (the normal type). There are cases, however, in which the values found tend to increase or decrease according as C_0 or L_0 increases. Persons for which such results are found do not fit in completely with the general picture as represented by formula (1), but on the other hand the deviations are not so large as to invalidate the formula. In each special case the log G value to be substituted in formula (1) is taken as the average of the 40 individually calculated values quoted above. Finally, according to the behaviour of G, the test persons can be divided into a number of types as indicated in *table I*. Type A is normal and was found to comprise 67 % of all the persons tested.

Table I. Survey of visual types.

Type	Dependence of G on		Numbers of persons (%)
	increasing luminance	increasing contrast	
A	—	—	67
A	—	—	67
B_1	increase	—	11
B_2	decrease	—	5
C_1	—	increase	11
C_2	—	decrease	1
B_1C_1	increase	increase	4
B_2C_2	decrease	increase	1

[4]) Here D_0 is expressed in minutes and the luminance L_0 in nits (cd/m²). In the articles mentioned in note [1]) the brightness (B_0) is expressed in millilamberts (1 mL = 3.18 nits).

[5]) The author has suggested the introduction for G of a unit called "snellen", in honour of the Dutch opthalmologist Herman Snellen (1839-1918).

The influence of age

It appears that the visual power G bears a simple relationship to the age of the test person. The test persons were divided into age groups and the average

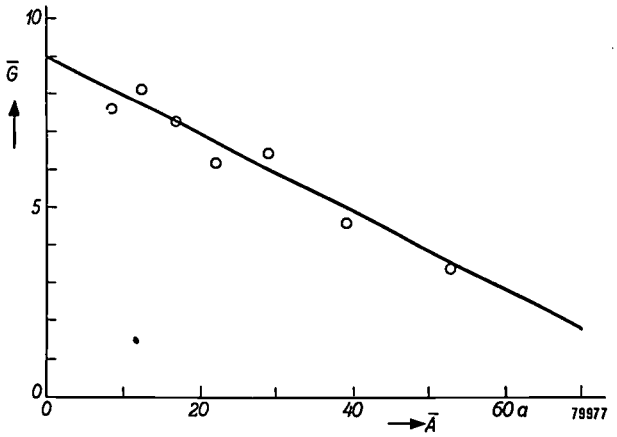


Fig. 6. The average value of the visual power $\bar{G}$ for various age groups, plotted against the average age $\bar{A}$ (in years) in these groups.

value of log G was calculated for each group. The mean value $\bar{G}$ thus obtained has been plotted in *fig. 6* against the average age $\bar{A}$ of the group. It can be derived from this graph that

$$\bar{G} = 9 - 0.1\,\bar{A} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

in which A is expressed in years.

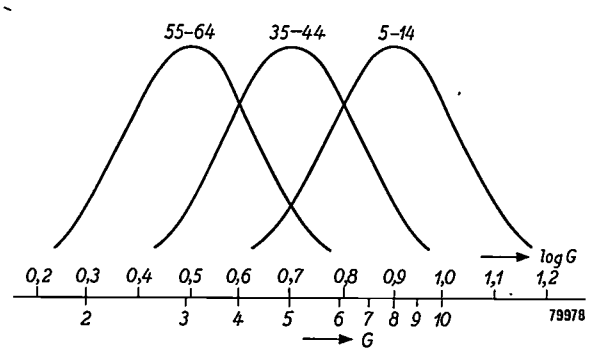


Fig. 7. Relative probability of G values for three different age groups. The figure was ob ained as follows: For each age group it was calculated for how many persons the value of log $G = g$ was situated in the intervals --0.025 to 0.025, 0.025 to 0.075, etc. From these data the average value of log G (g_m) for each age group (see fig. 6) was calculated, and also the standard deviation, i.e. the root mean square of the expression $(g - g_m)$. These standard deviations amounted for six successive age groups to 0.095, 0.15, 0.11, 0.15, 0.28 and 0.14 respectively. For the sake of simplicity it has been assumed that in each group the log G values have a Gaussian distribution, and that for all groups the standard deviation $s_m = 0.16$,i.e. the mean of the values mentioned above. The curves drawn here can thus be expressed by the formula.

$$\exp\left[-(g - g_m)^2 / 2s_m^2\right].$$

This formula has also been used to calculate the curves of figure 8.

Naturally this empirical formula may be applied only within the age limits for which it has been derived, i.e. 10 to 60 years. It shows that visual acuity drops a by factor $^1/_2$ when going from the age of 10 to 50, and, after some slight extrapolation, by a further factor of $^1/_2$ from 50 to 70 years. That even considerable extrapolation does not lead to absurd results appears from the fact that according to formula (2) $\bar{G} = 0$ for $\bar{A} = 90$ years.

Of course there is a certain amount of dispersion for G in each age group. This dispersion is shown by *fig. 7*. *Figure 8* shows that in general the steep drop of G with increasing age, as represented in fig. 6

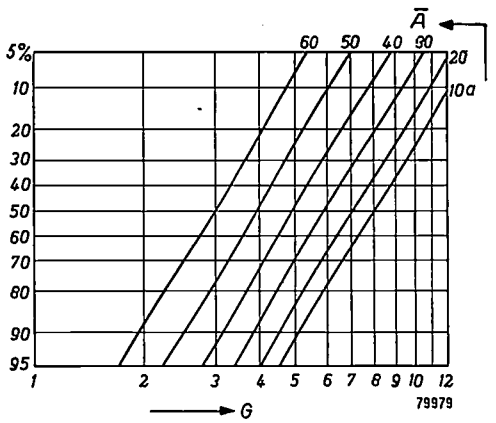


Fig. 8. Percentage of test persons for which the value of G exceeds the value plotted on the abscissa, with the average age $\bar{A}$ of the group as parameter.

and by equation (2), is only slightly affected by this dispersion. Here the percentage of test persons in a given age group whose G value exceeds a given value is indicated. The figure proves that only the best 10% in the 55-64 years group ($\bar{A} = 60$) and the poorest 10% in the 5-14 years group ($\bar{A} = 10$) can see approximately as well as the average ones in the 35-44 year group ($\bar{A} = 40$).

The variations of D_0

When reviewing once more the results of the whole test, one is struck by the fact that the value of D_0, the smallest detail observed, can vary so strongly according to conditions. *Table II* gives an idea of this. It contains the D_0 values calculated with formula (1) for various contrast and luminance values and for various values of visual power G, viz. for the average value of G in the age groups 5 to 14 years, 25 to 34 years, and 45 to 54 years, and for the highest and lowest G values found ($G = 14$, and $G = 1$ respectively).

Apart from this selection of numerical values, some graphic representations of formula (1) are given in *figures 9* and *10*, in order to illustrate the influence of the various variable quantities. In fig. 9*a*, *b* and *c*, three age groups (the same as in table II) are dealt with separately. When applying the result to an individual person, account must be taken of the fact that, owing to the dispersion, the effective age [6]) of a person may differ from his real age.

In fig. 10, L_0 has been plotted against the effective age for three values of the contrast and for two values of D_0, viz. 1′ and 10′ (0.00027 and 0.0027 radians). It follows from this that, in order to discern the same detail with increasing age, the luminance must be more strongly increased for small details than for larger details. This explains why elderly people are inclined (erroneously) to forbid children to read in waning light, telling them

[6]) By effective age A_{eff} is understood the age a person should be in order to have the G value following from formula (2) if G were substituted by G, and $\bar{A}$ by A_{eff}, thus

$$A_{\text{eff}} = 10(9 - G).$$

For G values over 9, A_{eff} is negative.

Table II. The smallest perceptible detail D_0 (in minutes of arc) for three contrast values C_0 and four luminance values L_0 at a visual power G = 1, 4, 6, 8 and 15.

L_0(nit) \ C_0 =	$G = 1$			$G = 4$			$G = 6$			$G = 8$			$G = 14$		
	0.1	0.32	1.0	0.32	1.0	1.0	0.32	1.0	0.1	0.32	1.0	0.1	0.32	0.32	1.0
1	26	13.5	2.7	6.5	3.4	4.1	2.1	1.15	1.15	3.3	1.7	0.9	1.8	0.96	0.5
10	13.5	9.1	4.9	3.4	2.0	12.5	1.2	1.3	0.8	1.7	1.0	0.6	0.95	0.6	0.35
100	9.5	5.8	3.7	22.2	1.4	0.95	1.4	0.9	0.6	1.1	0.7	0.45	0.6	0.4	0.25
1000	6.3	4.5	3.1	1.6	1.1	0.8	1.0	0.7	0.5	0.8	0.55	0.4	0.45	0.30	0.22

ERRATUM

VISUAL ACUITY IN CONNECTION WITH TELEVISION

In the November-December issue of this Review (Vol. **16**, p. 176), Table II of the article by G. J. Fortuin, Visual acuity in connection with television, has been badly misprinted. The editors tender their apologies and the table is here reprinted correctly in a form suitable for pasting over the previous table.

Table II. The smallest perceptible detail D_0 (in minutes of arc) for three contrast values C_0 and four luminance values L_0, at a visual power $G = 1$, 4, 6, 8 and 14.

L_0 (nit) \ C_0 =	$G = 1$			$G = 4$			$G = 6$			$G = 8$			$G = 14$		
	0.1	0.32	1.0	0.1	0.32	1.0	0.1	0.32	1.0	0.1	0.32	1.0	0.1	0.32	1.0
1	26	13.5	7.2	6.5	3.4	1.8	4.1	2.1	1.15	3.3	1.7	0.9	1.8	0.95	0.5
10	13.5	8.1	4.9	3.4	2.0	1.25	2.1	1.3	0.8	1.7	1.0	0.6	0.95	0.6	0.35
100	9.5	5.8	3.7	2.2	1.4	0.95	1.4	0.9	0.6	1.1	0.7	0.45	0.6	0.4	0.25
1000	6.3	4.5	3.1	1.6	1.1	0.8	1.0	0.7	0.5	0.8	0.55	0.4	0.45	0.30	0.22

that they are "spoiling their eyes", because they are no longer able themselves to discern small type under such conditions.

If, on the other hand, the luminance is a given value, then fig. 9 gives the answer to the value of D_0 for the various age groups. The latter conforms with the situation met in television. In television the brightness of the picture can be adjusted to a certain extent by turning a knob, but there is an upper limit to this because, owing to the discontinuity of the image, flicker may occur, and because, even if the other conditions remain unchanged, the critical flicker frequency (i.e. the frequency above which no flicker is seen) is higher as the brightness is greater. With the fluorescent materials commonly used in Europe and at a mains frequency of 50 c/s, the maximum permissible screen luminance can be taken at 100 nits (100 cd/m^2) [7]).

As the tests described above were carried out with the aid of Landolt rings and the details observed in a television image are of a different kind (brightness variations along the lines of the image) one cannot simply apply formula (1) to television. On the other hand the perception of details does not depend strongly on the nature of the detail. It is therefore still admissible to use formula (1) to get at least an impression of the variations in perception and their dependence on the various factors which affect them.

From Table II it can be seen that if $C_0 = 0.1$ and $L_0 = 100$ nits, D_0 may vary from 2.2′ ($G = 4$, $A_{\text{eff}} = 50$ years) to 1.1′ ($G = 8$, $A_{\text{eff}} = 10$ years), and in extreme cases even from 9.5 'to 0.6′. If we assume that formula (1) may be applied quantitatively to the case in question, this means that the distance a between the screen and the eye at which practically all details can be seen, varies from 1.5 metres ($A_{\text{eff}} = 10$ years) to 0.8 metre ($A_{\text{eff}} = 50$ years), and that in extreme cases a may even vary from 3.4 to 0.18 metres!

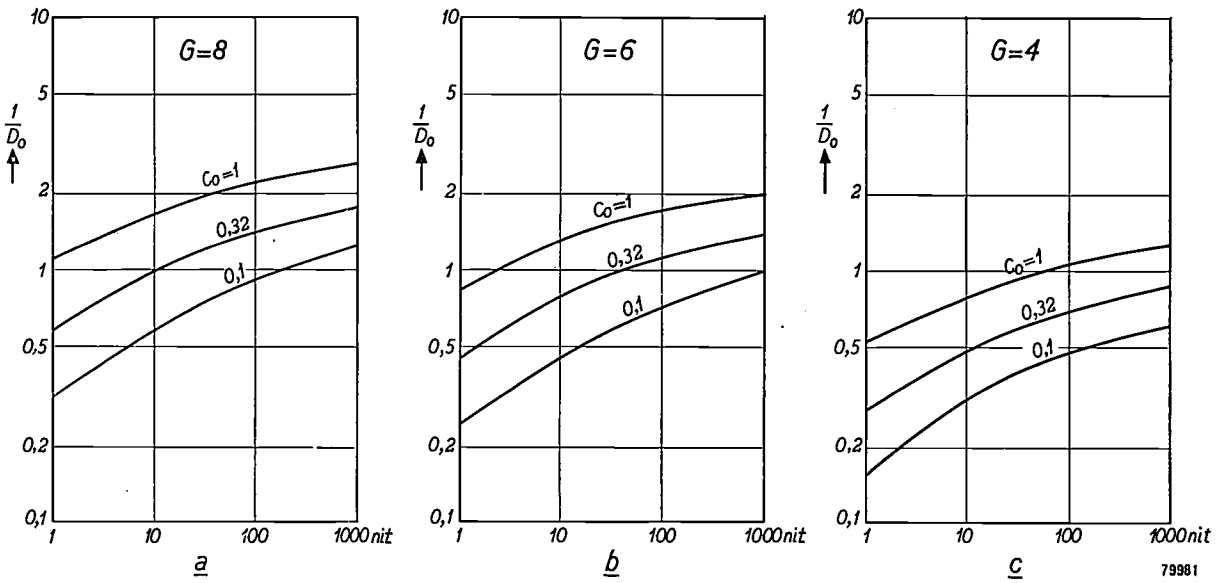


Fig. 9. The visual acuity ($1/D_0$) plotted on a double logarithmic scale against the luminance L_0 of the background, for three different values of the contrast ($C_0 = 0.1$, $C_0 = 0.32$, $C_0 = 1.0$). *a*) Age group 5-15 years; *b*) 25-35 years); *c*) 45-55 years.

The above conslusions can also be applied, mutatis mutandis, to the observation of a projected image (lantern slides, cinema projection). Here, too, with a view to the visibility of the details, it will be desirable to raise the luminance as much as possible, although here too there is an upper limit, this time set by the capacity of projector and lamp and the distance between projector and screen (in the case of cinema projection special measures are taken to avoid the inconvenience of flicker of the image [8]). It is a well-known fact that in general a screen lighting of 100 lux (screen luminance 20 to 100 nits, according to the reflection properties of the screen) is considered satisfactory. If we assume that the smallest details on a 4 × 3 m screen are 2.5 mm in size (this depends on the grain size of the negative with which the exposure was made and on the sharpness of the optical system of the camera) we

[7]) See J. Haantjes and F. W. de Vrijer, Flicker in television pictures, Philips tech. Rev. **13**, 55-60, 1951/52.

[8]) This is explained on p. 161 of the article by J. Kotte, A professional cine projector for 16 mm film, in this issue.

find that the optimum distance between observer and screen varies from 4 to 8 metres according to the observer's age, and from 1 to 18 metres if allowance is made for extreme cases.

We may conclude by pointing out that formula (1) should be used with some care for calculating lighting standards. In the first place the tests refer to threshold values, indicating a minimum which in practice should always be exceeded. It is difficult to say to what extent the minimum should be surpassed, because apart from perception the comfort of the subject plays also a part. Furthermore, the formula is the outcome of laboratory experiments, where *time* is not included as a factor. It will be clear, therefore, that conclusions drawn from such experiments should be applied to cases involving movement and rapidity of perception with great caution.

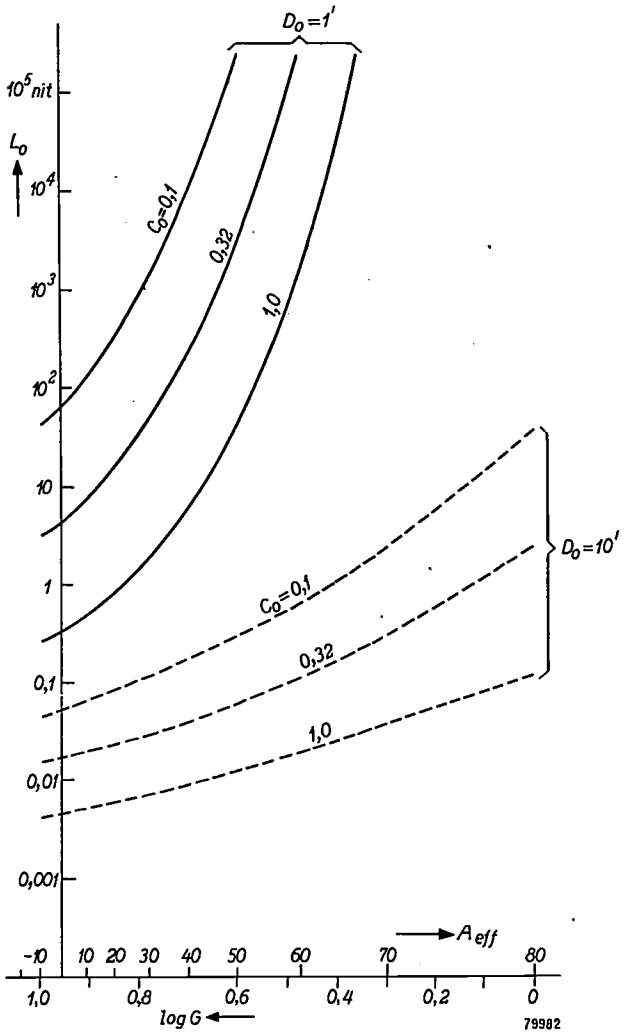


Fig. 10. Luminance L_0 required for perception of a detail $D_0 = 1'$ and $10'$ respectively at contrasts C_0 of 0.1, 0.32 and 1.0, as a function of the effective age A_{eff}.

Summary. In the observation of a television (or projection) image, the visual acuity of the observer is of importance. A measure for this is the smallest detail D_0 which can be discerned at a given contrast and a given luminance of the object. Tests carried out with Landolt rings on 228 persons in ages varying from 7 to 64 years revealed that the product GD_0 is a general function of contrast and luminosity, G being an individual constant (the visual power) for each observer, whose value, apart from a certain statistical variation depends on the age A of the observer in the following way: $G = 9 - 0.1\,A$. A description is given of the manner in which this average result has been obtained, and the restrictions are mentioned that must be taken into account, when applying it to an individual observer. The meaning of the relationship thus found is illustrated by a table and several graphs. It appears that the visual acuity increases by a factor of roughly 2, if a) the luminance increases by a factor 100, or b) the contrast increases by a factor 5. Under equal conditions the visual acuity of persons 50 years old is half that of a ten-year old child, while that of persons 70 years old is only one quarter. Although in the case of details of different kinds it is not permissible to draw quantitative conclusions from the relationship found, it still helps to give an impression of the extent to which the need for light increases with age and of the extent to which, at a given luminance and contrast, the perception of details is better with young persons than with old.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the administration of the Philips Research Laboratory, Eindhoven, Netherlands.

2099: H. J. G. Meyer and D. Polder: Note on polar scattering of conduction electrons in regular crystals (Physica **19**, 255-264, 1953).

The reciprocal relaxation time of conduction electrons due to polar interaction with piezoelectrically active acoustical vibration modes is calculated for two-atomic regular crystals. It is found to be proportional to $T^{1/2}$. It is shown for the case of ZnS (sphalerite) that below 150 °K the numerical value of this reciprocal relaxation time is comparable with the reciprocal relaxation time due to polar interaction with optical vibration modes. In crystals with sodium chloride structure the mass difference of the ions may in principle be responsible for polar scattering by acoustical modes. The corresponding reciprocal relaxation time is proportional to $T^{3/2}$. In the special case of LiF it becomes equal to polar scattering by optical modes at about 200 °K.

2100: H. A. Klasens: On the nature of fluorescent centres and traps in zinc. sulphide (J. Electrochem. Soc. **100**, 72-80, 1953, No. 2).

The impurity levels in the energy diagram of a zinc sulphide phosphor are considered to be localized S^{2-} levels lifted above the filled S^{2-} band due to the presence of monovalent positive or trivalent negative activator ions in the lattice. Electron traps are formed similarly by the substitution of S^{2-} ions by monovalent negative ions or of Zn^{2+} ions by trivalent positive ions. The energy produced when electrons recombine with trapped holes or when holes recombine with trapped electrons is either emitted directly as light or is first transferred to impurity ions. The elements of the iron group give rise to electron traps. The killing action of these elements is explained by assuming that the energy liberated by recombination between holes and electrons in these traps is transferred to the killer ions. The excited ions return to the ground state without radiation, because of the presence of many electronic levels between the excited and the ground state. The effect of heat and infrared radiation on the luminescence is discussed. It is shown that, in a phosphor, energy may be transferred by electrons through the conduction band or by holes through the occupied S^{2-} band.

2101: J. I. de Jong: The determination of formaldehyde in the presence of phenols and phenol alcohols (Rec. Trav. chim. Pays-Bas **72**, 356-357, 1953).

A method described by Pfeil and Schroth for the quantitative determination of formaldehyde appears to be applicable in the presence of phenols and phenol alcohols.

2102*: P. Cornelius: L'électricité selon le système Giorgi rationalisé (Ed. Dunod, Paris 1953, 116 pp.). (Electromagnetism according to the rationalised Giorgi system: in French.)

French translation of Dutch text (see these abstracts, No. **1803**). The author has extended and altered the text at a few points. The generator (dynamo) is treated from the viewpoint of an observer moving with the rotor (alternating flux) as well as from that of an observer at rest (Lorentz force on electrons in wire). The case of a beam of electrons in a magnetic field is considered. In the final chapter the author explains why he avoided discussions on quantity equations. (The equations in Giorgi units as given in the text may be considered either as numerical or as quantity equations.)

2103: H. Bremmer Eine einfache Näherungsformel für die Feldverteilung längs der Achse magnetischer Elektronenlinsen mit ungesättigten Polschuhen (Optik **10**, 1-4, 1953, No. 1). (A simple approximate formula for the field along the axis of magnetic electron lenses with unsaturated pole-pieces; in German.)

For the field $H(z)$ along the axis of a rotationally symmetrical magnetic electron lens with unsaturated pole-pieces, the formula

$$H = (0.4\pi)\,NI\left\{\varphi\left(\frac{z}{b}+\frac{s}{2b}\right)-\varphi\left(\frac{z}{b}-\frac{s}{2b}\right)\right\}\Big/ s$$

is given, in which s is the width of the slit, b the diameter of the bore, and $\varphi(x)$ a function which, depending on the value of x, can be expanded in a series in different ways. A table of $\varphi(x)$ for $0 < x < \infty$ is given. The results are in agreement with a (less simple) theory of Lenz.

2104*: J. A. Haringx: Stresses in corrugated diaphragms (pp. 198-213, C. B. Bienezo Anniversary Volume on Applied Mechanics, Stam, Haarlem 1953).

The stress components in the material of a corrugated diaphragm are determined, approximately, by replacing the diaphragm by a ficticious flat plate of similar properties as described in a previous paper. From the most unfavourable combination of these stress components that is ever possible, formulae for the maximum stresses were derived. These formulae were tested in a special case for which the stresses have been worked out in full by Grover and Bell, and are found to be in good agreement. Thus these formulae may be useful for design purposes.

R 216: F. van der Maesen, P. Penning and A. van Wieringen: On the thermal conversion of germanium (Philips Res. Rep. **8**, 241-244, 1953, No. 4).

Experiments were carried out regarding the conversion of n-type germanium into p-type by heat treatment at 800 °C. The results indicate that the conversion is due to the presence of acceptor impurities on the Ge surface prior to heating. As long as the surface is clean, its roughness seems to have no influence on the conversion. Experiments with pieces of Ge saturated by diffusion with Cu at 800 °C show that the Cu can be inactivated by heat treatment at 500 °C but that it remains present in some form throughout the material. A similar effect was found with Ni.

R 217: W. K. Westmijze: Studies on magnetic recording, III (Philips Res. Rep. 8, 245-269, 1953, No. 4).

Continuation of **R 213** end **R 214**. The recording methods leading to a linear relationship between the input signal and the recorded magnetization are discussed. Special attention is paid to the a.c. biasing method and its relation to ideal magnetization. This discussion is based on magnetic measurements in homogeneous fields. A magnetic model explaining the linearizing effect of the a.c. biasing field is discussed.

Next follows the calculation of the magnetic field that exists in and around a sinusoidally magnetized tape, the cases of longitudinal and perpendicular magnetization being treated separately. When the permeability of the tape is greater than unity, the demagnetizing field in the tape effects a decrease of the recorded magnetization. The flux in an ideal reproducing head is calculated for this case. It is shown that longitudinal and perpendicular magnetization produce the same flux in the head only when the permeability of the tape is equal to 1.

R 218: J. L. Meijering: Interface area, edge length, and number of vertices in crystal aggregates with random nucleation (Philips Res. Rep. 8, 270-290, 1953, No. 4).

The interface area, edge length, and numbers of faces, edges and vertices in an aggregate consisting of a large number of crystals are calculated for two models. In the first ("cell model") the crystals start to grow simultaneously and isotropically from nuclei distributed at random. In the second ("Johnson-Mehl" model) the nuclei appear at different moments, the rate of nucleation being constant. Corresponding calculations are made for plane sections of the aggregates and for two-dimensional aggregates. For the one-dimensional case the size-distribution curves are calculated. From a discussion of the results it is concluded that in the two- and three-dimensional cell models, the crystals are less equiaxial than in the corresponding Johnson-Mehl models.

R 219: A. H. Boerdijk: The value of the constant in Wien's displacement law (Philips Res. Rep. 8, 291-303, 1953, No. 4).

It is shown that the current definition of monochromatic intensity (I_λ) contains an arbitrary element. Other definitions are considered, based on a logarithmic wavelength scale (I_r) or on a frequency scale (I_ν). For black-body radiation the maxima of these intensities occur at different places in the spectrum, giving rise to three different constants in Wien's law. The question arises which scale should be used. It appears that the use of I_λ is a matter of convention only. There are logical arguments for using I_r, the corresponding Wien-constant being 0.3668 cm °K.

R 220 and **R 221:** F. C. Romeijn: Physical and crystallographical properties of some spinels (Philips Res. Rep. 8, 304-342, 1953, Nos. 4 and 5).

From X-ray measurements on simple and complicated spinels, regularities in the ionic distribution and lattice constants have been investigated. These regularities have been explained partly by general methods (Madelung potential, geometrical considerations) and partly by also taking into account the individual properties of the ions that are correlated with the distribution of electrons within the ion. The calculated correlation between the ionic distribution and the oxygen parameter u was found to be confirmed by experiment. The ultimate choice of the distribution, however, is governed by the individual properties of the ions, partly by their dimensions and partly by the distribution of electrons. Some physical properties of the compounds investigated have been correlated with the ionic distribution.

VOL. 16 No. 7, pp. 181-212 JANUARY 1955

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

INDUCTIVE AERIALS IN MODERN BROADCAST RECEIVERS

I. BRIEF HISTORY AND GENERAL DESCRIPTION

II. TECHNICAL ASPECTS OF INDUCTIVE AERIALS

by H. BLOK and J. J. RIETVELD.

621.396.677.5:621.396.62

*After the broadcast receiver had grown out of the amateur stage and began to find its way into every home, one of the problems arising was that of the aerials. In the large cities, where a large proportion of the population is housed, as it were, piled up in layers, a veritable forest of aerials began to appear on the roof tops. To combat this eyesore, one of the solutions sought was the system of one central aerial, feeding a number of receivers via a high-frequency booster *).*

The fact, however, that the normal roof aerial (T.V. aerials will not be considered here) has largely disappeared, is due to the fact that nowadays a large public is satisfied with a small indoor aerial, strung in the room. With the high sensitivety of the modern set, a reasonable signal strength can be obtained with such a room aerial, but all too often it is marred by local interference. Already before the war, it was realized that a properly designed inductive or frame aerial has relatively a far smaller sensitivity to interference than a (capacitative) room aerial. Receivers incorporating one or more inductive aerials have been manufactured on a large scale in the last few years. The development of the material Ferroxcube has been of great importance in this respect.

The first part of the article below gives a general review of the development of the inductive aerial. The second part enters further into various technical details.

I. BRIEF HISTORY AND GENERAL DESCRIPTION

Advantages of an inductive aerial over a capacitative aerial

When considering the course of development of the normal broadcast receiver over the last 15 years, the important part played by progress in component manufacture should be noted. The refinement of manufacturing methods and the application of new materials have made it possible to make many component parts smaller, cheaper, and often of a better quality.

The magnetic material Ferroxcube has contributed significantly to these developments. Not only does it permit the manufacture of excellent H.F. and I.F. transformers of small dimensions [1]), but it has also led to the design of frame aerials as an intrinsic part of the radio receiver.

In the course of the years various aspects of the frame aerial have been discussed in this Review. The application of its directional effect, for the location or direction finding of ships and aircraft [2]) and its use for the reduction of interference from unwanted transmitters [3]) are well known. The frame aerial is also ideal for measuring the field strength of not-too-short waves [4]), owing to the fact that it gives excellently reproducible results and that its effective height can be calculated with great accuracy.

*) J. van Slooten, The communal aerial, Philips tech. Rev. **1**, 246-251, 1936.

[1]) W. Six, Some applications of Ferroxcube, Philips tech. Rev. **13**, 301-311, 1951/52.

[2]) See for example Philips tech. Rev. **2**, 184-190, 1937.

[3]) P. Cornelius and J. van Slooten, Installations for improved broadcast reception, Philips tech. Rev. **9**, 55-63, 1947.

[4]) M. Ziegler, A recording field strength meter of high sensitivity, Philips tech. Rev. **2**, 216-223, 1937.

Aerials of this kind in broadcast receivers, are not of course used to determine the direction of the transmitter, or to apply a signal of accurately reproducible strength to the input of the receiver, but to obtain reception free of interference, in particular, local interference caused by electrical apparatus, such as motors, switches, etc. As demonstrated by Cornelius [5]), a frame aerial is less sensitive to such interference than an ordinary, capacitative aerial of the same effective height. This phenomenon can be briefly explained as follows: If a transmitter is received at a distance which is great with respect to the wavelength, then the receiver is situated in what is called the radiation field of the transmitter.

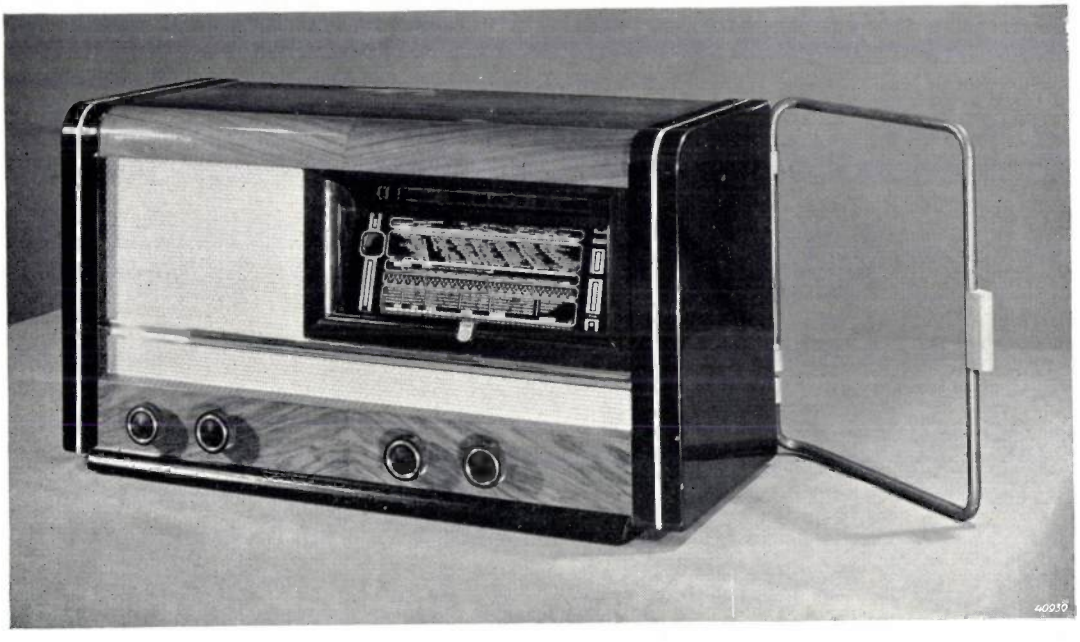


Fig. 1. Philips de-luxe receiver 902 A (1939), equipped with a single-loop frame aerial.

In this field there exists a certain fixed ratio between the strengths of the electric and the magnetic alternating fields at the position of the receiver, viz. the electric field strength E (expressed in volts per metre) is $4\pi c \times 10^{-7}$, i.e. approximately 120π times the value of the magnetic field strength H (in amperes per metre); c being the velocity of light (in metres/sec). Since a capacitative aerial reacts to the electric component of the radiation field, whereas an inductive aerial reacts to the magnetic component, it will be clear that a local source of interference will have the same effect on both types of aerials only if the ratio E/H between the electric and the magnetic field strengths of the disturbance is likewise 120π. It is found, however, that for waves longer than 200 m, the ratio E/H from nearby sources of interference is far greater than 120π. An inductive aerial, therefore, will receive correspondingly less of this type of interference than a capacitative aerial.

It follows from the above that it is desirable for an inductive aerial to be free of all capacitative sensitivity or, as it is generally expressed, it should show no "aerial-effect". Means to combat this effect have already been described in this Review [6]), viz: symmetrical construction (centre of the inductive aerial earthed), electrical screening, and reduction of the number of turns. The first remedy alone is insufficient, as symmetry can never be adequately realized in practice. The second has the drawback that the screening considerably increases the minimum capacitance (so that the wavelength range of an ordinary tuning capacitor becomes very limited). The third remedy, however, is very effective, because the aerial-effect is proportional to the self-inductance of the aerial, and thus to the square of the number of turns n, whereas the voltage induced by the alternating magnetic field is proportional to n. This led Cornelius to the construction of a frame with only one turn [6]). It was applied in the Philips

[5]) P. Cornelius, The sensitivity of aerials to local interferences, Philips tech. Rev. **6**, 302-308, 1941.

[6]) P. Cornelius, The aerial effect in receiving sets with loop aerial, Philips tech. Rev. **7**, 65-73, 1942.

receiver, type 902 A (*fig. 1*). This de-luxe set was marketed in 1939 and may be considered as the forerunner of the present frame-aerial receivers.

The general introduction of inductive aerials in mains-fed receivers [7]) is a consequence of a number of factors:

a) The use of all kinds of electrical appliances is steadily increasing, resulting in increased mains-interference.

b) The population in the large towns becomes more and more concentrated, so that a relatively greater number of wireless sets are used in surroundings with considerable interference.

c) Many local authorities have enacted legislation against the unlimited erection of roof aerials, so the use of (capacitative) room aerials has become more common. Although several broadcasting stations can be received with sufficient strength (thanks to powerful transmitters and sensitive receivers), a room aerial leaves much to be desired as regards freedom from interference.

d) Small receivers, with dimensions of about $20 \times 30 \times 15$ cm, have become very popular. One of their most attractive features is that they are easily transportable from one place to another, and they should therefore not be bound to one spot by an aerial connection.

e) The frequency allotment in the medium wave range according to the Copenhagen plan (1950) has not been universally accepted, with the result that many transmitters, especially in the evenings, are jamming each other's broadcasts. The directional effect of inductive aerials can be most useful to avoid interference of this type.

Development in Europe and the U.S.A.

Bearing these factors in mind it is of interest to note that the inductive aerial has developed along different lines in different countries. In Holland, after the appearance of the receiver 902 A with its single-loop aerial, several years passed before this principle was further applied in practice, due principally to interruption of development by the war. Moreover, the construction was considered too expensive and the hinged frame (fig. 1) did not enhance the appearance and limited the positioning of the set.

Paradoxically, the stimulus to further development came from receivers in the lowest price class. The dimensions of a set of this kind were small and the weight was low, as the tubes were directly fed from the mains without the use of a transformer. These properties made the set very suitable for use wherever mains supply was available, provided only it had its own aerial.

In the U.S.A., where receivers of this type rapidly became popular (at one time they sold for $ 10 to $ 15), they were equipped with a frame aerial. The coil of the first tuned circuit was in the form a fairly large conducting spiral, cemented to the rear panel of the set; an easily manufactured frame aerial was thus obtained, eminently suitable for mass production. This type of frame aerial worked very satisfactorily in the small American receivers, and is still being used on a large scale.

During the development of these small receivers in Europe it was soon found undesirable, however, to adopt this design without modification, due to the greater diversity of wavelengths in Europe than in the U.S.A. In the U.S.A., there is a large network of medium-wave stations (many of them transmitting identical programmes) but long-wave broadcasting stations are unknown. The small American receivers are therefore made for medium wave reception only, whereas European receivers require additional facilities for receiving long-wave stations. There is, moreover, the fact that in Europe short-wave reception is also considered desirable — the listening public has become accustomed to this facility and therefore demands it even in a small set.

If a small set is to receive these three ranges with a frame aerial, the best results are obtained if there is a separate frame for each range. This would mean incorporating three frames aerials in one set. Due to the compact construction of the whole, a certain inter-coupling of the frames is inevitable. A consequence of this is that if the set is tuned, for example, to a medium-wave station, the self-capacitance of the long-wave aerial forms a parasitic absorption circuit, which introduces losses. If, to avoid this, the long-wave aerial is connected in parallel with the medium-wave aerial, the total self-capacitance assumes such a value that the the range of an ordinary tuning capacitor does not cover the entire medium-wave band.

This applies particularly if the self-capacitance is already large due to electrical screening of the frame aerial. As mentioned before, such screening is desirable in order to reduce the sensitivity to interference (cf. the article [6])), especially with long-wave reception [8]). In this respect, too, the situation for small receivers is less favourable in Europe than in America.

[7]) Battery sets, to which quite different considerations apply, will not be discussed here.

[8]) L. Blok, Combating radio interferences, Philips tech. Rev. 4, 237-243, 1939.

All these difficulties made the American design less suitable for small European sets. Neither was the one-loop frame aerial suitable, since it necessitated an h.f. transformer, which at that time was too expensive a component. For these small receivers it was necessary to resort to a small capacitative aerial. This took the form of a strip of metal or metallized paper fitted on the inside of the cabinet, and served for all three wavelength ranges ("Philetta" 209 U).

With regard to their sensitivity to mains interference, the American type of multi-turn frame aerial and the "Philetta" type of capacitative aerial are nearly identical [5]) (*fig. 2*); it was found that in the small American receivers, the capacitative component of the total reception was about as great as that of the above capacitative aerial (approx. 12 pF). This demonstrates how seriously aerial-effect in a frame aerial can increase the sensitivity to mains interference. The performance of the receivers with the small capacitative aerial, may be summarized as follows:

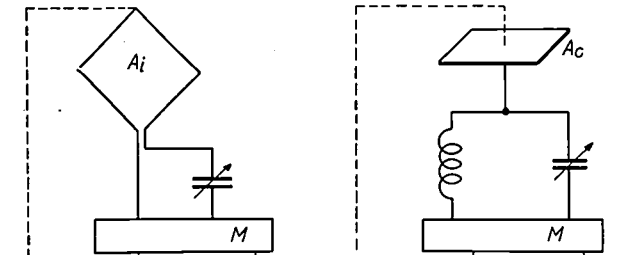


Fig. 2. Schematic representations of (*a*) receiver with inductive aerial (A_i), and (*b*) receiver with capacitative aerial (A_c). C_1 capacitance between aerial and surroundings, C_2 capacitance between chassis and earth, V_n sources of interference, M chassis.

a) In the short-wave range, surprisingly good; practically no mains interference.

b) For medium wavelengths, reception quality about equal to the small American sets with (non-screened) frame aerial: very good at night, but during the day, when listening to remote stations, a fairly high interference level which sounds like noise.

c) On long waves, generally excessive interference; in specific centres of interference, hardly any reception was possible, unless the field of the transmitter was extremely strong.

As is shown in *fig. 3*, mains interference can be divided into that from specific sources such as vacuum cleaners, switches, etc., and a nearly continuous spectrum of interference of uncertain origin,

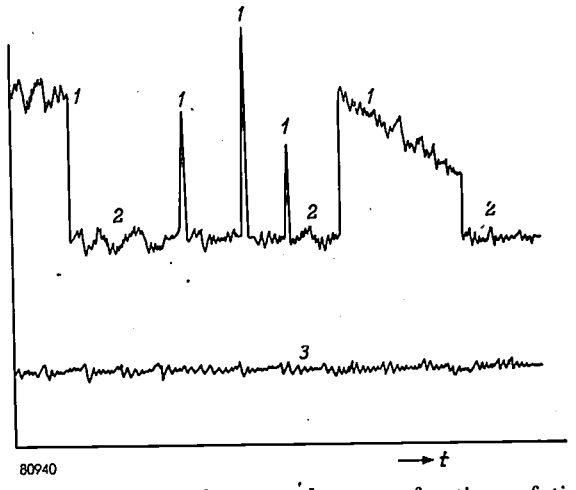


Fig. 3. Various interference voltages as functions of time (shown schematically). *1* interference from electric motors, switches, etc. *2* continuous interference pattern, often of unknown origin. *3* noise.

generally at a lower level. This spectrum of interference, mentioned under *b*, is similar to noise (hiss); in fact, it is difficult to distinguish between them. Actually, the interference spectrum sounds rather more "coarse grained" than the real noise and its intensity diminishes at increasing frequency [8]).

It has further been observed that outside densely populated areas the specific interference is considerably weaker, but that the continuous spectrum of interference may be substantially present without any apparent source.

"Philetta" receiver with single-loop frame

In order to improve reception in the medium and long ranges, a receiver was developed in 1949, (the BX 290 U, also in the "Philetta" class), incorporating a single-loop frame aerial. It was thus, about ten years after the introduction of the single-loop frame aerial in the de luxe set 902 A, that it made its come-back in one of the lowest-priced receivers. This course of events is an example of how a basically sound idea may be shelved for years until suitable technological facilities become available for its exploitation. Because of the small self-inductance of a single-loop aerial a step-up transformer is necessary to couple it to the tuning capacitor of the input circuit (*fig. 4*). In the 902 A set this transformer was a quite expensive component. However, by using a small Ferroxcube rod as a core [9]) it is possible to reduce the dimensions dras-

[9]) It should be noted that as early as 1946, Ferroxcube was used on a small scale for this purpose; cf. page 63 of the article referred to in [3]).

tically, and hence, also the price. This development has enabled the interference-free frame aerial of few turns to find a place in receivers in the lower price brackets.

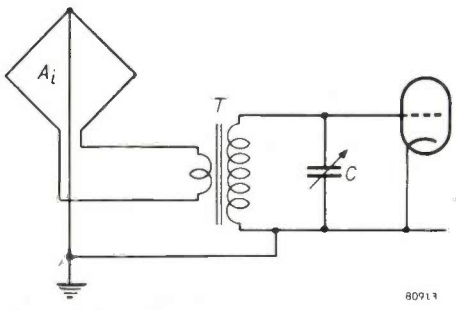


Fig. 4. A_i single-loop frame aerial. T h.f. step-up transformer. C tuning capacitor.

In the BX 290 U, the loop is in the form of an aluminium strip bent around the chassis (*fig. 5*). The small size of this radio means that the whole set can be easily turned; the need for a hinged construction is therefore dispensed with and the whole design is thereby considerably simplified. Fig. 5 shows how the frame has been built-in to the chassis. The mid-point of the loop is firmly screwed to the chassis, primarily for mechanical rigidity but also to further reduce the remaining small aerial-effect (see [6]).

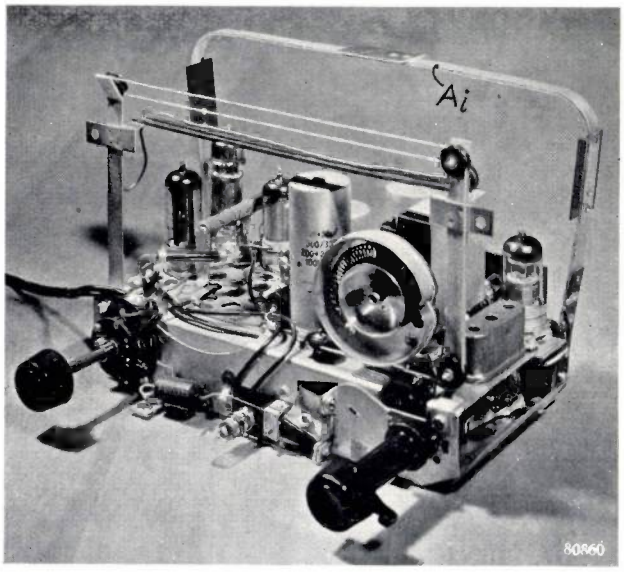


Fig. 5. Chassis of the Philetta BX 290 U receiver (1949). A_i single-loop frame aerial consisting of an aluminium strip.

Short waves are received with a small capacitative aerial (metal plate inside the receiver), which as mentioned above, is very effective for the short waves.

Modern inductive aerials

Frame aerials with few turns in larger receivers

After the single-loop frame had proved a success in the smaller sets, attempts were made to use it in larger receivers. The initial experiments, however, were disappointing: after tuning to moderately strong station, the "noise-like" interference was far stronger than with sets of the "Philetta" type, in spite of the greater loop surface in the larger set. The cause of this interference was the mains transformer, a component not present in the smaller sets, but necessary for several reasons in the larger receivers. It was found that the magnetic field of the mains transformer induced an interference spectrum in the frame aerial as soon as the primary was connected to the mains. Since the transformer consists of several layers of windings, each being coupled inductively and capacitatively to each other, the transformer contains a large number of oscillatory circuits. Many of these have resonance frequencies within the broadcasting range. When the transformer is connected to the mains, these oscillatory circuits are excited (even if a filter suppressing the high frequencies is incorporated between mains and transformer) and oscillate at their own fundamental frequencies. The result is a near-continuous interference spectrum, the amplitude of which diminishes at higher frequencies. The greatest trouble is therefore experienced in the long-wave range.

This interference can be effectively suppressed by means of a short-circuited copper strip surrounding the transformer (*fig. 6*). This strip serves as a short-circuited winding for the stray field of the transformer.

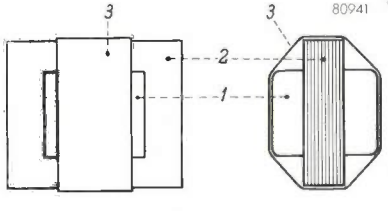


Fig. 6 Mains transformer (*1* coil, *2* core) provided with a closed copper loop (*3*) which prevents the stray field of the transformer from inducing an interference voltage in the frame aerial.

After the introduction of the short-circuit winding, there remained for the larger sets the problem of the directional effect. Since the primary reason for incorporating an inductive aerial in a receiver is the reduction of mains interference, the directional effect is considered a nuisance rather than an advantage. After the set has been tuned in to a certain station, it may involve an additional mani-

pulation, viz. turning the aerial to achieve the optimum signal. A hinged frame outside the cabinet as used with the 902 A (fig. 1), was not considered acceptable, and turning the whole receiver, whilst not objectionable for a small set, would hardly do for a large one.

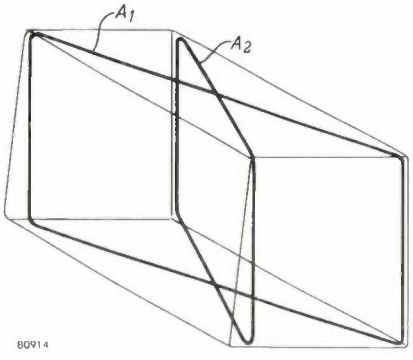


Fig. 7. Mounting of two crossed frame aerials (A_1, A_2) in a radio cabinet.

This difficulty was solved by providing the set with *two* fixed frame aerials, arranged at a certain angle to each other (*fig. 7*) with a switch for switching over from one aerial to the other, so that it is a simple matter to select the frame that gives the best reception from a given station. This system is applied, for example, in the receiver type BX 430 A-10 (*fig. 8*).

The angle between the two frames need only be small, as will be demonstrated in the second part of this article.

Each frame consists of two turns, for the following reason. The switch contacts (for switching over from one frame to the other and from medium to long-wave reception) are incorporated in the frame circuit. The contact resistance, although as low as a few mΩ per contact, would considerably impair the quality factor of a single loop circuit; in a two-turn frame, however, with its greater self-inductance and resistance, the influence of the contact resistance may be neglected. At the same time, the sensitivity to interference is not noticeably increased.

Both the single and two-turn frames can be classified as "low impedance frame aerials", in contrast with the Ferroxcube aerials, which will now be discussed.

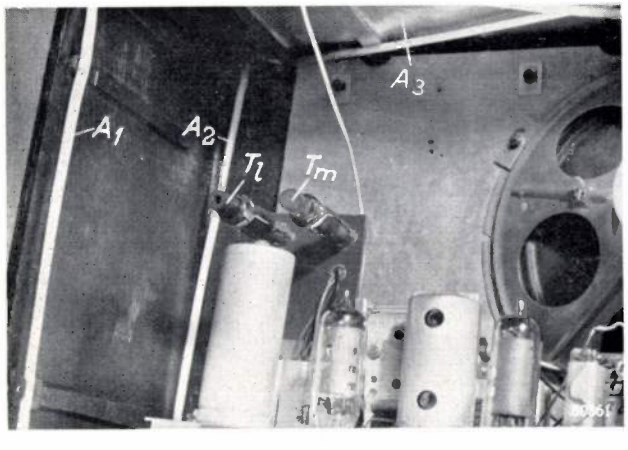


Fig. 8. Interior view of receiver BX 430 A-10. The two crossed frame aerials, (A_1, A_2, each consisting of two turns in this case) are partly visible. T_1 h.f. transformer for long wave reception, T_m h.f. transformer for medium wave reception, both having Ferroxcube cores. A_3 capacitative plate aerial for short wave reception.

Ferroxcube aerials (ferroceptors)

Another type of inductive aerial in which Ferroxcube plays an even more important part than in a frame aerial of few turns, consits of a coil wound on a Ferroxcube rod. A coil diameter of 6-15 mm is sufficient, owing to the exceptionally high permeability of the Ferroxcube; the magnetic flux of a far greater area is, as it were, concentrated in the core (see *fig. 9*).

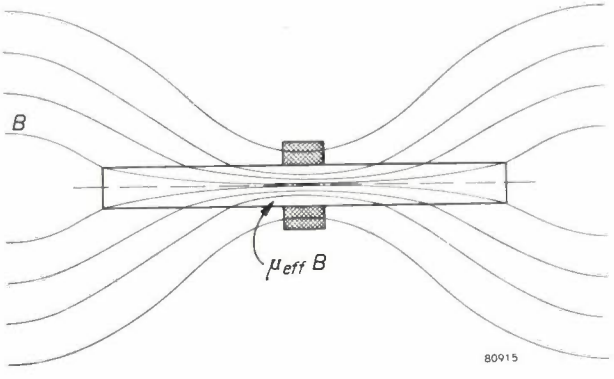


Fig. 9. A Ferroxcube rod concentrates an external magnetic field of induction B into a field of induction $\mu_{eff}B$ at the middle of the rod.

Without any intermediate transformer, the coil and the tuning capacitor constitute the first resonant circuit; the coil therefore has a far greater impedance than a frame of one or two turns. The shape of the coil no longer justifies the name of frame aerial, so the device is referred to as a Ferroxcube antenna or ferroceptor.

In Europe, Ferroxcube aerials were first manufactured on a large scale by "La Radiotechnique" of Suresnes. After meeting with great success in France they were also introduced in other countries. In the small medium/long wave Philips receiver BX 221 U-10, a single rod of Ferroxcube is used, fixed in a permanent position inside the set (*fig. 10*) and provided with one coil for long waves and another for medium waves. The slightly bigger set BX 321 A-10, has two fixed rods, one for long waves and one for medium waves; a capacitative plate aerial is used for short waves (*fig. 11*).

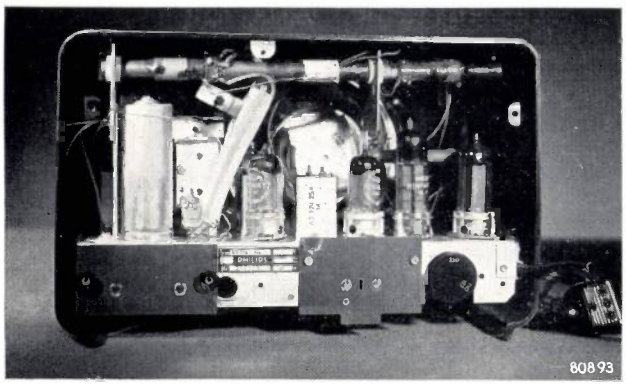


Fig. 10. Receiver BX 221 U-10, equipped with a long wave coil and a medium-wave coil on a single rod of Ferroxcube (ferroceptor).

In large sets, such as the BX 632 A and the BX 633 A, the two rods can be rotated in a horizontal plane by means of a knob (*fig. 12*).

The question arises as to how far the influence of the aerial-effect will be felt, in view of the fact that the coil has a fairly large number of turns. A favourable circumstance is that the dimensions of the coil are small, so that the capacitance with respect to the surroundings is small. A further reduction of capacitative reception can, if necessary, be effected by two measures: compensation of the aerial-effect and electrical screening.

Fig. 13 shows how the aerial-effect can be compensated. The interference voltage from, say, a source V_n in the mains is coupled by the capacitance C_a of the ferroceptor coil L. This voltage can be compensated by means of a second coil L', coupled to L and having a capacitance C_a' with respect to its surroundings; if C_a' has the correct value, the interference current through L' is such that it induces in L a voltage which cancels the former interference voltage.

Let us assume that the number of turns of L' is p times that of L, and let k represent the coefficient of coupling between the two coils. The condition for compensation is then approximately given by

$$k = -\frac{1}{p} \cdot \frac{C_a}{C_a'}.$$

Either k may be varied by altering the distance between the coils, or C_a' may be varied by adjustment of an earthed plate placed in the vicinity (fig. 11).

A minor disadvantage is that the equilibrium is disturbed by any slight capacitance change, so that a hand approaching the set may cause a noticeable rise of the interference level.

In the medium wave range the interference level is so much lower that the receiver BX 321 A-10 does not require any compensation, but only a some electrical screening. This is effected by placing the

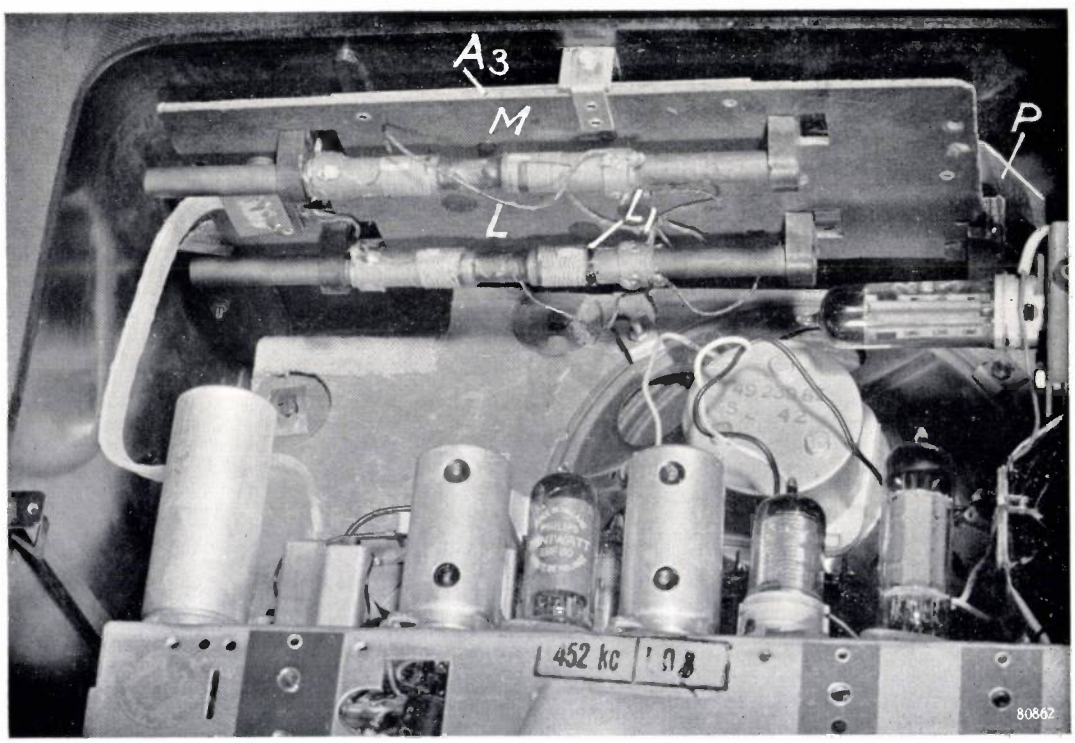


Fig. 11. The receiver BX 321 A-10 has separate Ferroxcube rods for receiving long waves (L) and medium waves (M). For long wave reception the aerial effect is compensated by means of a compensating coil L_1 and the adjustable plate P (cf. fig. 13).

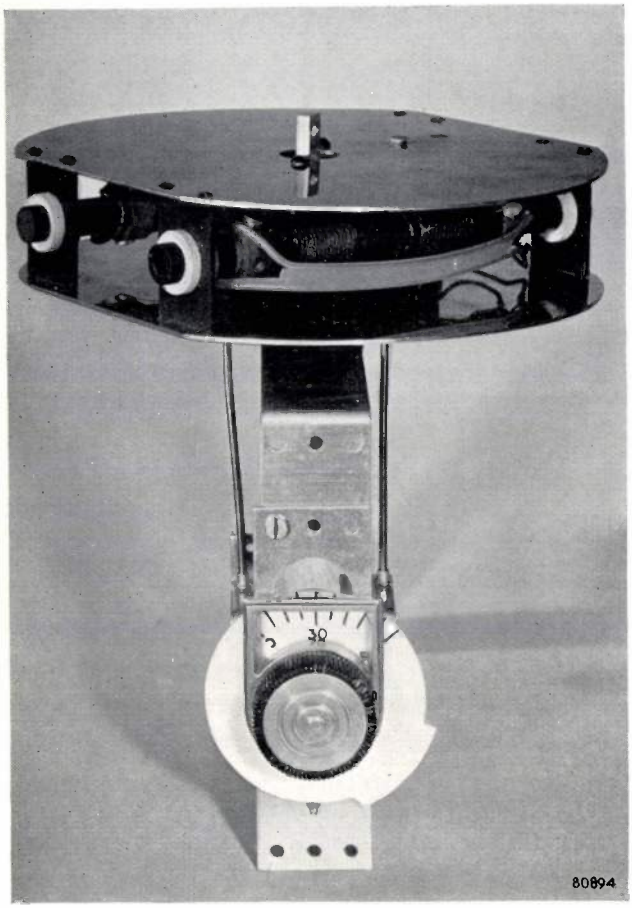


Fig. 12. Two rotatable Ferroxcube aerials (long and medium waves). Some sets (e.g. BX 632 A and 633 A) have a construction allowing the set to receive on a capacitative (outdoor) aerial in the zero position of the knob. When the knob is turned clockwise, this aerial is disconnected and the scale indicating the position of the Ferroxcube antennae lights up as a sign that reception now takes place exclusively via the Ferroxcube aerials. (In versions later than the one depicted, the scale range is more than 180°.)

ferroceptors between the chassis and the capacitive plate aerial (fig. 11); the latter is connected to the chassis during medium-wave reception. In the construction shown in fig. 12 the two rods are mounted between two earthed metal plates. Screening may be further improved by surrounding this assembly by a cage made of copper wire running parallel to the core, interwoven with nylon threads in the perpendicular direction (to avoid magnetic screening).

In contrast with the screening of ordinary frame aerials, the above-mentioned methods only slightly increase the self-capacitance, owing to the small dimensions of the coil.

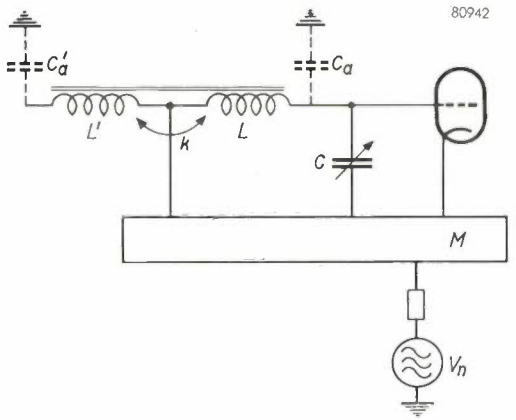


Fig. 13. Circuit for compensating the aerial-effect of a Ferroxcube aerial L. L' = compensating coil coupled with L (L_1 in fig. 11).
C_a and C_a' = capacitance of L and L' respectively with respect to the surroundings. At suitable values of C_a' and of the coupling coefficient k, no interference voltage will be produced across the tuning capacitor C (V_n = source of interference). M = chassis.

II. TECHNICAL ASPECTS OF INDUCTIVE AERIALS

This section deals with the directional effect and with three factors of particular importance in design: sensitivity, signal-to-noise ratio and effective height.

The directional effect

The signal voltage induced in a frame aerial, assuming the complete absence of capacitative reception, is proportional to $\sin\alpha$, where α is the angle between the perpendicular to the plane of the loop and the direction of the transmitter. Plotting the signal voltage in a polar diagram gives the well-known figure-of-eight curve (*fig. 14*).

At a value of $\alpha = 30°$ the voltage is reduced to half the maximum value, whilst the minima at $\alpha = 0$ are sharply defined.

In order to evaluate the directional effect, however, we are not so much concerned with the voltage

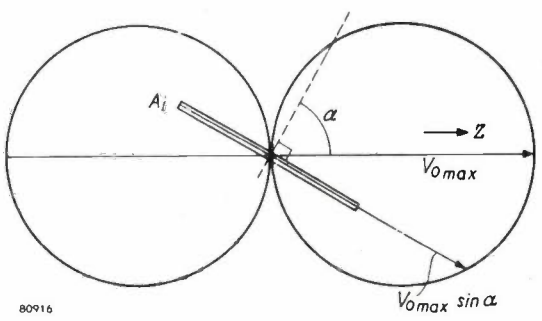


Fig. 14. The figure-of-eight polar diagram of an inductive aerial (A_i) having no capacitative component. The signal voltage in the aerial is $V_{0max} \sin\alpha$, α being the angle between the perpendicular to the plane of the loop and the direction Z of the transmitter.

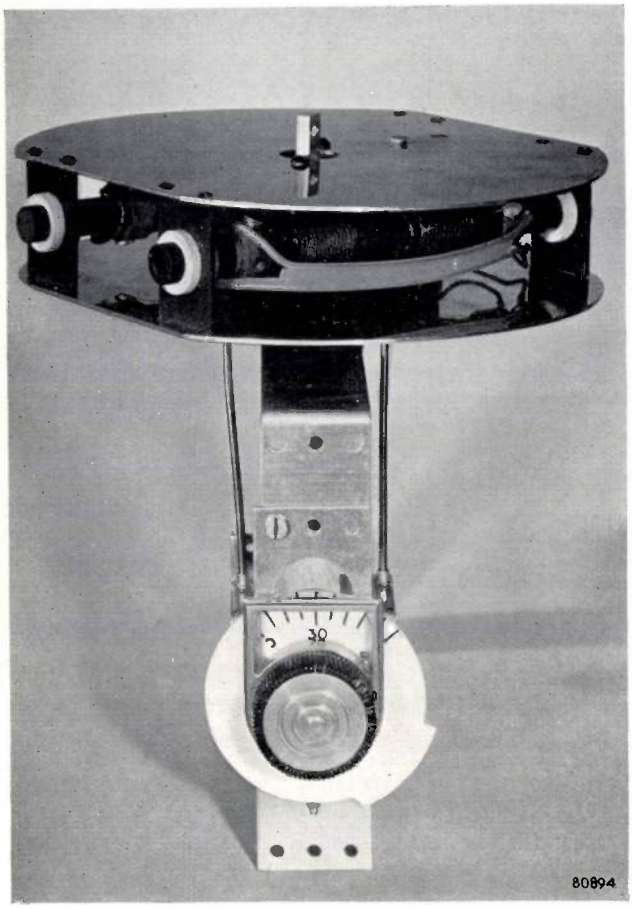


Fig. 12. Two rotatable Ferroxcube aerials (long and medium waves). Some sets (e.g. BX 632 A and 633 A) have a construction allowing the set to receive on a capacitative (outdoor) aerial in the zero position of the knob. When the knob is turned clockwise, this aerial is disconnected and the scale indicating the position of the Ferroxcube antennae lights up as a sign that reception now takes place exclusively via the Ferroxcube aerials. (In versions later than the one depicted, the scale range is more than 180°.)

ferroceptors between the chassis and the capacitive plate aerial (fig. 11); the latter is connected to the chassis during medium-wave reception. In the construction shown in fig. 12 the two rods are mounted between two earthed metal plates. Screening may be further improved by surrounding this assembly by a cage made of copper wire running parallel to the core, interwoven with nylon threads in the perpendicular direction (to avoid magnetic screening).

In contrast with the screening of ordinary frame aerials, the above-mentioned methods only slightly increase the self-capacitance, owing to the small dimensions of the coil.

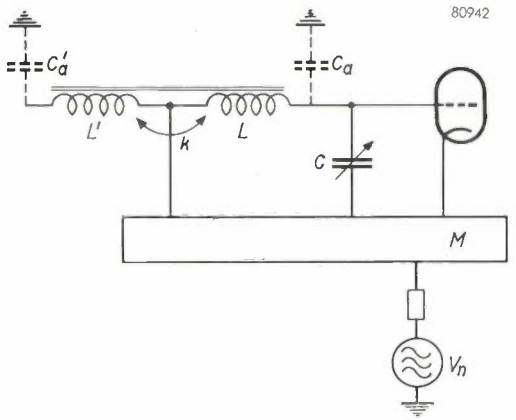


Fig. 13. Circuit for compensating the aerial-effect of a Ferroxcube aerial L. L' = compensating coil coupled with L (L_1 in fig. 11).
C_a and C_a' = capacitance of L and L' respectively with respect to the surroundings. At suitable values of C_a' and of the coupling coefficient k, no interference voltage will be produced across the tuning capacitor C (V_n = source of interference). M = chassis.

II. TECHNICAL ASPECTS OF INDUCTIVE AERIALS

This section deals with the directional effect and with three factors of particular importance in design: sensitivity, signal-to-noise ratio and effective height.

The directional effect

The signal voltage induced in a frame aerial, assuming the complete absence of capacitative reception, is proportional to $\sin\alpha$, where α is the angle between the perpendicular to the plane of the loop and the direction of the transmitter. Plotting the signal voltage in a polar diagram gives the well-known figure-of-eight curve (*fig. 14*).

At a value of $\alpha = 30°$ the voltage is reduced to half the maximum value, whilst the minima at $\alpha = 0$ are sharply defined.

In order to evaluate the directional effect, however, we are not so much concerned with the voltage

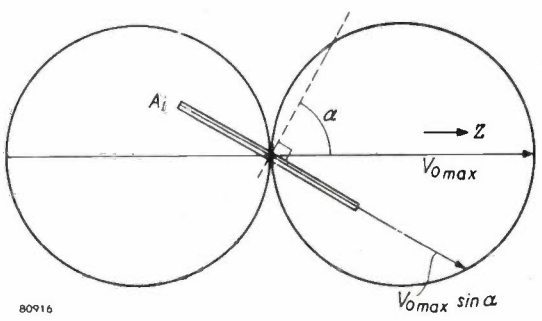


Fig. 14. The figure-of-eight polar diagram of an inductive aerial (A_i) having no capacitative component. The signal voltage in the aerial is $V_{0max} \sin\alpha$, α being the angle between the perpendicular to the plane of the loop and the direction Z of the transmitter.

induced in the frame as with the volume produced by the loudspeaker. Let us for the moment assume that the set operates without automatic volume control. We then get a nearer representation of the volume of sound by plotting the induced voltage in polar co-ordinates on a logarithmic scale. In *fig. 15*, curve *1*, the maximum sound volume ($\alpha = 90°$) is set at a level of 0 dB. At $\alpha = 30°$ the level is -6 dB, at $\alpha = 6°$ it is -20 dB.

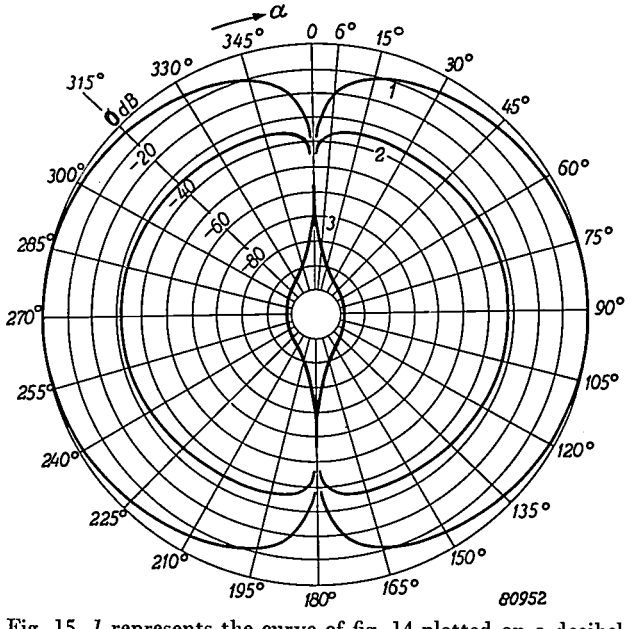


Fig. 15. *1* represents the curve of fig. 14 plotted on a decibel scale. *2* the same, taking into account the effect of the automatic volume control (cf. fig. 16). *3* level of noise and interference.

The automatic volume control should not, however, be left out of consideration. Its effect is shown in the curve in *fig. 16*, which can be regarded as representative for many modern receivers.

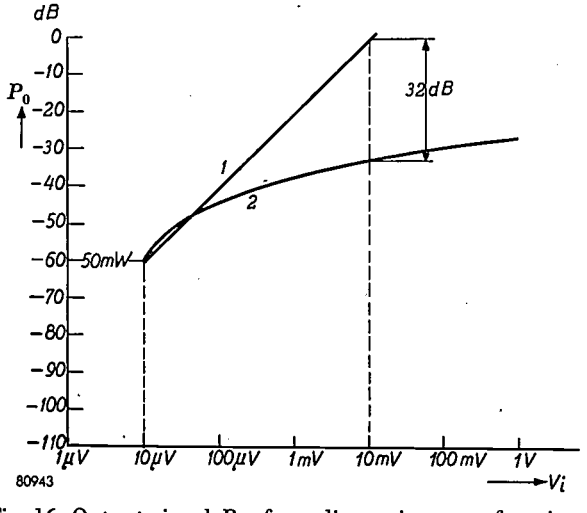


Fig. 16. Output signal P_0 of a radio receiver as a function of the input voltage V_i, *1* without a.v.c., *2* with a.v.c. At $V_i = 10$ mV the difference amounts to 32 dB. The vertical scale is made to correspond to the decibel scale of fig. 15.

Assuming a maximum input voltage ($\alpha = 90°$) of 10 mV, we see that the amplification is reduced by 32 dB by the A.V.C. At $\alpha = 30°$ the input voltage is 5 mV, i.e. 6 dB less than at $\alpha = 90°$; the reduction in amplification is then 28 dB , so that the sound level lies only $32 - (6 + 28) = 2$ dB below the level produced at $\alpha = 90°$. At $\alpha = 6°$ the difference is no more than 5 dB; cf. curve *2* in fig. 15. It will now be clear that the angle between two fixed frames (cf. Part I) can be quite small, for if one frame happens to be in the position of minimum reception ($\alpha = 0°$), the other gives a satisfactory signal if it crosses the former at even a small angle. Hence the two frames can lie along the diagonals of quite a flat cabinet: there is no need to make them smaller in order to include a larger angle.

Near the minimum, also the noise level has to be taken into account (under the term noise we include here also the continuous interference referred to in Part I). If it is assumed that in the position of the frame for maximum signal, the reception is just free of audible noise, then the noise level at the input is 66 dB lower than the signal level [10]) i.e. lower by a factor of 2000, and lies in fig. 15 at $-(66 + 32) = -98$ dB. The more α is reduced, the greater becomes the amplification by the A.V.C. and thus the absolute value of the noise increases. The result is represented by curve *3* in fig. 15. When α approaches zero the sound volume of the signal drops only slightly, but the noise increases rapidly. The best way to adjust the set for maximum reception is, therefore, by first setting the aerial for maximum noise (i.e. minimum reception) and then turning it through 90°.

Sensitivity and signal-to-noise factor

The sensitivity of a frame aerial may be defined as the signal voltage supplied by the aerial circuit to the grid of the first tube, per unit electric field strength E of the local radiation field. This grid voltage is Q times the voltage induced in the aerial (Q being the quality factor of the aerial circuit); the induced voltage is given by hE, according to the definition of the "effective height" (h) of an inductive aerial [11]).

[10]) M. Ziegler, Noise in receiving sets, Philips tech. Rev. **3**, 189-196, 1938, in particular, page 196.

[11]) The effective height of an inductive aerial has obviously no concrete physical meaning, but is only a quantity introduced for mathematical reasons. It is defined as the effective height of a capacitative aerial, which, placed in the same field of radiation, would produce the same signal voltage V_0 as the inductive aerial when oriented for strongest reception. The definition of the effective height of a capacitative aerial as $h = V_0/E$ is readily acceptable.

For the sensitivity S we find, therefore:

$$S = QhE/E = Qh. \quad . \quad . \quad . \quad . \quad (1)$$

We define the signal-to-noise *factor* N as the variable term in the signal-to-noise *ratio*, viz. the ratio of the signal power in the frame circuit for unit E, to the noise power in this circuit. It can be shown (see below, small print) that N is given by

$$N = Qh^2. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

If the field E is of unit strength, then the signal voltage V_s across the tuning capacitor is by definition S and hence

$$V_s = Qh.$$

If we introduce a parallel resistance R to account for the absorption of the signal, then this resistance will dissipate a signal power P_s given by

$$P_s = \frac{V_s^2}{R} = \frac{Q^2h^2}{R}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

The resistance R is responsible for the noise:

$$\overline{V_n^2} = 4KT \cdot \Delta f \cdot R,$$

in which $\sqrt{\overline{V_n^2}}$ = r.m.s. value of the noise voltage, K = Boltmann's constant, T = absolute temperature of the resistance R and Δf = frequency band under consideration. The noise power P_n is then:

$$P_n = \frac{\overline{V_n^2}}{R}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

From (3) and (4) we obtain for the signal-to-noise ratio:

$$\frac{P_s}{P_n} = \frac{V_s^2}{\overline{V_n^2}} = \frac{Q^2h^2}{4KT \cdot \Delta f \cdot R}.$$

If L is the self-inductance of the circuit and $\omega = 2\pi \times$ the resonance frequency, then $Q = R/\omega L$ and $Q\omega L$ may be substituted for R:

$$\frac{P_s}{P_n} = \frac{1}{4KT \cdot \Delta f \cdot \omega L} Qh^2.$$

In the case considered here, $4KT\Delta f.\omega L$ may be considered as a constant and Qh^2 is, therefore, the term which effectively determines the signal-to-noise ratio.

Both the quality factor and the induced voltage per unit of field strength can be calculated as well as measured. For a given design, therefore, both the sensitivity and the signal-to-noise factor can be calculated.

As mentioned above, Philips are currently using two types of inductive aerials: the frame aerial of low impedance (few turns) and the Ferroxcube antenna or ferroceptor. We shall now separately examine the two types with regard to sensitivity and signal-to-noise factor.

The low-impedance frame aerial

Fig. 17a shows a frame aerial of one or a few turns. It is coupled, via h.f. transformer T, to tuning capacitor C, which in turn is connected to the first amplifying tube. Self-inductances are designated by L and resistances by r; the indices $_{0, 1, 2}$, refer to the frame, the primary and the secondary coil respectively. For this diagram we may substitute the equivalent circuit of fig. 17*b*. Assuming that the transformer is ideal, having no leakage (coupling factor $k = 1$) and no losses, and that the self-inductances L_1 and L_2 are infinite (the ratio L_2/L_1, however, having a finite value), then

$$V = V_0 \sqrt{\frac{L_2}{L_1}}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

where V represents the e.m.f. of the equivalent generator (fig. 17*b*) and V_0 is the e.m.f. induced in the frame (fig. 17*a*). Assuming no transformer losses we may put:

$$Q = Q_0 = \frac{\omega L_0}{r_0}. \quad . \quad . \quad . \quad . \quad (6)$$

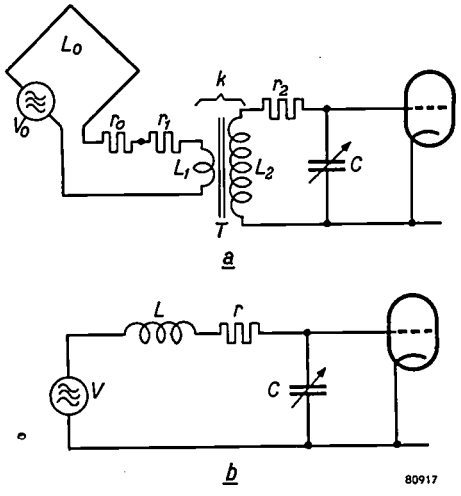


Fig. 17. *a*) L_0 frame aerial with low impedance, coupled to tuning capacitor C via an h.f. transformer T. The voltage induced in the frame aerial is V_0. The resistance of the frame, of the primary coil (self-inductance L_1) and the secondary coil (self-inductance L_2) are r_0, r_1 and r_2 respectively. The coupling coefficient is k.
b) Equivalent circuit of (*a*).

For the tuning, it is not the self-inductance L_0 of the frame but the "effective" self-inductance L (fig. 17*b*) which is the determining factor, since it is the latter in combination with C, which defines the frequency of resonance. It is thus more logical to take into account the effective height h of frame and h.f. transformer together than that of the aerial alone. We will define the effective height h of the combination as the value of V (not V_0) per unit field strength. For $E = 1$, equation (5) may be written as

$$h = h_0 \sqrt{\frac{L_2}{L_1}}. \quad . \quad . \quad . \quad . \quad (5a)$$

(h_0 is the effective height of the frame alone, i.e. the value of V_0 at $E = 1$).

As, however, the transformer is not ideal, the formulae actually applicable are more involved, viz.:

$$h = \frac{k}{1+\frac{L_0}{L_1}} h_0 \sqrt{\frac{L_2}{L_1}}, \quad \ldots \quad (5b)$$

and

$$Q = Q_2 \frac{\left(1+\frac{L_0}{L_1} - k^2\right)\left(1+\frac{L_0}{L_1}\right)}{\left(1+\frac{L_0}{L_1}\right)^2 + k^2\left(\frac{Q_2}{Q_1}+\frac{Q_2}{Q_0}\cdot\frac{L_0}{L_1}\right)} \quad \ldots \quad (6b)$$

Here Q_1 and Q_2 are the quality factors of the primary and the secondary respectively of the transformer.

The formulae (5b) and (6b) determine h and Q, the two factors giving the sensitivity S ($= Qh$) and the signal-to-noise factor N ($= Qh^2$). It is generally desirable that both S and N are large. The formulae show that h and Q, at given values of Q_0, Q_1 and Q_2, are dependent on only two quantities: the coupling factor k and the matching ratio L_0/L_1. *As regards sensitivity and signal-to-noise factor, it is therefore immaterial how many turns the frame has.* The number of turns may therefore be selected according to other criteria, e.g. the sensitivity to local interference and constructional considerations. As discussed in Part I, the sensitivity to local interference is proportional to the number of turns. For this reason the number is chosen as small as possible, preferably one, which is also convenient from the constructional point of view.

Further examination of (5b) and (6b) shows that if k is increased or the value of L_0/L_1 is reduced, h becomes greater, but Q smaller. Hence it may be expected that the Qh and Qh^2 as functions of k and of L_0/L, will show one or more stationary values. Analyzing Qh shows that it does indeed have a maximum at $k = 1$ and L_0/L_1 somewhere between 3 and 5, dependent on the various Q-values. Similarly Qh^2 is found to have a maximum at $k = 1$ and L_0/L_1 between 1.5 and 2.5 (likewise dependent on the Q-values). The value of k^2 can drop to approx. 0.7 before the maximum shows a substantial drop. *Fig. 18* shows Qh and Qh^2 as functions of L_0/L_1 for certain practical values of k, Q_0, Q_1 and Q_2. Since the maxima of the two curves do not occur at the same value of L_0/L_1, a suitable compromise value has to be sought. In this connection the following considerations should be taken into account. If the set itself contains a rather strong source of noise, there is little point in designing a frame aerial with the greatest possible signal-to-noise factor (Qh^2); in that case it is better to strive for maximum sensitivity (Qh). For a receiver having very little inherent noise it is preferable to design a frame aerial producing the least possible noise, if necessary by sacrificing a little of the sensitivity.

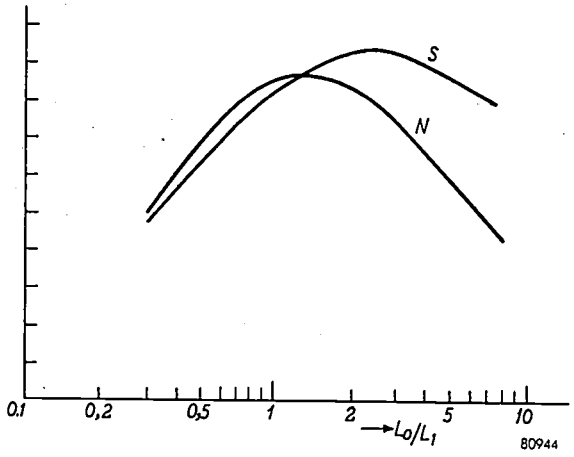


Fig. 18. Sensitivity $S = Qh$ and signal-to-noise factor $N = Qh^2$ plotted as functions of L_0/L_1, for $k^2 = 0.70$ and $Q_0 = Q_1 = Q_2$.

The Ferroxcube antenna (ferroceptor)

Consider a Ferroxcube rod (length l, diameter d, *fig. 19*), on which is wound a flat coil (width $a<d$; at a distance x from the middle). Fig. 19 also shows a measured curve of the effective relative permeability $\mu_x = B_x/H$, plotted as a function of x (B_x being the magnetic induction at position x due to the transmitter field H). For fairly small values of l/d this curve approximates to a parabola, which can be represented by

$$\mu_x = \mu_{\text{eff}}\left(1 - 3.6\,\frac{x^2}{l^2}\right).$$

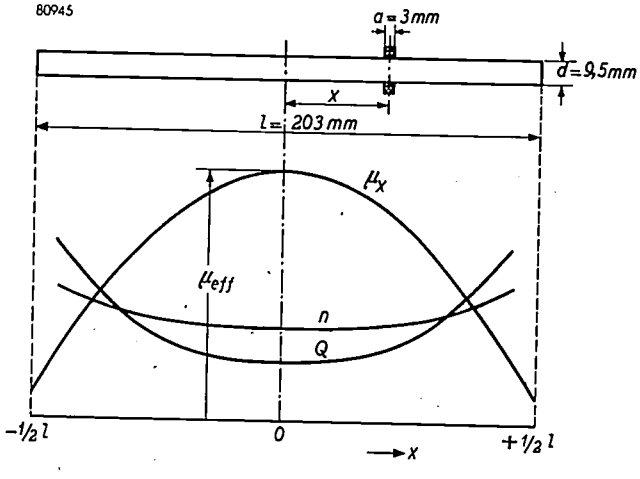


Fig. 19. Above: Ferroxcube rod of length l with flat coil, placed at a distance x from the mid-point. The curves represent μ_x, Q and n as functions of x, for a rod of Ferroxcube IV-B of the dimensions shown, wound with a coil of width $a = 3$ mm and a self-inductance of 200 μH at 1 Mc/s.

μ_{eff}, the maximum value of μ_x, determines the extent to which the Ferroxcube concentrates the transmitter field in the core. μ_{eff} depends on l/d and on the toroid permeability μ_{tor} of the material (i.e. the relative permeability measured on a closed toroid, see *fig. 20* [12]).

Fig. 19 also shows the number of turns n required for a coil of given self-inductance, and the quality factor Q of such a coil at a given frequency.

For the effective height h, the formula

$$h = \frac{2\pi\mu_x nA}{\lambda}, \quad \ldots\ldots \quad (7)$$

applies, A being the area of a turn and λ the wavelength.

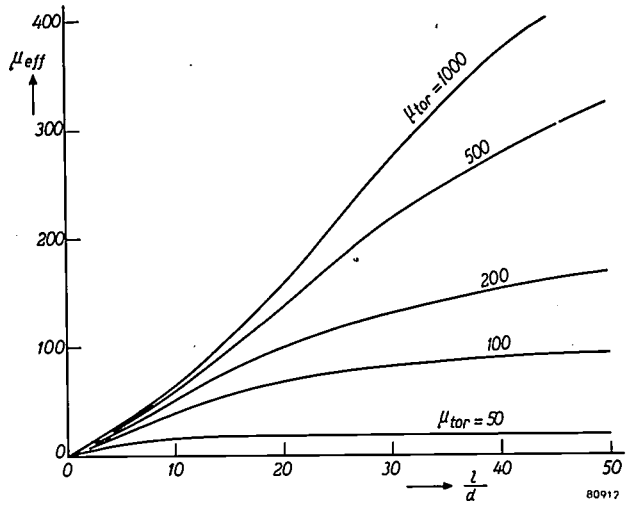


Fig. 20. μ_{eff} (the maximum value of μ_x) as a function of l/d, for various values of μ_{tor}.

Equation (7) is derived as follows: The transmitter field H induces in the coil an e.m.f. given by:

$$V = nA\omega B_x = nA\omega\mu_x\mu_0 H \quad \ldots\ldots \quad (8)$$

(μ_0 = permeability of free space). Now, ω may be written as $2\pi c/\lambda$, and in a radiation field $H = E/\mu_0 c$. Substituting, (8) becomes

$$V = \frac{2\pi\mu_x nA}{\lambda} E. \quad \ldots\ldots \quad (9)$$

By definition $h = V/E$. Substituting hE for V in (9) we arrive at (7), the required result.

Using (7), we find for the sensitivity:

$$S = Qh = \frac{2\pi}{\lambda}\mu_x nQ$$

and for the signal-to-noise factor:

$$N = Qh^2 = \frac{4\pi^2}{\lambda^2}\mu_x^2 n^2 Q.$$

[12]) For further details see H. van Suchtelen, Ferroxcube aerial rods, Electronic Applications Bulletin **13**, 88-100, 1952.

where μ_x, n and Q are the functions of x shown in fig. 19. With the help off these curves we can plot S and N as functions of x. This is shown in *fig. 21*,

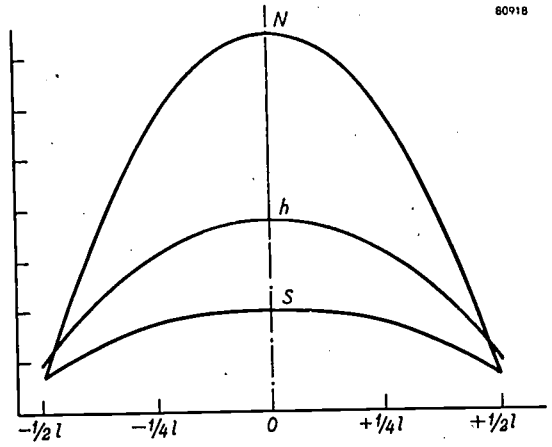


Fig. 21. S, N and h as functions of x for a narrow coil (cf fig. 19).

which also shows the curve for h. It can be seen that in the middle of the rod the sensitivity has a very flat maximum, whilst the signal-to-noise factor shows a fairly sharp maximum. This applies to a rod with a flat coil. As a further variable the coil *width* a may now be introduced. Both graphical analysis and actual experiments indicate that the optimum design is a coil of width $l/2$ situated in the middle of the rod (*fig. 22*). In practice, however, this construction cannot always be used for the following reasons:

1) The self-inductance should be adjustable, but near the middle of the rod a shifting of the coil has little effect. If one can indulge in the luxury of two Ferroxcube rods (one for long and one for medium waves), then either a small trimming coil may be added at the end of the rod, or the coil may be divided into two parts, situated a small distance apart on either side of the middle of the rod; the self-inductance is then adjustable by varying the degree of coupling between the two parts of the coil. e.g. by changing their separation.
2) In some cases considerations of price make it necessary to combine the medium and the long wave coil on a single rod (*fig. 23*). This presents some difficulties for the reception of medium

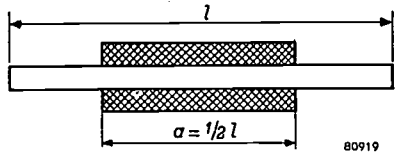


Fig. 22. The optimum shape for a ferroceptor coil is a "width" a equal to half the length l of the ferroxcube rod.

waves. If the long-wave coil remains open-circuited, its self-capacitance constitutes an absorption circuit for medium waves and thus introduces certain losses. There are two ways for avoiding this drawback:

a) The long-wave coil is short-circuited during the reception of medium waves. This has the disadvantage that a length *b* of the rod (*fig. 23*) is rendered useless, since the short-circuited coil prevents a field occurring in part *b*, particularly if the coil has a high quality factor. For the remaining part of the Ferroxcube rod the value of l/d is smaller, and consequently μ_{eff} and h are also smaller (see next section).

b) The long-wave coil is connected in parallel with the medium-wave coil during the reception of medium waves. The apparent shortening of the rod does not then occur (the effective height of the combination is even slightly greater than that of the medium wave coil alone). The fact, however, that parallel connection increases the self-capacitance considerably, often proves an insuperable drawback, and one has to resort to short-circuiting the long-wave coil as mentioned in (*a*).

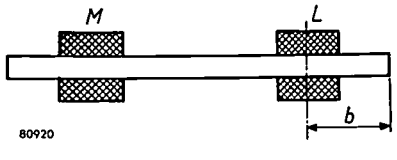


Fig. 23. Medium wave coil M and long wave coil L wound on a single rod of Ferroxcube. When, for medium wave reception, the long wave coil is short-circuited, part b of the rod becomes ineffective.

Effective height and rod dimensions of a Ferroxcube antenna

The effective height of a coil situated in the middle of the rod is proportional to $l\sqrt{d}$. This can be derived as follows.

The self-inductance L of the coil is proportional to n^2d, i.e.

$$n \propto \sqrt{L/d}. \quad \ldots\ldots\ldots \quad (10)$$

According to fig. 20, for not too large values of l/d, we have

$$\mu_{eff} \propto l/d. \quad \ldots\ldots\ldots \quad (11)$$

By applying (10) and (11) to equation (8), and putting $\mu_x = \mu_{eff}$ (since the coil is in the middle of the rod), we find that the e.m.f. induced in the coil, and hence also the effective height, is proportional to $l\sqrt{d}$.

From the fact that h is proportional to $l\sqrt{d}$ it is clear that the length l of the rod should be chosen as large as possible. When this length has been decided upon (as a rule determined by the width of the cabinet, or, for rotatable rods, by its depth), the thickness of the rod must be decided. Considering a constant length, the fact that h is proportional to $\sqrt{d}$ means that h increases only as the fourth root of the volume: in order to double h by using a thicker rod, 16 times as much Ferroxcube is required!

It will be obvious that the ratio l/d should be chosen as large as possible, having regard, however, to the magnetic and mechanical properties of the material. For Ferroxcube IV-D, e.g., with $\mu_{tor} = 40$, there is no point in choosing l/d greater than 15, since at greater values there is hardly any increase of μ_{eff} (fig. 20). For the materials with greater toroid permeability (Ferroxcube IV-B, $\mu_{tor} = 200$ and Ferroxcube III-B, $\mu_{tor} = 700$) l/d cannot be raised much above 25 in view of their brittleness.

Other points to consider are the core losses and the temperature coefficient of the permeability. As regards the core losses, Ferroxcube IV-B can be used up to higher frequencies than Ferroxcube III-B, but the latter has a smaller temperature coefficient. For rods exclusively intended for medium wave reception Ferroxcube IV-B is generally used, and for long wave reception mostly Ferroxcube III-B. If medium and long-wave coils are to be situated on a single rod (fig. 23), then either the whole rod is made of Ferroxcube IV-B, or the medium wave part is made of Ferroxcube IV-B and the long wave part of Ferrocxube III-B. The latter construction is shown in fig. 10; the Ferroxcube III-B part, whose permeability has the smallest temperature coefficient, is situated above the tubes, where most heat is generated.

Finally a commonly applied method for obtaining a greater effective height for a given volume of Ferroxcube may be mentioned. This consists of using a number p of parallel, identical Ferroxcube rods, each provided with a small coil (*fig. 24*). The coils are connected in series and the rods are spaced so far apart that there is little mutual interaction

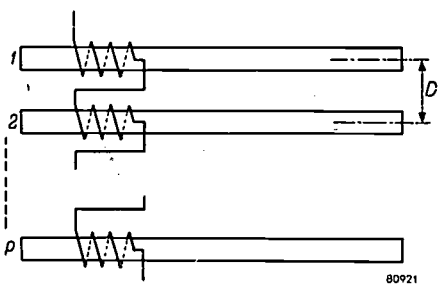


Fig. 24. Ferroxcube aerial consisting of p parallel Ferroxcube rods, each wound with one coil. The coils are connected in series.

between the coils (separation $D \geqslant d\sqrt{\mu_{\text{eff}}}$, e.g. 8 cm).

For a total self-inductance L, the self-inductance per coil must be L/p. If n_1 turns are required for one coil with a self-inductance L, then for p coils with the same total inductance $p^{-\frac{1}{2}}n_1$ turns are required per coil. In total this corresponds to $p.p^{-\frac{1}{2}}n_1 = p^{\frac{1}{2}}n_1$ turns linking the field. The voltage induced is therefore p times as great as in the case of a single rod (n_1 turns). Since the effective height h is defined as the voltage induced per unit field, we have $h \propto p^{\frac{1}{2}}$. The volume of Ferroxcube used is proportional to p, the number of rods used (l and d being constants); hence the effective height is proportional to the *square root* of this volume. This is more favourable than the *fourth root* relationship applying to a single rod. The application of a number of parallel rods thus produces a greater effective height or a saving in Ferroxcube. For practical reasons more than two rods are seldom used.

Summary. As early as 1939 Philips marketed a radio receiver (type 902 A) equipped with a single-loop frame aerial. This design was based on the fact that an inductive aerial is less sensitive to local interference than a capacitative aerial, and that the aerial-effect of an inductive aerial is diminished by reducing the number of turns. Until the introduction of Ferroxcube, however, frame aerials with few turns were not used on a large scale. Ferroxcube enables the dimensions of the h.f. transformer coupling the aerial loop to the tuning capacitor to be drastically reduced. A number of Philips sets employ two such frame aerials, set at a small angle to each other, for both medium and long wave reception; by means of a switch, the loop giving the most satisfactory reception of a certain station can be selected. Short waves are received by a capacitative plate aerial. Another form of inductive aerial is the Ferroxcube antenna or ferroceptor, consisting of a coil wound on a Ferroxcube rod. Various designs are dealt with. Part I of this article gives a brief history and a general description of these inductive aerials. The second part of the article examines in closer detail the directional effect, the sensitivity, the signal-to-noise factor and the effective height.

THE DIFFERENT TELEVISION STANDARDS CONSIDERED FROM THE POINT OF VIEW OF RECEIVER DESIGN

by W. WERNER.

621.397.62

The diversity of European television standards has been referred to more than once in this Review. From the point of view of those concerned with T.V. transmission, the most important consequence of this lack of uniformity is that it necessitates the use of "line converters" in any international exchange of programmes. The receiver designer, on the other hand, is concerned with the demand created in certain regions for sets able to operate in accordance with two or more different standards. Examples of such regions are the Saar, the districts along the Franco-German border, and, above all, Belgium. In Belgium two different standards are employed, one for Walloon, and one for Flemish transmissions; moreover, both these standards differ from those of the surrounding countries, i.e. France on the one side and Holland and Germany on the other.

As an introduction to an article on a four-standard receiver specially designed for Belgium, to be published in due course, the different regional standards and a number of associated problems will now be discussed.

A number of different standards for black and white (or "monochromatic") television are at present in force. Several European countries (including the Netherlands and West Germany), and also one or two countries outside Europe, have adopted what is known as the Gerber standard [1]), a 625-line system. Before this standard was accepted by the Comité Consultatif International des Radiocommunications (C.C.I.R.) (October 1950), an experimental transmitter designed by Philips to a standard having much in common with the American system was operated at Eindhoven. (The Philips transmitter differed in respect of the number of lines (567) and the numbers of complete pictures per second (25); the American system uses 525 lines and 30 pictures per second.)

Articles describing a TV receiver for the 567-line system [2]) appeared in this Review, and in another publication, in 1948.

Table I. Characteristics of the principal television standards now in force. N = number of lines per picture, f_r = frame frequency ($^1/_2 f_r$ = number of complete pictures per second), Δf = width of channel, f_s = frequency of sound carrier, f_v = frequency of vision carrier, A.M. = amplitude modulation, F.M. = frequency modulation.

Standard	N	$^1/_2 f_r$ c/s	Δf Mc/s	$f_s - f_v$ Mc/s	Pict. mod.	Sound mod.
U.S.A.	525	30	6	4.5	neg.	F.M.
Great Britain	405	25	5	3.5	pos.	A.M.
France	819	25	13.15	11.15	pos.	A.M.
"Gerber" (C.C.I.R.)	625	25	7	5.5	neg.	F.M.
Belgium	819	25	7	5.5	pos.	A.M.
	625	25	7	5.5	pos.	A.M.

[1]) So named after the Chairman of the C.C.I.R. sub-committee which proposed this standard; see Standards for the international 625-line black and white television system, C.C.I.R. Geneva, 10 October 1950.

In *Table I* some characteristics of the major television standards are listed. Certain consequences of these characteristics which are of particular interest to the designer of TV receivers will now be examined. The channel-width, method of picture-modulation, and system of sound-modulation will be discussed.

Channel-width

The following features are common to all the different standards.

a) Amplitude modulation and vestigial side-band transmission are employed for the video signal (*fig. 1*). *b*) The ratio between width and height of the picture is 4 : 3. *c*) Scanning is interlaced in an odd-even pattern. *d*) The number of complete pictures per second is 25 (except in the American system, where it is 30, so as to enable the frame frequency to be synchronized with the mains frequency, which is 60 c/s in America).

It will be seen from table I that in general the channel-width increases with the number of lines. The necessity for this may be seen as follows. The greater the number of lines employed, the higher the vertical definition of the picture. Logically, then, the horizontal definition should increase in the same proportion. Accordingly, the number of elements into which the picture is resolved during scanning increases as the square of the number of lines. The information contained in the video signal is proportional to the number of picture elements, and

[2]) P. M. van Alphen, J. de Gier, J. Haantjes, F. Kerkhof, H. Rinia and G. J. Siezen, Home projection television, Proc. Inst. Rad. Engrs. **36**, 395-411, 1948; Projection-television receiver, Philips tech. Rev. **10**, 69-78, 97-104, 125-134, 307-317 and 364-370, 1948/49.

the bandwidth should be proportional to this information.

However, the table shows that this general rule is not invariably followed. Whereas the French standard of 819 lines specifies a channel-width of 13.15 Mc/s, according to the Belgian standard for Walloon transmissions (likewise 819 lines) 7 Mc/s is

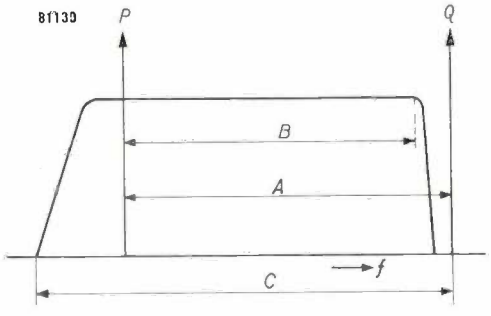


Fig. 1. Frequency spectrum of a television transmitter with asymmetrical side band ("vestigial side band"). f frequency, P vision carrier, Q sound carrier, A frequency-interval between the two, B bandwidth of complete side band transmitted, C channel width.

sufficient. Thus, in this Belgian system, some of the picture information is sacrificed for the sake of economy of channel-width.

As will be seen from fig. 1, and from the table, the frequency-interval between the vision and sound carriers should increase with the bandwidth.

systems may best be suppressed in the receiver. Such interference is in the form of pulses which, in unfavourable circumstances, may attain an amplitude far exceeding the peak value of the carrier. Interfering pulses of this type are shown at the points i in fig. 2*a* and 2*b*.

Visible interference in the form of spots

Pulse-shaped interfering signals appear on the picture as spots, which are mainly white when positive modulation is employed, and mainly black with negative modulation.

The luminance of the white spots associated with positive modulation may greatly exceed that of the brightest parts of the picture itself. If the particular interfering pulses are strong enough to drive the picture tube into grid current, the electron beam will fail to focus sharply, and the white spots will grow into bright discs. To avoid this, receivers for positive modulation are often fitted with limiters; such limiters are included in the circuit preceding the picture tube, to cut off any interference above a given level. In some cases, the cut-off level is variable.

As mentioned above, in the case of negative modulation, the spots produced by interfering pulses are mainly black. Occasional white spots also occur, but, provided that the "white" corresponds to

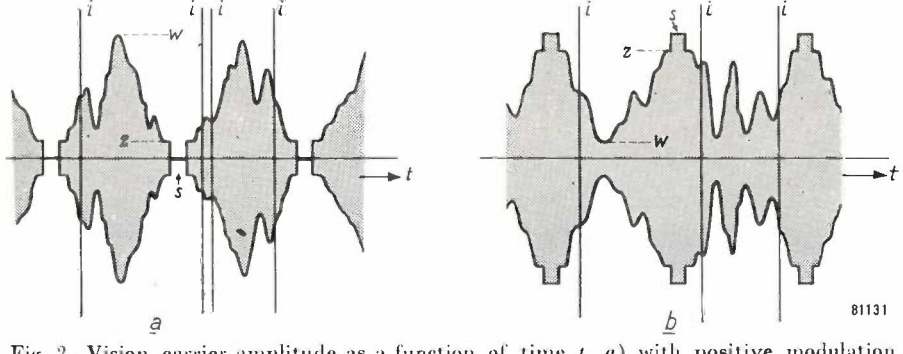


Fig. 2. Vision carrier amplitude as a function of time t, *a*) with positive modulation, *b*) with negative modulation. w level corresponding to white in the picture, z black level, s synchronizing pulses, i interference pulses.

Video modulation system

In principle, the amplitude of the vision carrier can be varied by the video signal in two ways, i.e. by positive or negative modulation (*figures 2a* and *2b*, respectively). Peak R.F. amplitude corresponds to "white" in the picture in the case of positive modulation and to "blacker-than-black" in that of negative modulation.

The direction of the modulation determines the measures by which interference from such sources as sparking commutator motors and car ignition

nearly zero modulation of the carrier, they cannot become very much brighter than the brightest parts of the picture itself (fig. 2*b*).

Effect of interference on synchronization

The horizontal and vertical deflection of the electron beam is synchronized by special signals included in the envelope of the transmitted television signal. Line synchronizing signals, for example, are rectangular pulses (fig. 2), whose steep leading edges initiate the flyback of the beam.

The interfering signals are likewise steep-sided; hence, of course, they are quite capable of disorganizing the synchronization completely, and thus mutilating the picture beyond recognition. Experience has shown that this is far more likely to happen when negative, than when positive, modulation is employed, owing to the fact that in negative modulation the interference pulses are predominantly in the same direction as the sync. signals, whereas in positive modulation it is the opposite. Accordingly, measures to minimize the effect of interfering pulses in the sync. circuit are particularly necessary in receivers designed for negative modulation, where the amplitude of such pulses may greatly exceed that of the sync. pulses. Such measures include the use of:

a) a so-called "flywheel circuit" for horizontal deflection;

b) an integrating circuit for vertical deflection;

c) a noise-inverter circuit, to act upon the synchronizing signals before they are separated.

The flywheel method of synchronization has been described in detail in an earlier issue of this Review [3]) and need not be discussed here. However, it is worth mentioning that in some cases this method is also employed in conjunction with positive modulation, for the following reason. A weak incoming signal is invariably associated with a relatively high noise level, which makes the edges of the sync. pulses irregular and vague. This upsets the timing of the electron beam in the picture tube, so that it starts to scan some of the lines at the wrong moment. The result is a horizontal displacement of the scanning lines relative to one another and "tearing" at the vertical sides of the picture, and at the vertical edges of individual objects in the picture. This "frayed" effect can be avoided by employing a flywheel circuit for horizontal deflection.

For the vertical deflection, an integrating circuit inserted between the sync. pulse separator and the saw-tooth generator is recommended, because pulses of short duration contribute virtually nothing to the output voltage of such a network, and are therefore unlikely to affect the frame synchronization.

In TV transmitters with negative modulation, the frame-synchronizing signal is preceded by so-called equalizing pulses [4]), whose function is to help to maintain the interlacing when an integrating circuit is used. Equalizing pulses are not transmitted in the British system (positive modulation), but integrating networks are nevertheless used in some British-made receivers to stabilize the frame synchronization during periods of heavy interference.

[3]) P. A. Neeteson, Philips tech. Rev. **13**, 312-322, 1951/52.

[4]) See fig. 1 in the article referred to in note [3]).

As already mentioned, the synchronization of negative modulation systems is in principle more sensitive to interference than that of systems in which positive modulation is employed. However, it is possible to remove this disadvantage of negative modulation by providing the receiver with a noise-inverter circuit; this changes the sign of those interfering pulses which extend above a certain critical signal-level. After passing through the video amplifier, the signal voltage is fed both to the picture tube direct and to the sync. separator via the noise inverter. Hence the signal and the interference pulses must pass through the latter to reach the sync. separator.

The above-mentioned critical level is established as accurately as possible at a value just above the peaks of the sync. signals. As regards synchronization, the negative modulation system then has all the favourable characteristics of positive modulation.

The amplified video signal is, of course, applied direct to the picture tube, by-passing the inverter circuit, so as to preserve what is really the most favourable feature of negative modulation, i.e. that the spots produced in the picture by interference pulses are mainly dark and quite small.

Effect of interference on automatic gain control

Television sets designed to receive signals on more than one channel are preferably equipped with A.G.C., that is, a system controlling the amplification of the receiver in such a way that the strength of the output signal is virtually unaffected by variations in the strength of the input signal.

The most suitable measure of signal strength is a particular level of the television signal not governed by the gradation of the picture. In the case of negative modulation, then, the peak of the synchronizing pulses is the obvious choice. In principle, a D.C. voltage extracted from the input signal of the video detector by means of a simple peak-voltage rectifier could be employed as a control voltage for the vision amplifier, but in practice this simple arrangement is rendered completely ineffective by strong interference, owing to the fact that the rectifier then responds to the relatively higher peaks of the interfering pulses rather than to those of the sync. pulses. This causes an undue decrease in amplification, which may even be sufficient to fade out the picture altogether. To avoid this, a gate valve is included in some circuits. The control grid of this valve is so biased as to pass current only during the sync. pulses. Any interference occurring in the intervals between sync. pulses is then entirely

innocuous. Provided that the rectifier has a fairly low time constant (i.e. smaller than approx. 5 × the line period), the effect on the control voltage of any interference happening to coincide with the sync. pulses will be negligible.

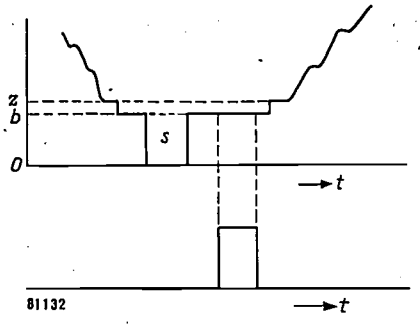


Fig. 3. *Above*: Vision signal with positive modulation, in the region of a line-synchronizing pulse (*s*). *z* black level, *b* blanking level.
Below: Keying pulse to operate the A.G.C. valve within the period of the blanking signal.

In the case of positive modulation, the level of the blanking signal (*fig. 3*), that is, the signal immediately preceding and following each sync. pulse to conceal the flyback of the scanning spot in the picture tube, may be employed as a reference level independent of picture gradation. The control voltage is extracted from this level by means of a gate valve operated intermittently by the successive line-synchronizing pulses. Since these pulses precede the blanking signals, they must be displaced slightly in time by means of a delay network. As in the case of negative modulation, interference other than that coinciding with the keying pulses cannot affect the control voltage.

Sound modulation

Frequency-modulated systems

Frequency modulation of the sound signal in a television system offers certain advantages as compared with amplitude modulation. Firstly, we have the well-known advantage of frequency modulation in general, that is, the relatively small amount of noise and interference involved. To this may be added, in the case of television, that frequency modulation enables the receiver to be so designed that the tuning is relatively less critical, so that a certain amount of frequency drift in the local oscillator is allowable and that microphony of the oscillator is inaudible.

The particular television system employed to ensure these advantages is known as the "intercarrier sound" system; the principle of this system may be explained with the aid of the block diagram shown in *fig. 4*. As will be seen from this diagram, the sound signal in the stages up to and including the video detector (D_{vid}) is amplified by the same R.F. and I.F. amplifiers as the vision signal. The mixing of these two signals produces at the output of the video detector a signal — the "intercarrier signal" — whose average frequency f_i is equal to the difference between the frequencies of the sound and vision carriers, that is, 5.5 Mc/s according to the Gerber standard, and 4.5 Mc/s according to the American system (Table I).

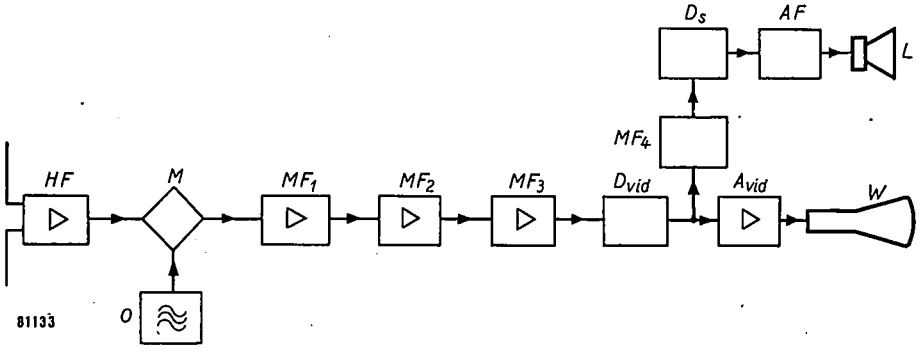


Fig. 4. Block diagram of a television receiver incorporating intercarrier sound. *HF* radio-frequency amplifier, *M* mixer stage, *O* local oscillator, MF_1, MF_2 and MF_3 first, second and third intermediate-frequency amplifier stages for vision and sound; D_{vid} video detector, A_{vid} video amplifier, *W* picture tube; MF_4 intermediate-frequency amplifier stage of sound channel (I.F. 5.5 Mc/s), D_s frequency detector of sound channel, *AF* audio-frequency amplifier, *L* loudspeaker.

This intercarrier signal varies in frequency with the sound modulation and in amplitude with the video modulation, since the sound and video signals applied simultaneously to the vision detector are respectively frequency-modulated, and amplitude-modulated.

The intercarrier signal passes via an amplifying stage (MF_4) to a frequency detector (D_s), which produces the audio signal and at the same time suppresses the (unwanted) amplitude modulation. A suitable rejection filter in the video amplifier (A_{vid}) prevents the intercarrier signal from reaching the picture tube and so interfering with the picture.

Television receivers are invariably tuned entirely by sound, this being far more critical than tuning to the vision signal since the sound channel covers a much narrower band. Receivers without intercarrier sound are especially critical in this respect by reason of the fact that the difference of the local oscillator frequency and the average frequency of the sound carrier must lie within the relatively narrow band (about 100 kc/s) covered by the particular I.F. amplifier in the sound channel. In receivers with intercarrier sound, on the other hand, the average frequency of the intercarrier signal (f_i) is fixed (5.5 or 4.5 Mc/s), regardless of the oscillator frequency; hence it is possible to detune the oscillator appreciably (e.g. 500 kc/s) without losing the sound. Naturally, such a receiver exhibits a similar insensitivity to deviations from the correct oscillator frequency arising from causes other than deliberate detuning, e.g. frequency drift produced by temperature variations in the local oscillator, or frequency modulation by microphony in the oscillator valve.

Against the above-mentioned advantages of intercarrier detection we must set certain disadvantages, some of which, however, can be avoided. For example, if the percentage modulation of the vision carrier is very high, the amplitude of this carrier corresponding to white in the picture will be very low; the same applies to the amplitude of the intercarrier signal. In the extreme case, i.e. 100% modulation, the intercarrier signal disappears during the scanning of white areas; hence the frequency detector has no signal to detect and the sound is temporarily interrupted. This interruption causes a highly irritating buzz (if the entire picture area be white, the television signal in the case of 100% modulation will consist solely of sync. pulses, and the frequency of the buzz will be the same as that of the frame-sync. pulses). It is prescribed, however, that the amplitude of the vision carrier may in no circumstances be less than a given fraction, e.g. 10%, of the maximum amplitude occurring during the sync. pulses (fig. 2*b*). Hence the above-mentioned buzzing noise can be avoided provided that the transmitting station complies with this standard, and that the receiver satisfies the two conditions which will now be defined. Firstly, the receiver must be so designed that even at the maximum depth of modulation the intercarrier signal is strong enough to ensure proper operation of the frequency detector. Secondly, the selectivity of that part of the receiver preceding the video detector must be such as to ensure that the amplification of the sound signal will invariably remain roughly 10 times (that is, about 20 dB) lower than that of the vision signal (*fig. 5*). (The

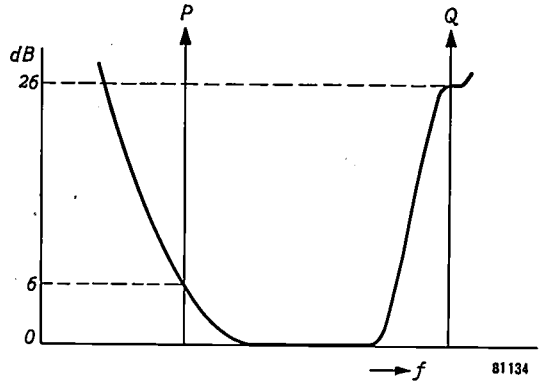


Fig. 5. Selectivity characteristic of a television receiver incorporating intercarrier sound. The input signal required to produce a given output signal is plotted logarithmically as a function of the frequency. *P* vision carrier, *Q* sound carrier. The amplification in the sound channel should be about 20 dB lower than that in the vision channel.

ratio required also depends on the strength-ratio of the vision and sound signals at the aerial terminals of the receiver, and on the extent to which the amplitude modulation is suppressed by the frequency detector.)

Unlike the above-mentioned intercarrier buzz, there is another disadvantage of intercarrier sound which cannot be avoided. This is the fact that failure of the picture transmitter is invariably accompanied by the total elimination of sound, so that any announcement concerning such a failure broadcast from the transmitting station is not heard by viewers whose sets are equipped for intercarrier detection. However, this is a minor drawback as compared with the advantages of the system. The intercarrier system is therefore employed in all Philips television receivers for TV systems with frequency-modulated sound.

Amplitude-modulated systems

Intercarrier sound is not applicable to systems with an amplitude-modulated audio channel owing to the fact that the intercarrier signal would here vary in amplitude with the sound as well as with the vision signal, thus preventing any separation of the two modulations. *Fig. 6* shows the method of detection employed in conjunction with amplitude-modulated sound. The mixer valve (M) produces an I.F. vision signal and an I.F. sound signal; the sound signal is amplified by a separate

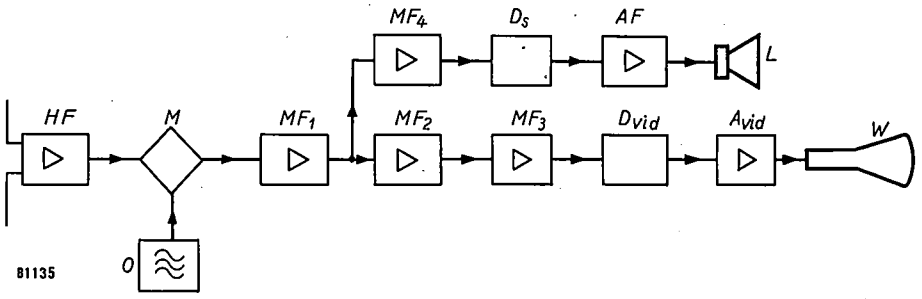


Fig. 6. Block diagram of a receiver for television systems with amplitude-modulated sound. *HF* radio-frequency amplifier stage, *M* mixer stage, *O* local oscillator, MF_1 first intermediate-frequency amplifier stage, for both vision and sound; MF_2 second and MF_3 third intermediate-frequency amplifier stages, for vision alone, D_{vid} video detector, A_{vid} video amplifier, *W* picture tube; MF_4 intermediate-frequency amplifier for sound, D_s amplitude detector for sound, *AF* audio-frequency amplifier, *L* loudspeaker.

I.F. amplifier (MF_4) and detected by an amplitude detector (D_s). In most cases, however, the first I.F. valve (MF_1) can be used to amplify both signals, without interference from cross-modulation; the subsequent I.F. stages for amplification of the vision signal (MF_2 and MF_3) must then include filters to adequately suppress the I.F. sound signal. In other words, the selectivity characteristic should be such that the amplification of the sound signal is at all times a factor of 150-250 (or 44-48 dB) lower than that of the vision signal (*fig. 7*). (Again, of course, the precise value of the factor depends upon the strength-ratio of the vision and sound signals at the aerial terminals of the receiver.)

The selectivity characteristic of the stages of the receiver up to and including the sound detector is shown in *fig. 8*. It is seen that tuning is far more critical than in the case of intercarrier sound. Detuning by one or two hundred kc/s is enough to eliminate the sound; hence steps must be taken to avoid frequency drift and microphony in the local oscillator.

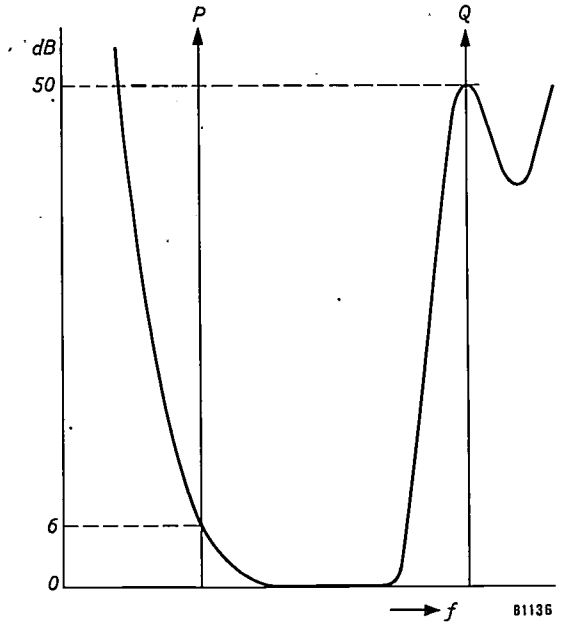


Fig. 7. Selectivity characteristic of the circuit up to the video detector in a TV receiver using A.M. sound. The amplification in the sound channel should be about 44 dB lower than that in the vision channel.

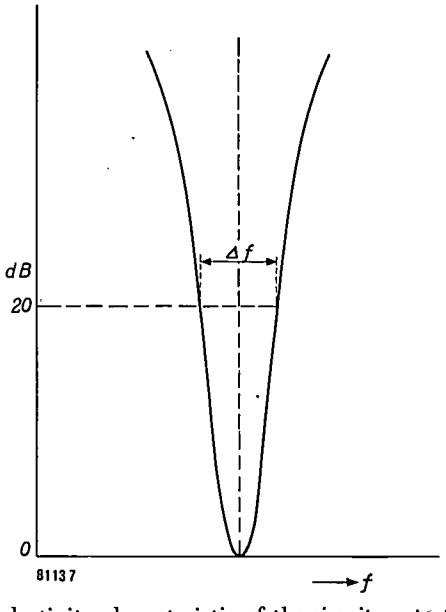
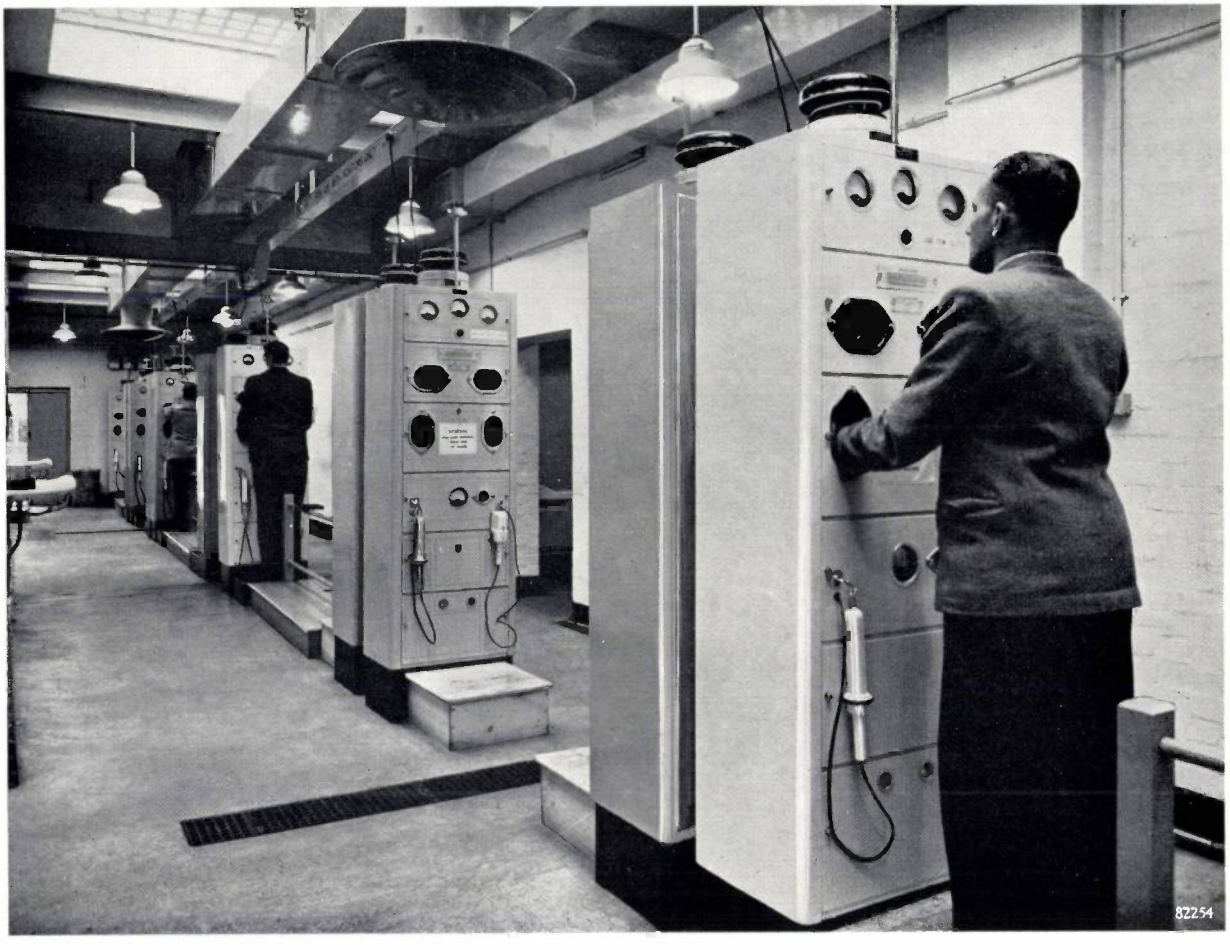


Fig. 8. Selectivity characteristic of the circuit up to the audio detector in a TV receiver using A.M. sound. At 20 dB the width Δf is about 300 kc/s.

The adverse effect of interfering pulses in the sound can be reduced by including in the audio-frequency stage an interference suppressor which limits the amplitudes to a certain level. However, this cannot be done as effectively as in the case of frequency modulation, unless very complex circuits are employed.

Summary. A survey of the principal television standards at present in force is followed by an analysis of certain associated problems which are of special interest to the designer of television receivers. The points considered are: 1) Channel width; 2) Positive or negative picture modulation in relation to the visibility of interference and the effect of interfering pulses on synchronization and automatic gain control; 3) The system of sound transmission, that is to say, by frequency modulation or amplitude modulation; in connection with frequency modulation, the method of "intercarrier sound" is discussed.



A RADIOACTIVE-CONTAMINATION MONITOR FOR THE HANDS, FEET AND CLOTHING

by A. NEMET *), R. B. STEPHENS *) and W. A. BAYFIELD *).

621.387.4:614.8-084.4

In laboratories and establishments concerned with the handling of radioactive materials, or where ionizing radiations are present, the personnel are subjected to a number of health hazards which require that certain precautionary measures are taken. Among these are the detection and measurement of the sources of harmful radiations by means of special instruments normally referred to as monitors.

This article describes a monitor for the protection of personnel, which was developed some years ago for use in atomic energy establishments in England. The instrument is interesting for the diversity of the counting devices employed to meet the unusual operating conditions.

Instruments for monitoring radioactive hazards may be conveniently classified into the following main types [1]):

1) *Personal monitors* for the measurement of accumulated radiation exposure (dosemeters) [2]).

2) *Survey meters* for measuring the radiation intensity in different parts of the laboratory. From the

*) Philips Balham Works Ltd., London

[1]) D. Taylor, Radioactivity surveying and monitoring intsruments, J. Sci. Instr. **27**, 81-88, 1950.
See also R. B. Stephens, J. Brit. Inst. Radio Engrs. **14**, 377-386, 1954.

[2]) A recently developed instrument of this type is described by N. Warmoltz and P. P. M. Schampers, Philips tech. Rev. **16**, 134-139, 1954/55 (No. 4).

radiation field plotted from these measurements, laboratory work can be planned so that the exposure of each individual worker is kept to a minimum.
3) *Contamination monitors* for the detection of radioactive materials. These instruments may be subclassified as
a) *Area contamination monitors* for checking the contamination of working surfaces, apparatus etc., and if necessary, the air in the laboratory.
b) *Personnel contamination monitors* for checking contamination on the hands, feet and clothing.

An area monitor has previously been described in this Review [3]). It is the purpose of this article to describe an instrument of the last-mentioned class, viz. a personnel contamination meter for monitoring the hands, feet and clothing [4]). This instrument was developed some years ago by the X-ray and electro-medical laboratory of the Philips Balham Works in collaboration with the Atomic Energy Research Establishment at Harwell. The title photograph shows a number of the instruments in use in an atomic establishment in England. (Photograph reproduced from "Britain's Atomic Factories" by permission of the Comptroller of Her Majesty's Stationery Office.)

Design requirements of the monitor

The monitor is required to check that personnel working with radioactive materials or employed in an area where such materials are handled are free from health hazards due to activity inadvertently picked up on the hands or clothing. The health hazard may be due to penetrating γ radiation from a source of activity lodged, for example, in the clothing, or the activity may be ingested, where the hazard will depend on the amount absorbed by the body, the type of activity and its half-life and also on the "biological half-life" (a measure of the time the substance will remain within the body). Contamination with certain α-emitting "bone-seeking" elements is especially dangerous since the red and white corpuscles of the blood are destroyed at their source by the ionization. The tolerance levels for such ingested materials are therefore very low.

The only protection against these hazards is thorough washing of the hands and removal of any active material from clothing. To check the efficiency of these cleansing processes the contamination monitor is then used:
a) To check that the contamination of the hands is below tolerance level. This is done separately for α and β-γ contamination, because the nature of the radiations is different and hence different types of detector are necessary.
b) To check that the clothing is free of contamination, again separately for α and for β-γ activity.
c) To check the level of activity on the soles of the shoes, due to active material picked up from the floor. This is done only for γ-contamination.

The essential features which distinguish the present instrument from other monitoring equipments are as follows:
1) A relatively large area (hands, clothing) has to be monitored.
2) The level of activity to be measured is low. In fact, it often corresponds to an activity of only the same order as the number of background counts due to cosmic rays and naturally-occuring radioactivity.
3) The relative positions of source and counter are not under very close control: for example, the active material may be lodged in crevices in the skin or under the fingernails.
4) The monitor has to be used by people of widely different intelligence levels and must therefore be fully automatic, foolproof and require no skill on the part of the user.

The problem of detecting and measuring low levels of activity under such conditions resolves itself into one of choosing the most suitable type of detectors and adapting their design to suit the given counting geometry.

The problem of establishing the tolerance levels, i.e. the maximum permissible activity ingested or deposited on the various parts of the body, is a difficult one and the levels have undergone changes during the past years as experience of physical and biological effects has accumulated. The instrument was therefore designed in such a way that the indication may be adjusted to conform with the prevailing tolerance levels. At the time the monitor was designed the following maximum permissible levels were adopted:
1) α-contamination on hands: 600 disintegrations per minute per hand.
2) β or γ-contamination on hands: 6000 disintegrations per minute per hand.
3) γ-contamination on feet: 500 000 disintegrations per minute per foot.
4) α-contamination on clothes: not specified; see below, p. 205.
5) β or γ-contamination on clothes: As 4), not specified, but see below, p. 204.

Choice of detectors

Owing to the different maximum permissible

[3]) G. Hepp, A battery operated Geiger counter, Philips tech. Rev. 14, 369, 1952/53.
[4]) F. S. Goulding and K. E. G. Perry, A hands and feet contamination monitor, Atomics 2, 43-50, 1951.

activities for α, β and γ, and the different counting geometry in the case of the hands, feet and clothing, the choice of the detectors for the various purposes poses a different problem in each case. These will now be discussed in turn.

Detectors for α-contamination on hands

The detection of α-activity on the hands might be achieved either with the aid of a scintillation counter or with a proportional counter. The former has the merit of a very high sensitivity and a short resolving time (although this is not important at low levels of activity). However, it is difficult to design a scintillation counter to cover a large counting area, an important requirement in this application.

The proportional counter, on the other hand, has the merit of simplicity of construction, sensitivity adequate for the purpose and reasonable stability during operation. It is also readily adaptable to large area counting. This type of counter has therefore been adopted for the α-monitoring of the hands.

It is true that proportional counters have certain disadvantages: microphony, small pulse height (of the order of 3 mV) requiring a high-gain amplifier, and high operating voltage (2.5 kV) which can lead to spurious pulses due to minute insulation breakdown in components. These points do not, however, outweigh its useful properties.

The proportional counter has the characteristic that the voltage pulse produced is proportional to the number of electrons initially produced, i.e. proportional to the energy of the incident particle. This is an extremely valuable feature, for, by suitable discrimination in the counting circuits, pulses of one kind may be counted in the presence of a high background of pulses of another kind but of lower energy. Thus in the present case, the use of these counters enables α-activity to be counted in the presence of a strong β-background.

In its simplest form the proportional counter consists of a cylindrical cathode surrounding a central anode wire of small diameter (see *fig. 1*). A source of high potential is connected across the counter via a resistor R in the anode lead. The dotted capacitor C represents the total capacitance of the counter and the resistor. The counter may be filled with air or other gas mixture.

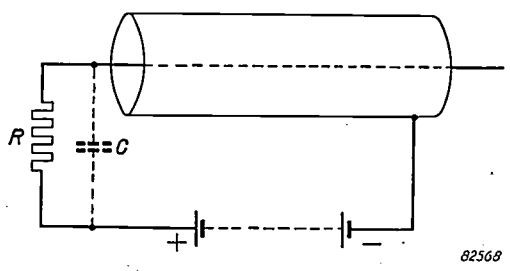


Fig. 1. Simple proportional counter. R resistance, C effective capacitance of counter.

Ionizing particles entering the counter and colliding with the gas molecules will produce free electrons which will be accelerated to the anode by the electric field. The collection of the electrons at the anode will give rise to a voltage pulse V across R of magnitude [5])

$$V = \frac{Ane}{C}$$

where n is the number of electrons formed in the initial ionizing event, e the charge on the electron, C the capacitance of the system and A the gas amplification factor.

The gas amplification arises because (assuming the anode voltage of the counter is sufficient) each initial electron entering the region of intense field near the anode wire will gain sufficient energy to produce further ionization by collision. Thus an avalanche of electrons is produced by each initial electron. The quantity A as defined above is the number of resulting electrons produced in this way.

It should be noted that if the applied voltage is small the gas amplification factor is unity and the counter becomes an ionization chamber. However by suitable choice of gas filling, pressure, voltage and diameter of anode wire, an amplification as high as 10^4 may be obtained.

For α-counting, some restriction is placed on the design of the proportional counter because an extremely thin window is necessary to minimize absorption. For this reason, air at atmospheric pressure has to be used as the gas filling. This of course means that the gas amplification is fixed since it is dependent on the choice of gas and the pressure.

The proportional counter used for monitoring the hands is of a flat construction, covering the whole area of the average hand. The design of this type of counter has been thoroughly treated by Simpson [6]). In the present instrument the construction is as shown in *fig. 2*. A number of fine tungsten

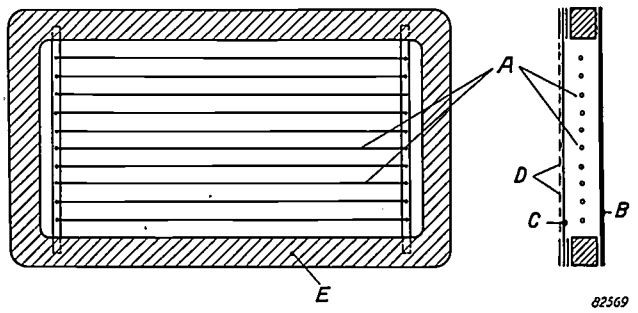


Fig. 2. Proportional counter of flat construction used for the counting of α activity on the hands. A tungsten anode wires 0.001″ (25 μ) diameter, B metal backing plate, C aluminized paper window 8 μ thick, D protective metal grid. The polystyrene frame E is about 5″ × 8″ (125 × 200 mm).

[5]) S. A. Korff, Electron and nuclear counters, Van Nostrand, N.Y. 1946, p. 34).

[6]) J. A. Simpson, Rev. sci. Instr. **19**, 733, 1948.

wires 0.001 inches (25 μ) in diameter are stretched between two nickel rods fixed in a moulded polystyrene frame. These wires form the anode of the counter. The frame is sandwiched between a metal backing plate and a paper window 8 microns thick. The cathode consists of the backing plate and the paper window which is aluminised to render it conducting. A strong metal grid protects the window.

Four of these counters are used in the equipment, two to each hand. Their exact arrangement and mounting which are described later, gives a net counting efficiency (number of counts per disintegration) of the order of 25 %. The α tolerance level of 600 disintegrations/min therefore corresponds to a rate of about 75 counts/min for each counter.

Detectors for β and γ contamination of hands

The levels of β and γ activity which must be detected are of lower order than can be conveniently measured with ionization chambers or proportional counters. The most suitable type of detector for this purpose is the Geiger-Müller tube. It is not proposed to discuss the action of the Geiger-Müller tube in detail as this subject has been adequately dealt with elsewhere [7]), [8]).

However, it may be noted in passing that in this tube the gas amplification is very high so that the initial ionization creates an avalanche of ions which spreads throughout the whole of the active volume of the tube. The result is that there is no longer any relation between the energy of the incident particle and the voltage pulse produced. The discharge has to be quenched, by either internal or external means.

In the contamination monitor, self-quenching ethyl bromide-filled Geiger-Müller tubes habe been used, having an operating voltage of 1100 volts and a plateau (range of anode voltage over which the counting rate does not vary appreciably) about 100 V long. This type of a counter was chosen because they were readily available at the time the equipment was designed and their behaviour is well known.

For the measurement of β- and γ-activity of the hands the following arrangement was used. Four thin-walled cylindrical counters, one on either side of each hand, are arranged to scan the hand from the fingertips to the wrists. The scanning period is thirty seconds, during which time a total of 100 counts may be expected from a "tolerance hand" (= 100 disintegrations per sec). The net counting efficiency is therefore about 3 %). This low figure is caused mainly by the geometry of the arrangement. The background counts from the four counters during this period would be about 80, so that the introduction of both, "tolerance hands" would increase the meter reading to rather more than twice the background value.

The net efficiency might have been raised by dispensing with the scanning system, and employing banks of counters above and below the hands sufficient in number to cover the whole area of the hands. This scheme, however, would suffer from the serious disadvantage that each counter would require to be separately adjusted in respect of its anode voltage and separately re-adjusted from time to time during the life of the tube as the quenching vapour was used up. About 16 counters would be involved, so this setting up and maintenance procedure would become very complicated.

Detectors for γ-contamination of feet

For measurement of the γ-activity on the shoe soles two large Geiger-Müller counters with an active length of about 10 inches (250 mm) and a diameter of $1^1/_4$ inch(30 mm), are fixed beneath the sheet steel platform on which the user of the instrument stands. Owing to their high energy the γ-rays are not appreciably attenuated by the platform.

The net efficiency of the arrangement is very low due partly to the poor geometry and partly to the inherently low efficiency of a Geiger-Müller counter for γ-rays. The geometry is such that the solid angle subtended by the counters is only a small fraction of 4π. Even so, activity corresponding to the tolerance level gives a meter reading (900 counts/min) which is double that due to background alone.

Detector for β-γ clothing probe

For β-γ monitoring of the clothing a similar type of Geiger-Müller counter is used to that for measuring the β-γ activity of the hands. The tube is mounted in the assembly shown in *fig. 3* and "scanning" of the clothing is done manually instead of automatically as in the β-γ hands unit. The net counting efficiency is such that 900 counts/min corresponds to the tolerance level, when the probe is held directly on the patch of activity. When not in use, the β-γ probe hangs on a hook on the left of the instrument (*fig. 4*).

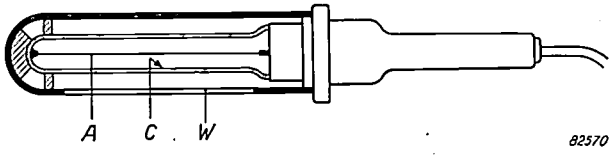


Fig. 3. β-γ clothing probe, consisting of a Geiger-Müller tube mounted in a protective assembly. *A* anode wire, *C* graphited glass cathode, *W* window in probe body.

[7]) H. Friedman, Proc. Inst. Radio Engrs. **37**, 791, 1949.
[8]) D. Taylor and J. Sharpe, Nuclear particle and radiation detectors, II. Proc. Instr. Electr. Engrs. **98**, Pt. II, 214, 1951.

Detector for a clothing probe

For α-monitoring of the clothing, again the possibility of using either an air proportional counter or a scintillation counter presents itself, both types having an adequate degree of sensitivity. An essential requirement of a clothing probe however is, that it should be able to withstand the rough handling it may be expected to receive when used by unskilled personnel. Proportional counters suffer from microphony due to vibration of the anode wires; this precludes their use in probes which will be subject to vibration and bumping during use.

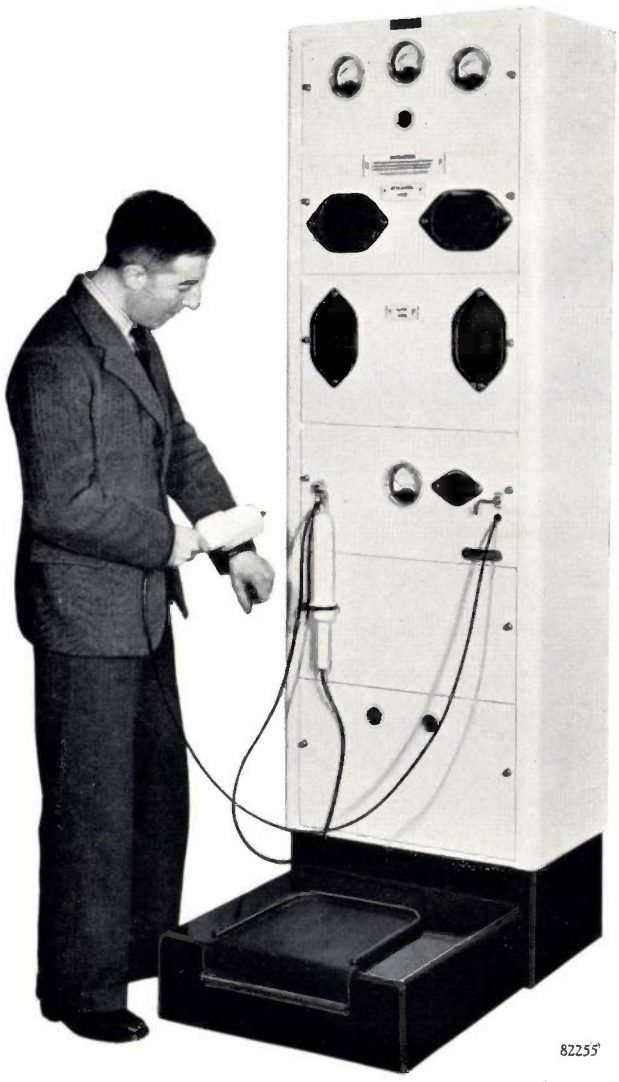


Fig. 4. Photograph of the monitor showing the α-probe in use.

A sketch of the scintillation counter used in the α-probe is shown in *fig. 5*. Essentially it consists of a thin light-tight window, a zinc sulphide screen and a photomultiplier. The window Al, whose function is to transmit α-particles but exclude external light from the photocell, consists of two sheets of aluminium foil each 8 microns in thickness. (Two thicknesses of aluminium are used to reduce the possibility of light leakage through pin holes, which are unlikely to be opposite each other.) The zinc

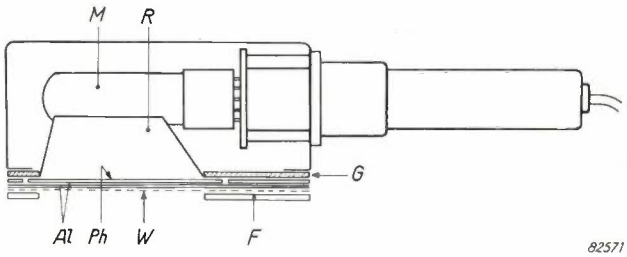


Fig. 5. Construction of the α clothing probe. *F* frame, *G* rubber gasket, *W* wire mesh protective cover, *Al* two aluminium foils each 8 μ thick, *Ph* phosphor (zinc sulphide), *R* reflector, *M* photo-multiplier.

sulphide screen *Ph* is activated with silver and coated on a "Perspex" sheet. The incident α-particles cause scintillations to occur on the screen; the light is reflected on to the photosensitive surface of the photomultiplier by means of a prism-shaped aluminium reflector *R*.

The net counting efficiency of this probe is such that about 300 counts/min corresponds to the tolerance level, when the probe is held directly on the active spot.

Operation and circuitry

Before discussing the actual construction of the instrument and its measuring circuits, it is instructive to first consider the measuring procedure and the various operations involved.

Measuring procedure

The user first pushes both hands through the vertical slots of the α-hand unit (see title photograph and fig. 4). The fingertips operate microswitches which energize a motor; this drives the vertically disposed α-counters on to the hands via a cam and lever mechanism, spring-mounted to allow for differing hand thicknesses. When the counters have closed on the hands, another microswitch energizes relays which switch off the motor and, after a short delay, trigger a timing circuit for the 15-second counting period. A warning light "Test now on" is also switched on. The integrated counts are indicated separately for each hand on the meters at the top left and right of the instrument (fig. 4). The meter readings build up until the end of the counting period, when the warning light extinguishes. The reading now on the meters indicate the α-contamination on the hands. Removal of the hands cancels the meter readings and causes the motor to open the counters in readiness for the next count. Should the hands be removed before the end of the 15-

second counting period an alarm rings. The alarm also rings if the α-contamination of either hand exceeds the tolerance level. The user must then wash his hands afresh and re-check with the monitor.

For measurement of the β-γ contamination, the hands are pushed through the horizontal slots in the β-γ hand unit (fig. 4). Switch bars at the rear of this unit are depressed by the fingertips to start the operation. The Geiger-Müller counters, mounted on a carriage, are then driven slowly so as to "scan" the hands. At the beginning of the scan, cam-operated contacts open-up the β-γ hand counting circuit (and also the γ-feet counting circuit, see below). At the end of its travel, the carriage carrying the counters switches off the motor and cuts out the counting circuit. A synchronous motor is used, so that the time taken for the scanning, and hence the duration of the counting period, is accurately determined; the period is adjusted to be 30 seconds. The β-γ contamination of the hands is indicated separately for each hand, on the same meters at the top of the instrument. Again, the alarm bell sounds if the contamination exceeds the tolerance level, or if the hands are removed before the end of the counting period.

The γ-count for the feet takes place simultaneously with the β-γ hands count, and the integrated count over the 30 second period is registered on the centre meter at the top of the instrument.

The probe units which come into operation only when removed from their hooks, register their counts on the meter on the probe panel, half-way down the instrument. (fig. 4).

Measuring circuits

Fig. 6 shows a block diagram of the counting and timing unit. The α-counters and the β-γ counters for the hands and the γ foot counters are connected via amplifiers to the counting circuits of this unit. The α-probe (scintillation screen + photomultiplier) and the β-probe (G.M. tube) are connected via amplifiers to a logarithmic count-rate meter and a loudspeaker.

The counting and timing unit provides for the control of the various operational sequences, the counting of pulses over a specified time interval, and registering the integrated count as a meter reading. Only three counting circuits are required for the five counting channels since the α-activity and the β-γ activity of the hands are not measured at the same time. Accordingly two of the counting circuits (with their associated meters) serve for both the α-measurement and the subsequent β-γ measurement.

Each counting circuit consists of a monostable flip-flop [9]) which serves as a pulse shaper, followed by an integrating circuit which feeds the indicating meter. A pulse-shaper is essential when using this integrating circuit: without it the varying pulse heights would be integrated into a meter reading quite unrelated to the number of counts. The flip-flop also serves as a discriminator and rejects pulses below a certain height (e.g. noise): a pulse amplitude of about 5 V is necessary to trigger the flip-flop circuit.

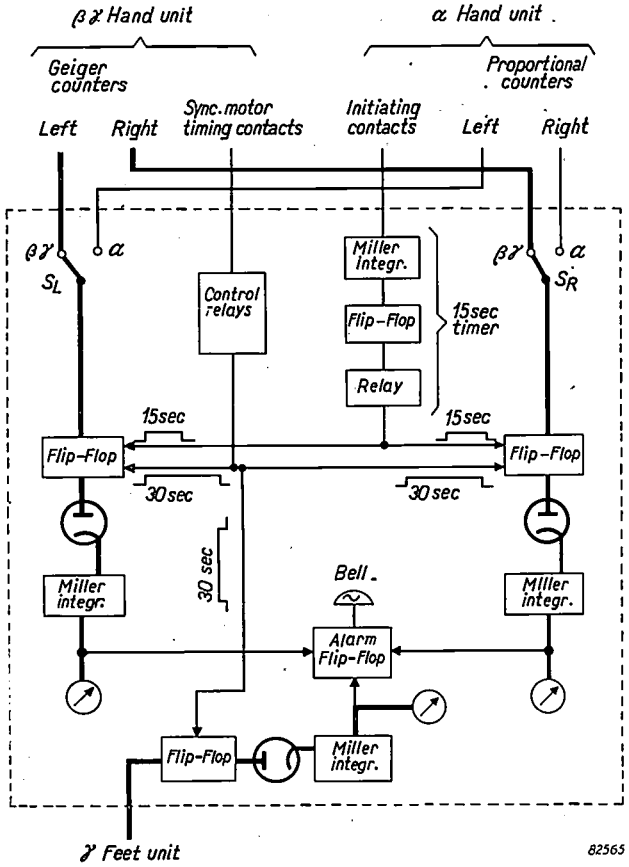


Fig. 6. Block diagram of the counting and timing unit. Three separate counting channels are provided. One of these counts γ pulses from the feet unit only, while the other two count either the α-activity or the β-γ activity of the hands, according to the position of the switches S_L, S_R. The bold lines indicate the routes of the counting pulses (S_L and S_R connected to β-γ unit). The counting period for α-activity of the hands is controlled by a 15 second timing unit based on a Miller integrator circuit. The counting period for the β-γ activity of the hands (and the γ activity of the feet) is controlled by a synchronous motor. An alarm circuit is connected to all three counting Miller integrators.

Integrated count circuit

The integration is performed by a Miller type of circuit [10]) illustrated in *fig. 7*. The circuit consists

[9]) See for example O. S. Puckle, Time bases and scanning generators, Chapman and Hall, 1951 (2nd Edn.) pp. 72-85.

[10]) See for example, Philips tech. Rev. 12, 328-330, 1950/51, or O. S. Puckle, loc. cit., p. 340.

essentially of a pentode, which is normally cut-off by a switch which keeps the grid at a potential some 10 V. below that of the cathode. A capacitor C couples the anode to the grid. At the beginning of the counting period the switch is opened leaving the grid floating. The leakage currents are so small that the grid voltage does not drift appreciably—insufficient, anyway, to allow the valve to conduct,

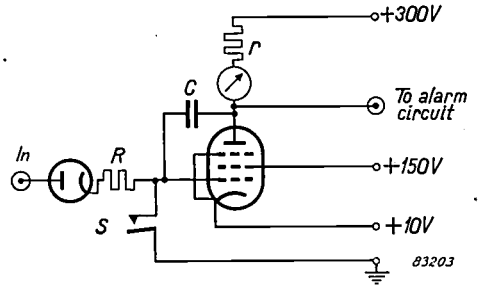


Fig. 7. Miller integrator circuit.

in the absence of pulses on the grid. The shaped counting pulses are fed to the grid via an isolating diode and the resistor R. Each pulse gives rise to an increment of charge on capacitor C which causes the anode current of the valve to increase by a fixed amount.

The charge entering the capacitor per pulse is $V_p t/R$ where V_p = pulse amplitude and t = pulse duration, so that the increase in potential across the capacitor is $V_p t/RC$ per pulse. Due to the gain of the valve the greater part of this voltage change occurs at the anode, so that the drop in anode voltage is nearly proportional to the increment of charge due to the incoming pulse. Hence the current through the meter is given by

$$I = N \cdot \frac{V_p t}{RCr},$$

where N is the total number of pulses arriving since the beginning of the counting period.

Three such integrated count circuits are used (fig. 6): one for the left hand (α or β-γ), one for the right hand (α or β-γ) and one for the feet (γ only). During the β-γ counting operation, the time constants (RC values) of the Miller integrators common to both α and β-γ channels are automatically changed to correct the meter scales for the difference in counting efficiency and in α and β-γ tolerance levels.

The 15 second timer

A Miller integrator circuit is also used for the accurate timing of the counting period for the α-activity of the hands. In this case the Miller pentode is normally held in the conducting state (*fig. 8*) but the cathode current passes entirely through the screen resistor R_4, because the suppressor grid is held some 150 V below the cathode potential so that no anode current can flow. When the hands are inserted into the instrument and the counters have closed on the hands, the switch S is opened and the suppressor rises to cathode potential. Anode current then flows and the screen current drops sharply. The voltage drop at the anode is fed via C to the grid, causing the valve to become nearly cut-off. At the same time, the rise of screen voltage triggers the flip-flop and energizes the timing relay which starts the counting period. The voltage on the grid begins to rise again due to the current entering via R_1 but again, due to the gain in the valve, the anode voltage drops more rapidly, at a rate determined to a first approximation by C and R_1. When the anode bottoms (at about 20 V above cathode voltage) the grid voltage rises rapidly and the screen current increases. The screen voltage therefore drops and the flip-flop returns to its normal state, opening the timing relay and ending the counting period. The duration of the "run down" of the Miller circuit for the α-counting is chosen to be 15 sec.

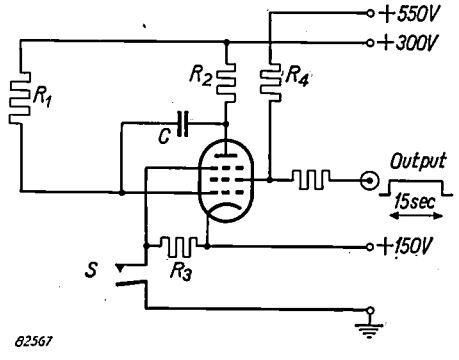


Fig. 8. Miller integrator circuit used for timing in 15 sec timer.

The alarm circuit

The anodes of the Miller integrator valves of the three counting circuits are coupled to another flip-flop. If any of the Miller valves suffers an excessive drop in anode voltage, the flip-flop is triggered, causing the alarm bell to ring. The flip-flop is so adjusted that it is triggered whenever the activity measured by any of the three counting circuits is above tolerance.

The alarm bell is also coupled via relays to the hand-operated switches on the α-hand unit and the β-γ hand unit, in order to give audible warning if either of the hands are removed during a counting period.

Construction and layout

The α-hand unit

The layout of the α-hand unit is seen in *fig. 9*. The fingertip-operated microswitches can been seen at the rear of the counters. The timing relay energized by the 15 second timer (fig. 6) removes the negative bias on the flip-flop circuits, permitting the counting pulses to pass through.

The output pulses (approx. 3 mV) from each pair of counters are fed to the input of a three-stage feedback amplifier with a gain of about 1600. This raises the pulse amplitudes to the 5 V necessary to trigger the flip-flops feeding the Miller integrators.

Due to the low level of the signal pulses at the counters, great care has to be taken to avoid spurious counts due to such causes as microphony in amplifiers and in the counters themselves, leakage across the insulators and accumulations of dust and small hairs in the counters. The effects of microphony in the system have been greatly reduced by giving the amplifier a low-frequency cut-off at about 10 kc/s. The insulators are cleaned with alcohol during assembly and the counters blown out with warm clean air to remove dust and any small hairs. Leakage due to condensation on the counters is minimized by heating resistors placed under them, which maintain a local ambient temperature of about 40 °C.

The β-γ hand unit

Fig. 10 shows an upper view of the β-γ hand unit. At the back of the flat grids are the switch-bars which start the counting operation. The cam-opera-

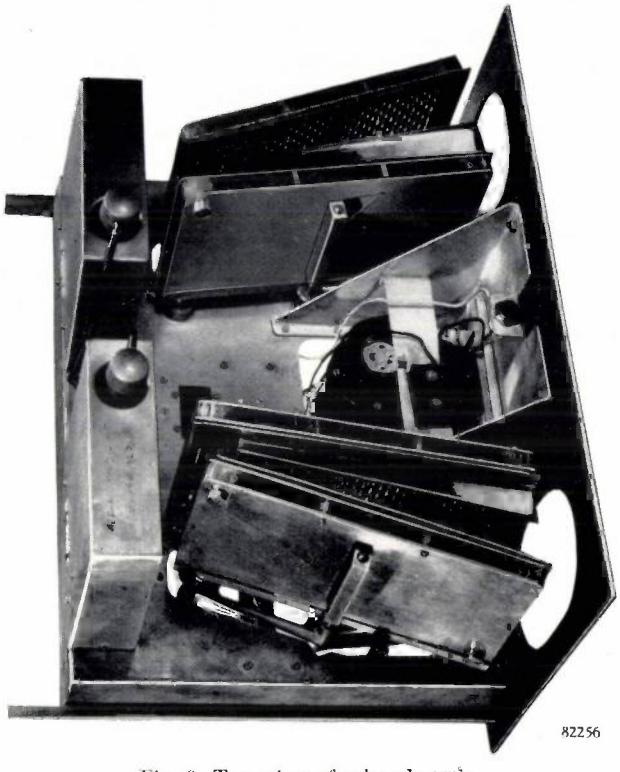


Fig. 9. Top view of α hands unit.

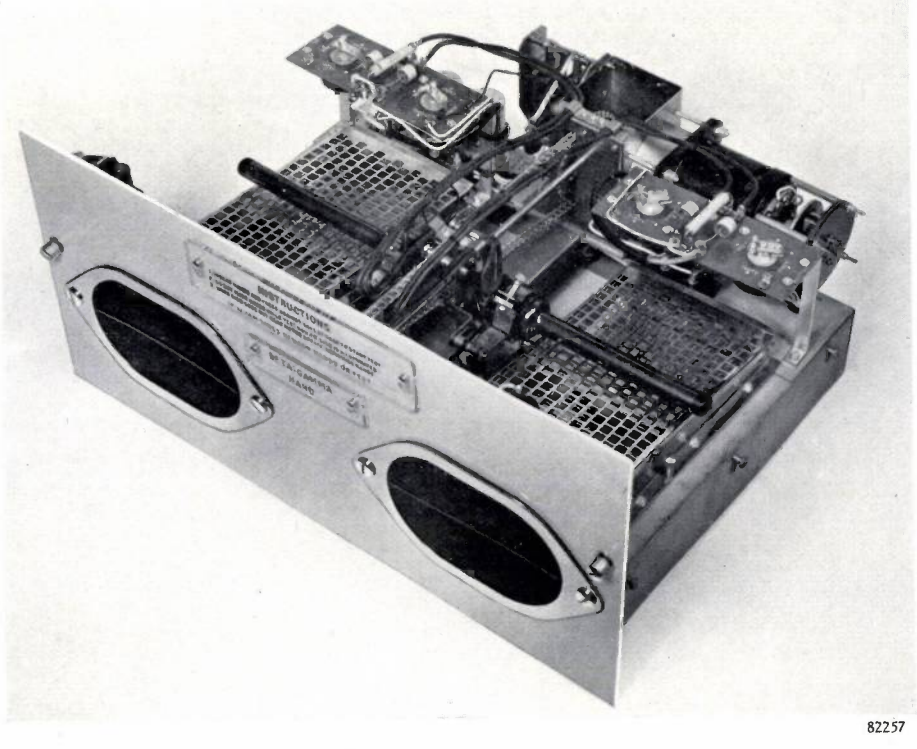


Fig. 10. Top view of β-γ hands unit.

ted contacts (at rear of unit) start the counting period proper by removing the negative bias on the flip-flops.

Since Geiger-Müller counters are used in this unit, the counter pulses are already fairly large and only little amplification is required before they are fed to the counting circuit. In fact a gain of about 5 is adequate and this is provided by a single valve in each channel.

As in the case of the β-γ hands counters, the pulses from the Geiger-Müller tubes measuring γ activity of the feet are large enough to require only one stage of amplification ($\times 5$). The counting and alarm circuit (fig. 6) are similar to those of the hand-counting channels.

The probe unit

In the probe unit a logarithmic counting rate circuit is employed for both the α-probe and the β-γ probe, as mentioned earlier. This permits counting over a wide range of activities without the necessity of changing the range (a useful feature in an instrument to be operated by unskilled persons). The circuit consists of a flip-flop so arranged that the duration of the current pulse from the normally cut-off valve diminishes as the frequency of the input pulses increases. This results in an approximately logarithmic scale giving reasonable reading accuracy over a counting rate range of 500 : 1. The meter itself is calibrated not in counts/sec but in "times tolerance". Full-scale on the meter corresponds to an activity of $100\times$ tolerance level. A single stage of amplification is used to give the audible indication of the count rate on the loudspeaker.

When the β-probe is removed from its hook (fig. 4), a microswitch raises the voltage on the Geiger-Müller tube by about 300 V so bringing it on to the correct operating point of its characteristic. Counting then begins. Only a single stage of amplification is required for the pulses to trigger the count-rate circuit. By maintaining the counter voltage about 300 V below normal while not in use, the life of the counters is conserved. This procedure has been adopted with all the Geiger-Müller tubes in the apparatus.

When the α-probe is unhooked, a microswitch energizes a relay which switches over the input of the logarithmic count rate circuit so that it receives pulses from the amplifier of the photomultiplier. The relay also changes the time constant of the flip-flop so that the duration of its current pulses is altered to give a meter scale conforming with the α-tolerance level. The anode resistance for the multiplier is situated in the main unit, so that the same cable carrying the high tension supply also carries the output pulses fed to the ratemeter. A length of concentric cable of low capacitance is used so as to obtain as large a pulse height as possible.

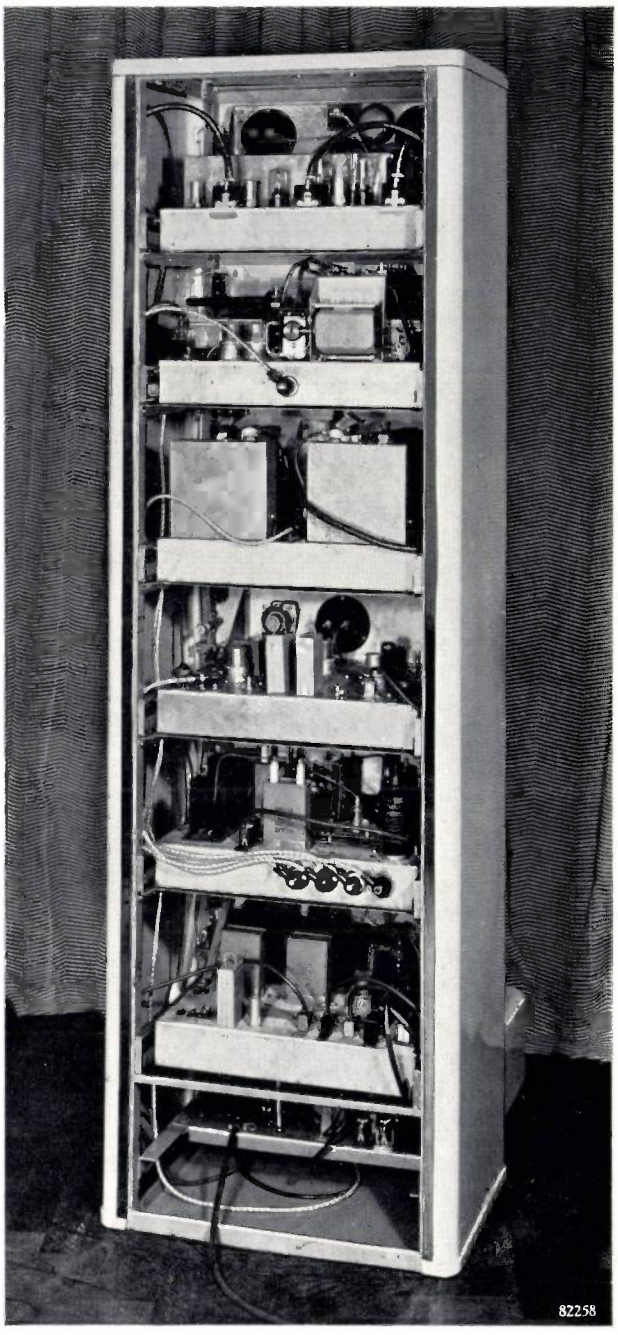


Fig. 11. Rear view of complete instrument. From top to bottom: counting and timing unit, β-γ hands unit, α hands unit, probe unit, E.H.T. power pack for counters (2500 V and 1200 V), power pack for valves (300 V), power pack for relays (50 V). The platform in front of the instrument contains the Geiger Müller counters for measuring γ-activity on the feet.

General layout

The layout of the complete instrument can be seen from the rear view shown in *fig. 11*. The various

counting and indicating units occupy the top four chassis. At the bottom of the rack is a 50 V power unit supplying the relays of the instrument. Immediately above this is the 300 V D.C. stabilized power supply for the valves of the monitor. This circuit is of conventional design, using a hard valve rectifier and a series-parallel type of stabilizer [12]).

The E.H.T. power pack (third unit from bottom) provides the stabilized D.C. supplies for the various counters, viz. a nominal 1200 V supply for all the Geiger-Müller counters and the photo-multiplier of the α-probe and a nominal 2500 V supply for the proportional counters. Both these supplies are also stabilized by means of series-parallel stabilizers. The intermediate anodes of the photo-multiplier are supplied from a high resistance potential divider connected across the 1200 V supply.

For ease of servicing, all the units of the instrument are mounted on sliding arms so that they can be withdrawn from the front of the apparatus. Interconnections between the units are made by plug and socket connections, with sufficient spare cable to allow the units to be pulled forward. When the units have been so pulled forward, a system of interlocks ensures that the high voltage supplies are disconnected.

All labels and escutheons are made detachable so that they can be easily decontaminated. The lines of the apparatus have been made as smooth as possible so as to avoid crevices where contaminated dust might lodge and be difficult to remove.

Thanks are due to the Director of the Atomic Energy Research Establishment at Harwell for permission to publish this article.

Summary. The need for an instrument to monitor the hands, feet and clothing of personnel working with or near radioactive materials is explained with reference to the health hazards involved and the safe tolerance levels. After an outline of the requirements of such a monitor, the detectors for the different measurements are discussed. Proportional counters are used to measure α-activity on the hands; Geiger-Müller counters are used for β,γ-activity on the hands and the clothing, and for γ-activity on the feet; for α-activity on the clothes a scintillation counter is used. Some of the timing and counting circuits are described, notably the Miller integrator circuit on which the hands and feet measurements are based. The probe units for measurement of activity on the clothing use a logarithmic count-rate circuit. Special emphasis is laid on the automatic operation of the instrument, which requires no skill on the part of the user. The construction of some of the units and their layout are briefly treated.

[12]) See for example, H. J. Lindenhovius and H. Rinia, Philips tech. Rev. 6, 54-61, 1941; or F. A. Benson, Voltage Stabilizers (Electronic Engineering, London, 1950).

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can de obtained free of charge upon application to the administration of the Philips Research Laboratory, Eindhoven, Netherlands.

R 222: W. K. Westmijze: Studies on magnetic recording IV (Philips Res. Rep. **8**, 343-366, 1953, No. 5).

Concluding article of the series **R 213, 214** and **217.** A comparison is made between numerical values given by the formulae derived earlier and the experimental data. Special consideration is given to the relation of the output to the biasing current and to the frequency and the distortion as a function of signal current and biasing current. Finally, changes of a magnetic recording with time are discussed; in particular the magnetization by weak magnetic fields from adjacent layers of tape. Measurements of this so-called print effect are compared with the magnetic lag calculated from the Brownian movement of particles in a field of force. The logarithmic increase of the print effect with time is explained in this way.

R 223: H. van de Weg: Quantizing noise of a single integration delta modulation system with an N-digit code. (Philips Res. Rep. **8**, 367-385, 1953, No. 5).

In order to compare the delta-modulation system (ΔM system) with the pulse code modulation system (PCM system) as regards the bandwidth required to obtain a certain signal-to-quantizing-noise ratio, the quantizing noise is calculated for a single-integration ΔM system. The possibility that the difference between signal and approximation may be coded in a more-digit binary code is also considered. In accordance with Bennett, the information signal is chosen to consist of a noise function, in order to obtain average quantizing-noise spectra that are smooth functions of frequency. Because in ΔM the signal is not limited in amplitude but in slope, the signal-to-noise ratio depends on the frequency

spectrum of the applied signals in another way than is the case with PCM. It turns out that for speech transmission, the ΔM system has more favourable properties than PCM, if the same sampling frequency and number of digits are taken in both cases. An analytical expression is derived for the quantizing-noise spectrum that is valid for an arbitrary information signal.

R 224: P. Zalm and H. A. Klasens: A new 6100-Å band in zinc orthosilicate activated with manganese (Philips Res. Rep. **8**, 386-392, 1953, No. 5).

When phosphorus is incorporated together with manganese into zinc orthosilicate by heating at relatively low temperatures (820 °C), a phosphor is obtained with an emission band at 6100 Å next to the normal green Mn emission at 5250 Å. The properties of this new emission band are discussed and compared with those of similar bands in other manganese-activated zinc silicates.

R 225: A. Bril, H. A. Klasens and P. Zalm: A new method to determine short decay times of phosphors excited with ultraviolet light (Philips Res. Rep. **8**, 393-396, 1953, No. 5).

A new and simple method is described to determine decay times of phosphors using a pulsed cathode-ray tube with ultraviolet-emitting phosphors of short decay time as light source. The decay time of Sb-activated halophosphates was found to be 5×10^{-6} sec.

R 226: W. de Groot: Some remarks about the so-called Crova wavelength (Philips Res. Rep. **8**, 401-410. 1953, No. 6).

It is shown that it is of minor importance whether the Crova wavelength λ_c is defined with the aid of Planck's or Wien's radiation formula, both definitions leading to approximately the same result. If $V(\lambda)$, the relative spectral luminous efficiency, is approximated by a Pearson function $A\lambda^{-p}\exp(-q/\lambda)$, λ_c is a linear function of T^{-1}. It is shown that the relative luminous efficiency of a perfect radiator is approximately maximum if $\lambda_c T = c_2/4$ or $\lambda_m T = c_2/4$, in which λ_m is the wavelength for which $V(\lambda) = 1$. It is further shown that the theoretical value of the maximum spectral luminous efficiency, K_m, as deduced from the definition of the candela, depends on the values assigned to c_1 and to the ratio c_2/T_{Au} (T_{Au} = melting temperature of gold). To a first approximation,

$$\frac{dK_m}{K_m} = -\frac{dc_1}{c_1} + \frac{1}{\lambda_c(T_{Pt})}\, d\,\frac{c_2}{T_{Au}}.$$

The luminance of a perfect radiator in given in terms of λ_c.

R 227: F. K. du Pré: The counting loss of a Geiger counter with periodic arrival rate of quanta (Philips Res. Rep. **8**, 411-418, 1953, No. 6).

A formula is derived for the counting loss of a Geiger counter with arbitrary dead time, exposed to periodic X-radiation of arbitrary waveform. An "extended" dead time is assumed. In particular, the small-loss case is considered, since then the results are independent of the type of dead-time mechanism. The small-loss corrections obtained by different authors are shown to be special cases of the general formula.

R 228: F. W. Gundlach: Laufzeiterscheinungen der Diode im Anlaufstromgebiet (Philips Res. Rep. **8**, 419-426, 1953, No. 6). (Transit-time phenomena in the exponential region of the diode characteristic; in German.)

Where a diode is in the exponential region of its characteristic (minute electrode distances, negative anode voltage), the convection current at its anode may be calculated by considering the transit-time phenomena at small a.c. voltages. The conductance of the diode is ascertained; it turns out the electrons which do not hit the anode provide a particularly strong contribution. The discrepancies of earlier publications are pointed out. The values calculated agree reasonably well with the measurements published by other authors.

R 229: H. C. Hamaker: The efficiency of sequential sampling for attributes, Part II. Practical applications (Philips Res. Rep. **8**, 426-433, 1953, No. 6).

The practical consequences of the theory developed in part I (see **R 208**) are considered. It is shown that the simplified equations (valid under Poisson conditions) can be used to construct a slide rule from which the operating characteristic of a sampling system can be read off at once when the parameters h_o and p_o are given. Furthermore, graphs are constructed which permit and easy and straightforward transition from the parameters p_o and h_o to the more common parameters such as AQL, AOQL, etc.

R 230: J. L. H. Jonker: The angular distribution of the secondary electrons of soot (Philips Res. Rep. **8**, 434-440, 1953, No. 6).

The angular distribution of the secondary electrons of soot is measured with the same equipment as was used before to measure poly-crystalline

nickel (see **R 175**). The results for soot and for nickel are compared.

H 231: H. A. Klasens, P. Zalm, F. O. Huysman: The manganese emission in ABF_3 compounds (Philips Res. Rep. **8**, 441-451, 1953, No. 6).

A number of compounds of the composition ABF_3 were made with A = Na, K, Rb, Cs and B = Mg, Zn, Cd, Ca, Sr. These substances have the perovskite structure when the corrected tolerance factor 1.15 $(R_A + R_F)/(R_B + R_F)\sqrt{2}$ lies between 0.9 and 1.1 (R_A etc. are ionic radii). They are fluorescent under cathode-ray excitation when activated with Mn. The colour of the Mn emission varies between orange and green. It shifts to shorter wavelengths with increasing Mn-F distance. An attempt is made to explain this shift.

R 232: J. Volger, J. M. Stevels and C. van Amerrongen: The dielectric relaxation of glass and the pseudo-capacity of metal-to-glass interfaces, measured at extremely low frequencies (Philips Res. Rep. **8**, 452-470, 1953, No. 6).

The dispersion of the dielectric constant of glass in the L.F. region has been investigated. The main relaxation time may be explained as the reciprocal transition probability of Na^+ ions jumping between adjacent interstices. Extremely high electrode capacitances have been detected, which are frequency and temperature dependent. These values are in qualitative agreement with a simple formula derived for the dynamic values of double-layer capacitances.

R 233: A. van Weel: Measurements of phase angles (Philips Res. Rep. **8**, 471-475, 1953, No. 6).

Phase angles are measured from the frequency variation of an oscillating circuit, caused by introducing the unknown phase angle into the closed loop of the oscillator circuit. High sensitivity, good accuracy and direct indication on a calibrated scale are some of the features of the system (cf. Philips tech. Rev. **15**, 307-316, 1953/54).

VOL. 16 No. 8, pp. 213-240 FEBRUARY 1955

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

THE THEORY AND CONSTRUCTION OF GERMANIUM DIODES

by J. C. van VESSEM.

621.314.632:537.311.33

The effective use of germanium diodes does not necessarily require a knowledge of the theory underlying the rectification process in these diodes. For the user, the diode characteristic is sufficient as a starting-point. Nevertheless, the editors of this Review feel that many readers will be interested in what happens inside the germanium crystal. Accordingly, current views on the mechanism of crystal rectification and certain details of the design and manufacture of point-contact germanium diodes are discused in the present article as an introduction to the subsequent article on the application of point-contact germanium diodes.

The present insight into the processes of rectification and transistor action in germanium crystals is largely due to Shockley and his associates, whose descriptions of these processes, some scientific and others written in a more popular style, have appeared in various publications [1]).

One of the most important features of these crystals is that conduction takes place in them through the movement of two different types of charge-carrier, known as "free (or excess) electrons" and "holes"; these carriers are mostly generated in pairs (one of each type) and may likewise vanish in pairs by recombination. Rectification takes place under certain conditions as a result of a subtle interplay between charge movements produced by electric fields and diffusion, and the above-mentioned processes of generation and recombination. This will now be explained more fully.

Electrical conduction in germanium

Free electrons and holes in pure germanium

Germanium, like silicon, selenium and tellurium, belongs to a group of semi-conductors owing their conductivity exclusively to electrons, since, as we shall see later, what is known as conduction by "holes" is in fact also a transport of charge by electrons. Semi-conductors in which the charge transport is effected by ions also exist, but belong to another group which will not be discussed here.

In a crystal of silicon or germanium, each atom is situated at the centre of gravity of a tetrahedron formed by the four adjacent atoms to which it is bound. Each bond between two atoms is attributable to an electron-pair of one valence electron from each atom (a so-called covalent, or homopolar, bond). Thus the four valence electrons of each atom are all involved in similar bonds. A two-dimensional representation of such an array will be seen in *fig. 1.*

Thermal vibrations in the crystal lattice may eject an electron from its position. However, such vibrations can only release valence electrons, since the other electrons of the atom are bound too tightly to the nucleus. In fact, the probability that even a valence electron will be so ejected is very small, since it must not only have sufficient energy, but the local situation must be such that an escape-path may be formed. This is equivalent to saying that for release to occur it must be possible for the laws of conservation of both energy and momentum to be satisfied. For the same reason, however, the probability that an electron will drop back to the original site when once released is likewise small; it is then virtually isolated. Nevertheless, such an electron tends to linger in the vicinity of the vacant

[1]) See W. Shockley, Transistor electronics: imperfections, unipolar and analog transistors, Proc. Inst. Rad. Engrs. 40, 1289 - 1313, 1952. Also W. Shockley, Electrons and holes in semiconductors, D. van Nostrand, New York and London, 1950.

site (or hole), owing to the presence at this site of a non-compensated nuclear charge, which attracts it. Because the two charges attracted to one another are in a dielectric (i.e. germanium, where the dielectric constant $\varepsilon_r = 16$), the attraction is weak; even at room temperature the electron is more or less free to drift through the crystal at random under the influence of thermal agitation.

The holes are also affected by thermal agitation. A hole may easily become occupied by a valence electron from an adjacent bond, leaving another hole at the original site of the electron thus displaced. Again, a valence electron may move into the new hole, and so the process will continue. In effect, then, the original hole roams through the crystal.

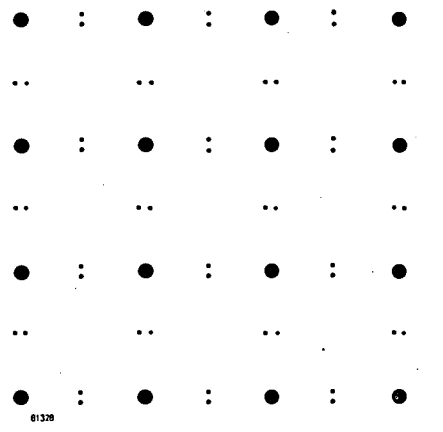


Fig. 1. Two-dimensional representation of the germanium crystal lattice. The large dots represent germanium atoms, and the smaller ones valence electrons paired to form homopolar bonds between each atom and its four immediate neighbours.

The conductivity of pure silicon and germanium

Given an electric field in a piece of pure silicon or germanium, free electrons and holes will both contribute to the flow of current. The contribution of the free electrons requires no explanation; that of the holes arises as follows. The holes tend to become occupied by those valence electrons which, in moving into them, submit to the thrust exerted by the field. Since the electrons move against the field, the holes move with it, i.e. they behave like positively charged particles.

The precise conductivity of a material depends (amongst other things) on the number of free electrons and holes in it. Now, the generation of free electrons and holes is continuous but, at the same time, recombination, that is, the filling of holes by free electrons, is also taking place. In the equilibrium state, the number of charge-carriers generated and the number recombined per unit time and volume are exactly equal. The respective concentrations n (for negative) and p (for positive) of free electrons and holes, identical in pure silicon or germanium, are invariably so small as compared with the concentration of atoms that the rate of generation of charge carriers is not noticeably affected by the number of valence electrons already released; hence the rate of generation may be considered as a constant (C). On the other hand, the number of recombinations is proportional to the concentration of free electrons as well as to the concentration of holes, and may therefore be defined as cnp, c being a constant. Accordingly, the equilibrium values of n and p are governed by the following equations:

$$c\,n\,p = C \quad \ldots\ldots\ldots \quad (1)$$

and

$$n - p = 0, \quad \ldots\ldots\ldots \quad (2)$$

giving:

$$n = p = \sqrt{C/c} \quad \ldots\ldots\ldots \quad (3)$$

The constants C and c are both functions of the temperature; a rise in temperature will produce sharp increase in C, but only a relatively small increase in c, thus increasing the conductivity. This is a characteristic property of all semi-conductors.

The bond between valence electrons and nuclei is slightly weaker in germanium than in silicon; hence germanium is the better conductor at a given temperature. In diamond, the crystal structure of which resembles that of silicon, the valence electrons are bound so tightly to the nucleus that the number of charge-carriers generated is negligible. Diamond is therefore an insulator.

The conductivity of germanium, though high compared to diamond, is still quite low. At room temperature, pure copper is roughly 10 million times better conducting than pure germanium.

The effect of "impurities"

Germanium is adapted to the purpose of rectification by dissolving small quantities of certain elements ("impurities") in it. For example, arsenic is a suitable impurity. The arsenic atom, which can be substituted for a germanium atom in the germanium lattice, has five valence electrons instead of four. Hence one of these five electrons can find no place in the germanium valence-bond structure and will at first wander more or less at random in the vicinity of the arsenic atom. The latter, having lost one of its electrons, now has a non-compensated nuclear charge. Owing to the effect of the dielectric, the attraction between this positive nuclear charge and the fifth valence electron is rather weak. Even at room temperature, then, thermal vibrations

readily free the surplus electron so that it is in no way distinguishable from others released from valence bonds.

Impurities of the same type as arsenic, that is, those which can supply a free electron, are known as donors. Other donor elements (all pentavalent) are phosphorus, antimony and bismuth.

The addition of a trivalent element, such as boron or indium, creates an entirely different situation. One of the bonds is then deficient of one electron; hence there is a vacant site into which an electron from an adjoining bond can move. The energy level of the electron thus transferred is somewhat higher in the new site than in a normal bond, owing to the fact that here its charge is not compensated by a positive charge in the nucleus, but this extra energy is readily supplied by thermal agitation. A valence electron moving into the vacant site in a boron atom leaves a hole behind. Again by reason of the dielectric, the attraction between the hole and the negative boron ion is weak. Hence the hole easily diffuses away and so prevents the extra electron captured by the boron atom from returning. Eventually, of course, another hole will approach the boron atom, thus enabling the captive electron to escape, but on an average the electron will remain in this energetically less favourable position for some time.

Because they capture electrons, trivalent impurities are described as acceptors. Acceptors, then, promote the generation of excess holes.

Germanium incorporating both donors and acceptors

A type of germanium with which we are very often concerned is that in which both donors and acceptors are present. Contrary to what may be expected, the numbers of free electrons and holes do not both increase, owing to the fact that these different carriers recombine in the manner already described. We shall now go more deeply into this question.

The mechanism of the direct generation and recombination of free electrons and holes as a result of thermal agitation in the crystal, which was tacitly assumed to be the only active mechanism in the case of pure germanium, is likewise active when both donor and acceptor impurities are present. However, other mechanisms of generation and recombination may be present. Now a theorem of statistical mechanics states that, if a phenomenon is promoted simultaneously by different independent mechanisms, the resulting equilibrium concentration will be independent of the number or nature of these mechanisms [2]). (As a possible second mechanism we have, for example, a process of generation in which a valence electron is first captured by a donor which has emitted its own excess electron, and then released, recombination taking place via the same intermediate stage, but in the opposite direction.) The existence of more than one active mechanism will affect the speed with which equilibrium is attained but not the equilibrium itself (catalytic action).

Since the concentrations of the impurities are relatively very low (e.g. 1 impurity centre per 100 million atoms) they do not for practical purposes affect the direct processes of generation and recombination. Hence these direct processes are indeed independent, and the equilibrium concentrations can be computed on the basis of this mechanism alone.

Accordingly, formula (1) is still valid. We can re-write this equation:

$$p\,n = C/c = N_i^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where N_i represents the equivalent values of n and p in pure germanium (suffix i for "intrinsic").

Equation (2), on the other hand, takes a different form. At room temperature, almost all the donors have produced electrons, and almost all the acceptors holes. Accordingly, the donors and acceptors (concentrations N_d and N_a respectively) are incorporated in the crystal lattice as fixed positive and negative charges, the combined charge-density of which is $N_d—N_a$. The holes and free electrons (motile charges) give a charge-density $p - n$, and, since the crystal is neutral, we have:

$$p—n = —(N_d—N_a) \quad . \quad . \quad . \quad . \quad (5)$$

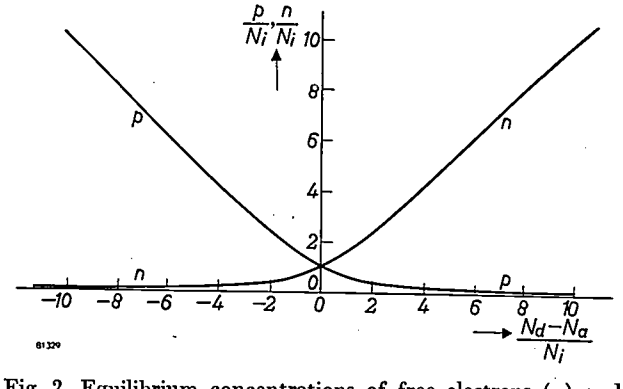


Fig. 2. Equilibrium concentrations of free electrons (n) and holes (p) in a semi-conductor, plotted as a function of the difference in concentration of donors and acceptors ($N_d—N_a$). The concentration (N_i) of free electrons in the pure semi-conductor ($N_d = N_a = 0$) is taken as the unit of concentration for both co-ordinates.

[2]) The proposition is based on the "principle of detailed balancing".

Formulae (4) and (5) establish p and n completely as functions of N_i and $N_d - N_a$. In *fig. 2*, p and n are plotted against the excess of donors $N_d - N_a$. For both co-ordinates, N_i is taken as unity. Germanium with an excess of donors produces a high concentration of free electrons and a low concentration of holes. It conducts primarily by electrons, and is therefore known as n-germanium. On the other hand, germanium with an excess of acceptors produces a high concentration of holes and a low concentration of free electrons; it conducts mainly by holes (which behave like positive charge-carriers) and is therefore known as p-germanium.

The p-n transition zone

Let us now consider the case of a germanium crystal in which the concentrations of donors and acceptors are so distributed as to create a localized region of transition from p-germanium to n-germanium. The associated variation of $N_d - N_a$, i.e. the difference in the concentrations of donors and acceptors, at such a "p-n junction" will be seen in *fig. 3a*. In the following discussion, it is assumed that all the quantities involved depend solely upon the x co-ordinate. Fig. 3*b* shows the distribution of the concentrations of free electrons and holes corresponding to the local values of $N_d - N_a$ given in fig. 2. This curve is not consistent with a state of equilibrium. A concentration gradient of free electrons, and a concentration gradient of holes is set up in the transition zone. These gradients produce diffusion currents of free electrons travelling to the p-region and of holes travelling to the n-region, thus creating a deficit of negative charge in the n-region and an excess of negative charge in the p-region. Accordingly, a region of positive space-charge is formed on the right-hand side of the junction and a region of negative space-charge on the left-hand side, in other words an electric double layer, roughly as shown in fig. 3*c*, is formed. This gives rise to an electric field, which drives the free electrons back to the n-region, and the holes back to the p-region (fig. 3*d*). Equilibrium is reached when the electric field is just strong enough to prevent any further diffusion of charges. The above-mentioned electric double layer sets up the potential difference between the n and p regions, shown in fig. 3*e*.

An expression for this potential difference, which will be found very useful when we discuss the process of rectification, will now be formulated.

The density w_n of the free electron stream in the positive x-direction may be expressed as:

$$w_n = -\mu_n E n - D_n \frac{dn}{dx} \quad . \quad . \quad . \quad . \quad . \quad (6)$$

The first term in the right-hand part of this equation represents the contribution of the electric field (E). This contribution is, of course, proportional to the field and to the concentration (n) of the free electrons. The proportionality constant μ_n is known as the mobility of the electrons; the negative sign prefixed to this term is due to the negative electron charge.

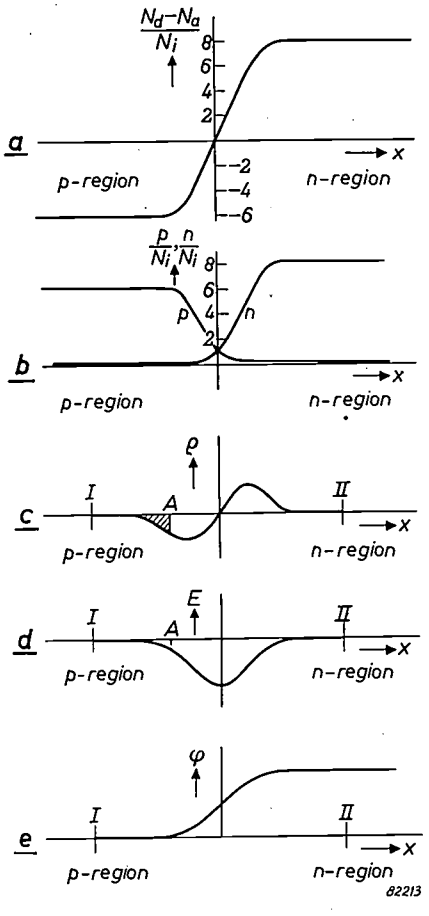


Fig. 3. *a*) Variation of the difference in concentration ($N_d - N_a$) between donors and acceptors at a p-n junction. In the p-region $N_d - N_a < 0$ (excess of acceptors), whereas in the n-region $N_d - N_a > 0$ (excess of donors).
b) Variation of the concentrations of free electrons and holes, n and p respectively, derived from fig. 2, for the variation of $N_d - N_a$ shown in (*a*). In fact, however, diffusion considerably modifies this distribution: the concentrations of both holes and free electrons actually vary more gradually, owing to the diffusion of free electrons to the p-region and holes to the n-region.
c) Space charge density arising from the diffusion of free electrons to the p-region and holes to the n-region. This space charge distribution forms an electric double layer. The hatched area represents the charge which has crossed section A in the p-region.
d) Variation of the electric field E within the electric double layer indicated in (*c*). The hatched area in (*c*) is a measure of the electric field at section A.
e) Potential at a p-n junction. The electrostatic potential relative to an arbitrary zero level is represented by φ.

The second term in the above equation represents the contribution of the diffusion; this is proportional, but opposite, to the concentration gradient dn/dx.

The proportionality constant D_n is the diffusion constant of the free electrons.

In the case here considered (no applied voltage across the crystal), $w_n = 0$ in the stationary state; from (6), then, we have:

$$E = -\frac{D_n}{\mu_n}\frac{1}{n}\frac{\mathrm{d}n}{\mathrm{d}x} \quad \ldots\ldots\ldots \quad (7)$$

If φ be the electric potential, $E = -\mathrm{d}\varphi/\mathrm{d}x$. Considering two sections I and II, one on each side of the transition zone, that is, one in the p, and one in the n germanium (see fig. 3c), and taking n_I and n_{II} as the respective equilibrium concentrations of free electrons at these sections, we find the potential difference φ_0 between the n-zone and the p-zone by integrating (7) from I to II, as follows:

$$\varphi_0 = -\int_I^{II} E\,\mathrm{d}x = \frac{D_n}{\mu_n}\int_I^{II}\frac{1}{n}\,\mathrm{d}n = \frac{D_n}{\mu_n}\ln\frac{n_{II}}{n_I} \quad . \quad (8)$$

Here, then, we have the desired expression for the potential difference.

Similarly, but taking the density of the hole current as a starting point, we deduce that:

$$\varphi_0 = \frac{D_p}{\mu_p}\ln\frac{p_I}{p_{II}} \quad \ldots\ldots\ldots \quad (9)$$

This expression is equivalent to (8) by reason of the fact that, from (4), $p_I n_I = p_{II} n_{II}$ and, in general, $D_n/\mu_n = D_p/\mu_p$. The latter equation is known as the Einstein relation, since it was he that first gave a general proof for it.

Putting

$$\mu_n/D_n = \mu_p/D_p = a,$$

(8) and (9) may be written

$$\frac{n_{II}}{n_I} = \frac{p_I}{p_{II}} = \mathrm{e}^{a\varphi_0} \quad \ldots\ldots \quad (10)$$

To calculate the variation of concentrations n and p in the transition zone as a function of x, and so establish the thickness of this zone, cannot be done in a simple way. With the aid of Poisson's equation, however, it is possible to make a qualitative analysis of the process whereby the diffusion and the electric field reach equilibrium, and so to ascertain that the thickness of the transition zone decreases according as the conductivity increases.

If ϱ be the charge density, ε_0 the dielectric constant of a vacuum, and ε_r the relative dielectric constant of germanium, Poisson's equation in the one-dimensional case here considered is:

$$\frac{\mathrm{d}^2\varphi}{\mathrm{d}x^2} = -\frac{\varrho}{\varepsilon_0\varepsilon_r}.$$

Since $E = -\mathrm{d}\varphi/\mathrm{d}x$, the above equation may also be written:

$$\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{\varrho}{\varepsilon_0\varepsilon_r}.$$

Hence:

$$E(x) = \int_{-\infty}^{x}\frac{\varrho}{\varepsilon_0\varepsilon_r}\,\mathrm{d}x. \quad \ldots\ldots \quad (11)$$

Accordingly, the hatched area in fig. 3c is a measure of the field strength at section A. If the space charge regions expand, as they will if the diffusion predominates over the electric field, the size of this hatched area will increase, since it represents the charge which has traversed the cross-section A; hence the field strength at A will increase (see (11)). On the other hand, the concentration gradients will decrease, since the variation of the hole concentrations and those of the free electrons, will become more and more gradual. The diffusion will decrease as these gradients decrease. Thus the electric field will gain and the diffusion diminish until a state of equilibrium is established. Given a high conductivity, a small field will be sufficient to maintain a steep concentration gradient in equilibrium. The charge layer will then be quite thin. Similarly, it is seen that if there is a difference in conductivity as between the region on the left and that on the right of the p-n junction, the spread of the space-charge region will be smallest in the zone of high conductivity. In practice, the thickness of the space-charge regions at p-n junctions is of the order of 1 μ.

Rectification at a *p-n* junction

Hypothetical p-n junction in the absence of generation and recombination

It is an apparently incongruous fact that *without the generation and recombination of free electrons and holes, the flow of current in either direction across a p-n junction would be impossible.* This will now be explained more fully, after which it is only a short step to an explanation of the rectification at a p-n junction.

First let us consider the case of a material containing charge carriers of only one type, which are densely concentrated on the left, and sparsely on the right, of a given plane. Here we have simply a good conductor adjoining a poor one. As is well known, when a voltage is applied, a stationary state is established, the condition for continuity of the current flow being satisfied by virtue of the fact that in each conductor the electric field strength is inversely proportional to the conductivity, that is, weaker in the good, than in the poor conductor. Such a state arises from the accumulation of a space charge at the boundary plane between the conductors. This space charge is responsible for the difference in field strength.

However, we are more interested in the case involving two types of charge carrier, i.e. holes and free electrons. Since we exclude the processes of generation and recombination, the electrons and holes must satisfy the continuity condition separately when a current flows, in the stationary state. From the point of view of the holes, the essence of this condition is that the field must be weaker in the

p-germanium than in the n-germanium, whereas the condition for a continuity for the flow of free electron current is precisely the opposite, i.e. that the field be stronger in the p-germanium than in the n-germanium. The only state in which this contradiction does not arise is that in which no current flows and the field strength outside the transition zone is zero. Accordingly, this state must prevail. The whole of the applied potential difference will then be across the p-n junction. As the field strength outside the p-n junction is zero, its gradient dE/dx is also zero and since $\varrho = \varepsilon_0 \varepsilon_r \, dE/dx$ (Poisson) holds good for the space charge density, this signifies that outside the junction the crystal is electrically neutral ($\varrho = 0$).

In the above argument it is assumed by implication that in the stationary state, and at a given distance from the p-n junction, an electric field alone will be responsible for any charge-displacement which may take place, that is, that no concentration gradients, which would give rise to diffusion, are present. This supposition is plausible in itself; hence we do not propose to furnish the proof. Let it suffice to say that a situation in which the concentration gradients do not approach zero as the distance from the p-n junction increases, is found to be incompatible with Poisson's equation, and therefore inconceivable.

The p and n concentrations in the absence of an applied voltage

Equation (7) holds good when no current flows. Accordingly, if the potential difference between n and p germanium be increased by an amount $\Delta\varphi$ through the application of a voltage, integration of formula (7) from section I to section II should produce the result $\varphi_0 + \Delta\varphi$. It can only do so if the equilibrium concentrations of the free electrons and holes change. We shall employ dashed symbols to distinguish the modified concentrations from the original ones.

Integration of equation (7) produces:

$$\varphi_0 + \Delta\varphi = \frac{D_n}{\mu_n} \ln \frac{n_{II}'}{n_I'},$$

which, with the aid of (10), can be put in the form:

$$\frac{n_{II}'}{n_I'} = \frac{n_{II}}{n_I} e^{+a\Delta\varphi} \quad \ldots\ldots\ldots \quad (12)$$

A similar formula holds good for the modified concentrations of holes:

$$\frac{p_{II}'}{p_I'} = \frac{p_{II}}{p_I} e^{-a\Delta\varphi} \quad \ldots\ldots\ldots \quad (13)$$

Moreover, it follows from the neutrality condition that the changes in the concentrations of the free electrons and holes must be equal.

In the case of p-germanium, then:

$$n_I' - n_I = p_I' - p_I, \quad \ldots\ldots \quad (14)$$

and in that of n-germanium:

$$n_{II}' - n_{II} = p_{II}' - p_{II}. \quad \ldots\ldots \quad (15)$$

Equations (12) to (15) determine the four modified equilibrium concentrations. These equations are easily solved. In *fig. 4*, the concentrations as calculated in a given case are plotted as a function of $\Delta\varphi$. It is seen that a given increase in the potential difference between the n and p regions (say, $a\Delta\varphi = +2$) produces only a small drop in concentration. On the other hand, an equivalent decrease in this potential difference ($a\Delta\varphi = -2$) produces a very considerable rise in concentration. This is by no means surprising, because the values of the concentrations have a natural lower limit, that is, they can not fall below the zero level of the so-called minority charge carriers (i.e. the free electrons in the p-region and the holes in the n-region). On the other hand, increases in concentration are not so limited. As

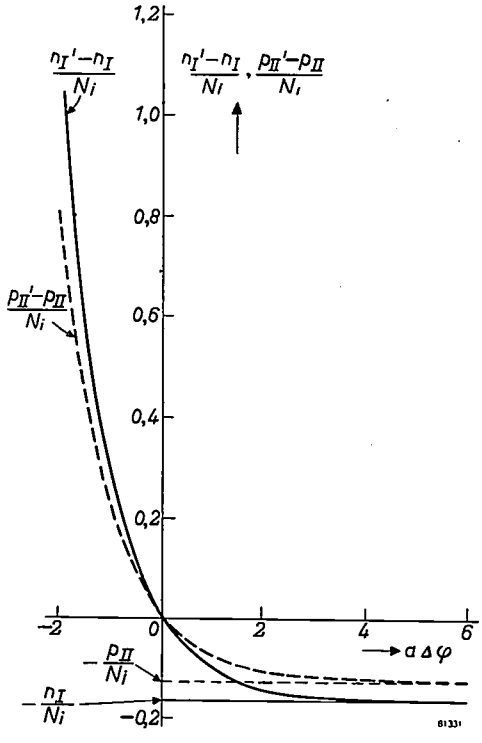


Fig. 4. The variation in concentration of holes and free electrons in the p-region (solid line: $p_I' - p_I = n_I' - n_I$) and in the n-region (dotted line: $p_{II}' - p_{II} = n_{II}' - n_{II}$), respectively, as functions of $\Delta\varphi$, that is, of the change in potential difference between the n and p regions due to an applied voltage. These curves refer to the p-n junction represented in fig. 3. At room temperature $a \approx 40$ volt^{-1}; hence $a\Delta\varphi \approx \pm 40$ when $\Delta\varphi = \pm 1$ Volt. For the ordinate, N_i is again taken as the unit of concentration.

will be explained later, this difference in behaviour arising from a difference in the polarity of $\Delta\varphi$ is responsible for the rectification at a p-n junction.

For not too large negative values of $a\Delta\varphi$, the changes in concentration conform to a simple power of e. Eliminating n_{II}' from (12) with the aid of (15), we write:

$$n_I' = n_I\, e^{-a\varphi\Delta}\left(1+\frac{p_{II}'-p_{II}}{n_{II}}\right).$$

Now as n_{II} is a majority carrier, $n_{II} \gg N_i$. Hence for negative values of $a\Delta\varphi$ which are not so large as to make $(p_{II}'-p_I)/N_i$ very much larger than unity (see fig. 4), $(p_{II}'-p_I)/n_{II}$ may be neglected in comparison to unity. In that case we have $n_I' = n_I\, e^{-a\Delta\varphi}$ and hence:

$$n_I'-n_I = p_I'-p_I = n_I(e^{-a\Delta\varphi}-1). \quad . \quad . \quad . \quad (16a)$$

Similarly, we find that:

$$p_{II}'-p_{II} = n_{II}'-n_{II} = p_{II}(e^{-a\Delta\varphi}-1) \quad (16b)$$

Physical picture of non-conductance of the crystal

The process leading to the state discussed in the preceding section may be envisaged in the following manner. If a voltage be applied to the germanium then, initially, a current will flow which in the p-region will be composed almost entirely of holes, and in the n-region almost entirely of free electrons. To avoid the complications associated with the end contacts of the crystal, it is assumed that the applied voltage is so regulated as to maintain a constant potential difference $\varphi_0 + \Delta\varphi$ between two cross-sections II and I, situated at positions intermediate between the p-n junction and the end contacts in the n and p regions respectively, (*fig. 5*). If the spontaneous field, so that a nett surplus of holes flows across the junction. In a similar way, a nett surplus of electrons flows in the opposite direction. As already mentioned, however, this flow cannot persist: it appears that the holes entering the junction from the p-region and the electrons entering it from the n-region attenuate the double-layer at the junction (fig. 3c) and so reduce the spontaneous potential difference across the junction until there is no longer any potential drop in the p and n regions outside (fig. 5a). The field outside the junction is therefore zero, for the whole potential drop $(\varphi_0+\Delta\varphi)$ now occurs across the junction. At the same time another, important effect takes place. The nett surplus of holes flowing across the junction from the p-region diffuse into the n-region and so increase the

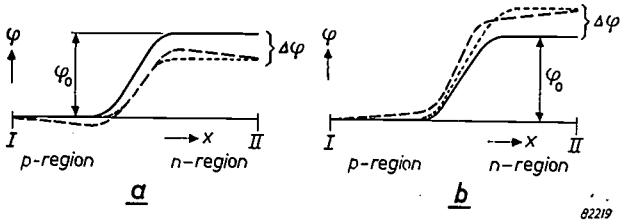


Fig. 5. The potential as a function of position at a p-n junction with an applied voltage, in the absence of generation and recombination. In both diagrams, the curve drawn as a solid line refers to the state before the voltage is applied; the potential difference between the n and p regions is then spontaneous (φ_0). The curves drawn as broken lines apply to the instant when the voltage is applied, and those drawn as dotted lines refer to the subsequent equilibrium state in which no current flows. A potential level of zero is assumed in section I. In (a) the potential of the n-region is reduced, and in (b) it is increased, by the applied voltage.

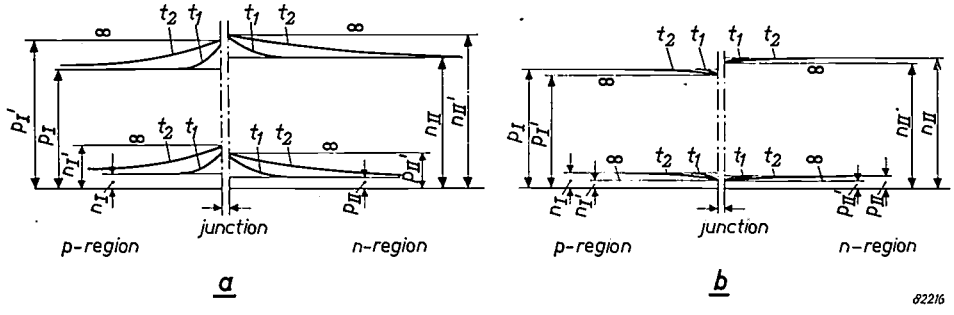


Fig. 6. Three successive stages, viz. at times t_1 and t_2, and at the end of an infinitely long period, during the change in the concentrations of holes and free electrons produced by an applied voltage. Again, it is assumed that generation and recombination do not occur. The changes in concentration develop outwards from the p-n junction, thus gradually reducing the concentration gradients to zero. If generation and recombination do occur, a state roughly equivalent to one of the intermediate stages will remain.
a) The applied voltage reduces the potential of the n-region relative to that of the p-region, and drives holes from the p-region to the n-region.
b) The same voltage applied in the reverse direction. The decreases in concentration then produced are very much smaller than the increases in case (a).

$\Delta\varphi$ be negative, a field exists which urges holes in the p-germanium towards the junction and those in the n-germanium away from it. At the junction itself, the applied field upsets the previous equilibrium between diffusion of holes $(p \rightarrow n)$ and the reverse flow due to concentration of holes there. As this increased concentration penetrates through the n-region, however, the concentration gradient which drives it gradually diminishes, until all diffusion ceases (fig. 6a). The invading holes, form a space charge in the n-region,

but this is rendered ineffective by an influx of free electrons from the opposite direction; hence the concentrations of the different charge carriers increase simultaneously to the values already calculated (fig. 4). A similar process takes place in the *p*-region, where the free electrons now take the part played by the holes in the *n*-region.

Conversely, by reversing the applied voltage we strengthen the existing electric double layer. The potential jump across the *p-n* junction then increases, again reducing the field outside the transition zone to zero (fig. 5*b*). The shortage of holes in the transition zone resulting from the large-scale migration of holes to the *p*-region now produces in the *n*-region a diffusion current of holes flowing towards the *p-n* junction, thus causing a decrease in the hole-concentration in the *n*-region. This decrease penetrates throughout the *n*-region and so reduces the gradient (fig. 6*b*). Eventually, then, the gradient is eliminated altogether and the diffusion current therefore stops. A simultaneous decrease in the concentration of free electrons prevents the formation of an effective space charge. Again, a similar process takes place in the *p*-region, where the holes and free electrons exchange roles.

An important feature of these processes is that the changes in concentration build up in the crystal from the transition zone.

Conduction as a result of generation and recombination

It is by reason of the generation and recombination of free electrons and holes that in reality a current can flow when a voltage is applied across the crystal. As we have already seen, when $\Delta\varphi$ is negative the concentration of holes and free electrons increase, the origin of this increase being the *p-n* junction (fig. 6*a*). From the start of this increase, the rate of recombination will exceed the rate of generation. This is incompatible with a non-conducting stationary state, since the fact that electrons and holes are eliminated more rapidly than they are generated necessitates a constant influx of these charge carriers, that is, a flow of current.

For example, let us consider what happens in the part of the *n*-region close to the *p-n* junction. Holes diffuse into the *n*-region, creating an enhanced hole-concentration there. As this enhanced concentration penetrates deeper into the *n*-region, so the number of holes eliminated per second by recombination increases, whereas the propagating concentration gradient, and therefore the supply of holes, decreases.

This process leads to a state in which the relatively high concentration extends so far into the *n*-region that the influx and elimination of holes and free electrons are precisely in equilibrium. The necessary supply of free electrons is maintained by virtue of the fact that the electric field outside the *p-n* transition zone does not now disappear completely, contrary to the case when generation and recombination were supposed not to occur. Some of these electrons recombine with holes in the *n*-region and others diffuse into the *p*-region, where they recombine with holes newly introduced by the electric field. The holes so introduced which evade recombination, pass the *p-n* junction and so maintain the diffusion current of holes in the *n*-region.

Similarly, a positive $\Delta\varphi$ produces a stationary state in which regions of relatively low concentration extend from the *p-n* junction into the *p* and *n* regions in this case, the rate of generation in these regions exceeds the rate of recombination. This state corresponds roughly to one of the intermediate stages shown in fig. 6*b*.

In a stationary state of this kind, a current must flow to provide an outlet for the excess free electrons and holes generated.

Magnitude of the current; rectification

Whether $\Delta\varphi$ be positive or negative, minority charge carriers are propagated by diffusion, and majority charge carriers by the electric field. Of course, the electric field also affects the minority carriers, but their concentration is so much lower than that of the majority carriers that an electric field producing an appreciable current of majority carriers will produce virtually no current of minority carriers. For the latter, then, diffusion is the only effective means of propagation. For this reason, it is simpler to consider only the minority charge carriers and, in fact, this enables conclusions to be reached without considering the more complex behaviour of the majority carriers.

The current in each section is the sum of the free electron and the hole components; hence this is true, for example, at the section of the *p*-region adjoining the transition zone. Now, the transition zone itself is invariably so narrow that the number of charge carriers generated and recombined in it is negligible. Accordingly, the hole current (produced by the field) in the boundary plane at the *p*-side is equal to the hole current (produced by diffusion) at the *n*-side of the junction. Hence the total current (= hole current + electron current) may be expressed as the sum of the diffusion currents of the minority charge carriers (electrons in *p*-region and holes in *n*-region) at the two boundary planes of the transition zone.

A diffusion current is proportional to the concentration gradient. If deviation from the equilibrium concentration is large, the rates of elimination and generation of the charge carriers will also be large and hence a steep concentration gradient will be set up. The diffusion current will then be correspondingly strong. Now, the curves shown in fig. 4 are still valid, to a first approximation, for the changes in concentration in the boundary planes of the transition zone. Accordingly, the diffusion currents occurring in these two boundary planes will be appreciably stronger at a given negative value of $\Delta\varphi$ than at the equivalent positive value, i.e. rectification takes place. In the former case, which is the "forward direction" the applied voltage drives holes from the p-region and free electrons from the n-region to the p-n junction (see fig. 5a). When the applied voltage is reversed ($\Delta\varphi$ positive) holes are driven from the n-region and electrons from the p-region; conduction is then much smaller, and in the "reverse direction".

A quantitative analysis of the process can also be made without undue difficulty. The effect of the electric field on the minority charge carriers may be neglected. Hence from the point of view of the free electrons in the p-region, equation (6) is reduced to:

$$w_n = -D_n \frac{\mathrm{d}n}{\mathrm{d}x}. \quad \ldots\ldots\ldots \quad (17)$$

Given two sections, a distance $\mathrm{d}x$ apart, in the p-region (*fig.* 7), we find that the number of free electrons entering this element through the left-hand boundary plane exceeds the number leaving it through the right hand boundary plane, per second and per unit area, by an amount $-(\mathrm{d}w_n/\mathrm{d}x)\mathrm{d}x$, which, from (17), is equal to $D_n(\mathrm{d}^2n/\mathrm{d}x^2)\mathrm{d}x$. Hence the excess influx per unit volume per second is $D_n\mathrm{d}^2n/\mathrm{d}x^2$.

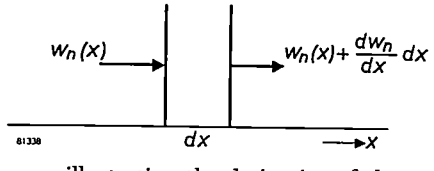


Fig. 7. Diagram illustrating the derivation of the expression for the concentration gradient of minority charge carriers.

In equilibrium, this difference is equal to the numerical surplus of free electrons eliminated (by recombination), over free electrons added (by generation), per second and per unit volume. For the present, let us consider only the direct processes of generation and recombination; the number of carriers generated is then constant, viz C (see earlier), and the number of recombinations is $cnp = c\,(n_I + \Delta n)\,.\,(p_I + \Delta p)$, where Δn and Δp represent the deviations from the respective equilibrium concentrations n_I and p_I in the p-region. Bearing in mind that n_I (minority charge carriers) is very low, and that $cn_Ip_I = C$ (see (1)), we find that the number of recombinations exceeds the number of carriers generated by $cp_I\Delta n$. Other mechanisms of generation and recombination likewise produce a surplus of recombinations, proportional to Δn; in general, then, we may define this surplus as $b\Delta n$, b being a constant. Since, furthermore, $\mathrm{d}^2n/\mathrm{d}x^2 = \mathrm{d}^2\Delta n/\mathrm{d}x^2$, we have:

$$D_n \frac{\mathrm{d}^2\Delta n}{\mathrm{d}x^2} = b\Delta n.$$

The boundary conditions of this differential equation are: $\Delta n = 0$ for $x = -\infty$, and (see (16a)) $\Delta n = n_I\,(e^{-a\Delta\varphi} - 1)$ for $x = 0$ (assuming that the p-n junction is located at $x = 0$). The solution is:

$$\Delta n = n_I\,(e^{-a\Delta\varphi} - 1)\,e^{x/L_n},$$

where $L_n = \sqrt{D_n/b}$ is known as the diffusion length for free electrons.
For $x = 0$, we have:

$$\left(\frac{\mathrm{d}n}{\mathrm{d}x}\right)_{x=0} = \left(\frac{\mathrm{d}\Delta n}{\mathrm{d}n}\right)_{x=0} = \frac{1}{L_n} n_I\,(e^{-a\Delta\varphi} - 1).$$

Hence the diffusion current of free electrons at the p-n junction is (see (17)):

$$w_n = -n_I \frac{D_n}{L_n}\,(e^{-a\Delta\varphi} - 1).$$

Similarly, we find that the diffusion current of holes at the p-n junction, is:

$$w_p = p_{II} \frac{D_p}{L_p}\,(e^{-a\Delta\varphi} - 1).$$

Accordingly (if $-e$ is the charge on the electron), the total current (i) is:

$$i = ew_p - ew_n = \left(\frac{D_p}{L_p} p_{II} + \frac{D_n}{L_n} n_I\right) e\,(e^{-a\Delta\varphi} - 1). \quad . \quad (18)$$

When $\Delta\varphi = +\infty$, the current i approaches the saturation value $-i_s$ of the leakage current (the minus sign is added to denote current in the reverse direction i.e. in the negative x-direction):

$$-i_s = -\left(\frac{D_p}{L_p} p_{II} + \frac{D_n}{L_n} n_I\right) e,$$

which enables us to write formula (18):

$$i = i_s\,(e^{-a\Delta\varphi} - 1).$$

The formation of p-n junctions

Such p-n junctions as we are now considering may occur in several ways. They sometimes occur as a result of a non-uniform distribution of impurities during solidification of a melt of germanium. Again, they may be formed deliberately in many ways, e.g. by adding the desired impurities suddenly to the molten germanium during the process of crystallization.

For some purposes, crystal diodes are made by establishing a p-n juncton, such as discussed above, in a single crystal of germanium. A more common form of the crystal diode, however, is the point con-

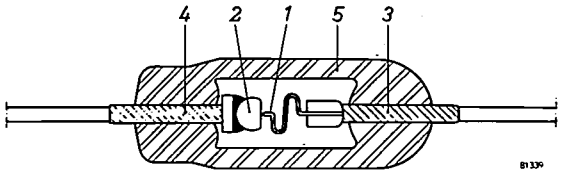


Fig. 8. Construction of a point-contact germanium diode of the latest type. The pointed contact spring *1* presses against the germanium crystal *2*. The contact spring and the crystal are secured to supporting pins *3* and *4*, sealed into a glass bulb *5* as protection against the atmosphere.

tact diode (*fig. 8*). It is now generally accepted that rectification in a modern point-contact diode (fig. 8) is also attributable to a *p-n* junction [3]). After the application of the point contact to the germanium crystal, several current pulses of high intensity are passed through the diode (a process known as "forming" the contact). This enables impurities to diffuse from the crystal surface or out of the contact spring into the crystal and so form a *p-n* junction around the contact point. One may also picture the process as an effect of the current pulses on the local distribution of donors and acceptors near the point, with the formation of a *p-n* junction.

Even without the forming process, any contact between a metal and a semi-conductor will act as rectifier. No entirely adequate explanation of the mechanism responsible for this rectification has yet been given, and we shall not pursue the matter further here, since it plays only a minor role in the point-contact diode: once the contact has been "formed", rectification at the *p-n* junction so established will dominate over any other mechanisms. *Fig. 9* shows the rectification characteristics of a point-contact germanium diode before and after "forming".

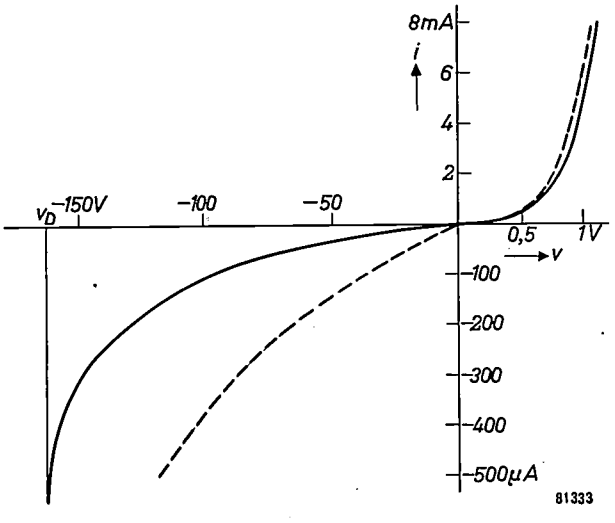


Fig. 9. The rectification characteristic (current i as a function of the applied voltage v) of a point-contact germanium diode (type OA 71). v_D is the breakdown voltage, at which $dv/di = 0$. The dotted curve is the characteristic of the same diode before the forming process. Note the difference in the scales used for the voltage in the forward and reverse directions and also for the forward and leakage currents.

As mentioned above, the point-contact is not the only form of *p-n* junction with rectifying properties. However, the capacitative effect of such a contact is small (stray capacitance of the order of one or two pico-farads); hence point-contact diodes offer certain advantages for high-frequency operation. On the other hand, diodes with a *p-n* junction of relatively large area (and consequently higher stray capacitance), are useful for other purposes, e.g. power operation at heavier currents.

The rectification characteristic

As will be seen from fig. 9, the leakage current increases with the negative bias until, at point v_D, the dynamic resistance dv/di of the diode is zero. The voltage to produce this effect is known as the breakdown voltage of the diode, since, unless the current be limited by a suitable ballast resistor, breakdown will occur when the voltage accros the diode reaches the value corresponding to v_D. The breakdown voltage varies between different types of diode; for example, $v_D \geqq 30$ V for the OA 60, whereas $v_D \geqq 120$ V for the OA 55. Germanium diodes with breakdown voltages higher than 300 V are occasionally produced, but so far no one has succeeded in producing diodes with this value in a normal production run.

The primary cause of breakdown is heat generated by the current flowing in the crystal. This liberates free electrons and holes, which reduce the internal resistance of the crystal and so increase the leakage current, thus generating more heat; hence still more free electrons are released and so the process continues until the internal resistance of the crystal is virtually zero. However, if the same voltage be applied in short pulses instead of continuously, this "avalanche" will not have time to develop. According to the above theory, then, the breakdown voltage v_D is higher under these conditions.

The breakdown voltage of Philips diodes are measured statically, that is, under the most unfavourable conditions. In practice, a certain safety margin must be preserved; this is accomplished by imposing two restrictions on the reverse voltage. These restrictions are that the reverse *peak* voltage must not exceed a certain value and that the *average* reverse voltage must not exceed another specified value.

In the forward direction, the increase in current with voltage is fairly gradual within the range up to about 0.5 V; beyond this range it is rapid and nearly linear (fig. 9). The current in the forward direction at 1 volt is usually given as the characteristic quantity for a diode; it represents the average increase in current per volt within the range between 0 and + 1 volt. At + 1 volt, of course, the actual slope of the characteristic in mA/volt is considerably greater than this average value.

[3]) See M. C. Waltz, Proc. Inst. Rad. Engrs. **40**, 1483 - 1487, 1952, and Makoto Kikuchi and Tomio Onishi, J. appl. Phys. **24**, 162-166, 1953.

The manufacture of point-contact germanium diodes.

The basic material is spectrochemically pure germanium dioxide (GeO_2), obtained as a by-product of certain zinc ores. It is reduced in a hydrogen atmosphere to pure germanium, which, if necessary,

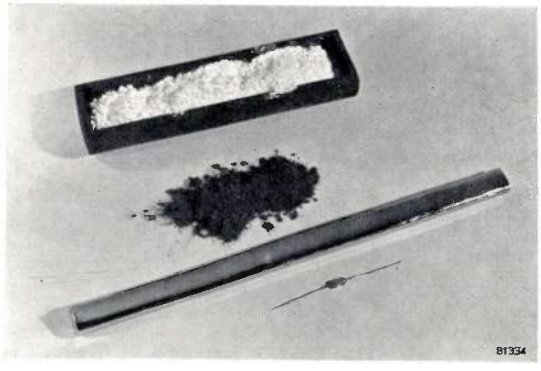


Fig. 10. A boat (of graphite) containing pure germanium dioxide powder. In front of the boat is a sample of the black, metallic germanium powder obtained by reducing the oxide. A rod of germanium metal, which is a single crystal, and a germanium diode are shown in the foreground.

can be further refined by fractional crystallization (*fig. 10*). At this stage, a pre-determined quantity of a suitable impurity may be added to, and melted with, the germanium. The next step is to divide the germanium into small pieces and solder them to nickel supporting-pins. Special measures are taken to prevent rectification at the contacts formed during this process. The top face of the crystal is then ground smooth; since this grinding seriously affects the crystal orientation of the top layer, the latter is removed by etching with hydrofluoric acid. The supporting pin carrying the crystal is then sealed into a glass bulb fitted with lead-in caps of fernico, and a similar pin carrying a thin spring of tungsten wire is inserted in the bulb from the opposite end so that the spring contact presses firmly against the crystal. The second pin is then also sealed into the bulb (see *fig. 11*). Either a piece of polycrystalline germanium, or a single crystal may be employed, according to the particular diode properties required. A considerable improvement of the rectification characteristic is effected by passing several strong current-pulses through the germanium ("forming", see above).

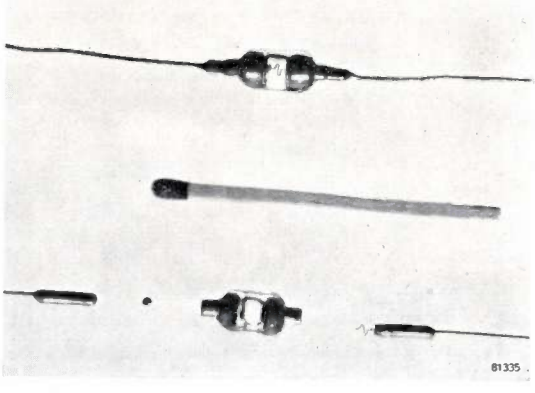


Fig. 11. The components of a point-contact germanium diode shown full size. Note the support-shanks, one of which carries a fine contact-spring. Between these shanks lie the glass bulb and the germanium crystal. An assembled diode is seen at the top of the photograph.

The process of soldering the supporting pins in the fernico lead-in caps of the bulb creates a special problem, owing to the fact that the etched germanium surface is readily contaminated by the vapour formed during this process. Such contamination would shorten the life of the diode. For this reason, a protective layer of plastic lacquer, easily pierced by the contact spring, is applied to the germanium surface immediately after the etching process. This layer satisfactorily precludes the above-mentioned effect without detracting from the quality of the diode.

However, in a new type of germanium diode recently introduced, the metal electrode-holders are sealed direct to the glass (fig. 8). The advantage of this method is that it avoids soldering, and so ensures an exceptionally high degree of stability. On the other hand, for technical reasons, the lead-in pins employed in this process must be thinner than those of the glass-fernico diodes already described; this decrease in pin-thickness slightly reduces the heat-dissipation from the crystal and so necessitates a slight lowering of the maximum limits for current and voltage.

In fact, the maximum permissible peak and average values of the reverse voltage for this new type of diode are limited to the fixed values 75% and 50%, respectively, of the breakdown voltage as statically measured.

Behaviour of germanium diodes at high frequencies

Germanium diodes are characterised by what is known as the "hole-storage" effect. We have already seen that in a germanium diode operated in the forward direction, a relatively high concentration of charge carriers occurs in the vicinity of the *p-n* junction. If the voltage applied to the diode be reversed (reverse direction), the leakage current so produced will at first greatly exceed the value indicated by the characteristic, owing to the fact that this relatively high concentration must be reduced to the lower concentration consistent with the normal leakage current (see *fig. 12*).

It can be deduced from fig. 12*b* that the average leakage current in the negative period increases with

the frequency of the a.c. voltage. In a circuit, then, any increase in this frequency will reduce the detection efficiency and the equivalent attenuation resistance[4]). This explains the unsatisfactory results obtained when a standard diode (type OA 50, 51, 53, 55 or 61) is employed, for example, as a video detector. In fact, a special diode (OA 60, and OA 70 in the new series) has been developed for such purposes. When so employed, this diode exhibits

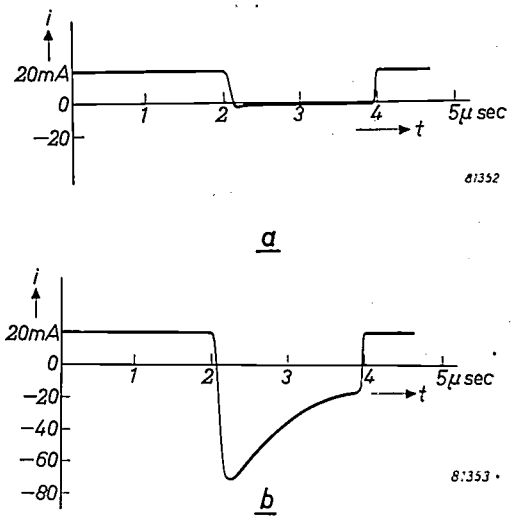


Fig. 12. "Hole storage" effect in two types of point-contact germanium diodes. The curves are taken from oscillograms. In both cases the forward current is 20 mA and the reverse voltage 20 V.
a) Diode showing virtually no hole storage effect.
b) Diode showing considerable hole storage. The leakage current fails to attain its static value even 2 μsec after the application of the reverse voltage. If the period of the voltage pulse in the reverse direction be short, the average value of the leakage current during this period will be high.

dynamic properties considerably better than those of the other, statically superior, diodes.

Difficulties from "hole storage" may also occur in the circuits of calculating machines, where it is important to ensure that only pulses of the required square shape are admitted.

[4]) See the following article on p. 225 of this issue: J. Jager, The application of germanium point-contact diodes.

Behaviour at high temperatures

As the temperature rises, more and more electrons are released from the normal bonds. These electrons, and the holes generated with them, increase the conductivity and so reduce the internal resistance of the diodes.

At temperatures in the region of 65-75 °C, the number of charge-carriers thus generated so far exceeds the number of those introduced artificially, as to virtually suppress the differences between standard diodes such as the OA 50, 51, 55 and 56. It is then necessary, of course, to lower the maximum limit for voltages in the reverse direction as compared with those valid at room temperature (that is, up to 25 °C). The maximum voltage ratings of diodes of the new series, i.e. OA 70, OA 71, and so on, are valid for an ambient temperature up to 60 °C; hence it will be necessary to reduce these ratings only if the ambient temperature exceeds 60 °C.

In general, the ambient temperature of the germanium should be kept as low as possible, and operation at temperatures above 60 °C is to be avoided, since rectification, although it does not cease entirely at such temperatures, cannot then be guaranteed.

Summary. This article discusses at some length the mechanism of rectification in germanium diodes. Pure germanium conducts by electrons and "holes", the latter behaving like positively charged particles. It is possible to change the concentrations of these different charge carriers appreciably by introducing certain foreign elements. The addition of pentavalent elements (donors) increases the number of free electrons and reduces the number of holes, thus producing what is known as *n*-germanium (conduction by negative particles). Conversely, the addition of trivalent elements (acceptors) increases the number of holes and reduces the number of free electrons. The material thus produced is known as *p*-germanium (conduction by positive particles). Given both a *p*-region and an adjoining *n*-region in a single germanium crystal, rectification will take place when a current flows through the crystal across this junction. The explanation of this effect is rather complex; it depends on a subtle interplay between the generation and recombination of electrons and holes and their movements as a result of diffusion and electric fields. It is nowadays generally accepted that rectification in point-contact diodes is also attributable to a *p-n* junction. The probable mechanism of formation of such a junction in a diode of this type is briefly outlined. Particulars are also given of the manufacture of point-contact diodes, their rectification characteristics, and their high-frequency and high-temperature performance.

THE APPLICATION OF POINT-CONTACT GERMANIUM DIODES

by J. JAGER.

621.314.632

Now that germanium diodes are becoming generally available, there is a marked tendency to substitute these small, convenient components for vacuum diodes wherever possible. While it is true that their special properties enable germanium diodes, and point-contact germanium diodes in particular, to be used advantageously for the rectification of relatively low voltages and currents, the vacuum diode is often retained, even in cases where a superficial examination suggests the use of a germanium diode. The purpose of the present article is to give an idea of the relative advantages and disadvantages of germanium and vacuum diodes.

Germanium and vacuum diodes

As compared with the vacuum diode, the germanium diode offers a number of conspicuous advantages, viz:

1) It contains no filament. This advantage is particularly valuable in apparatus where there is no other reason to employ heater power, or where a filament would necessitate the use of long supply leads. Also, of course, it eliminates the problem of the filament as a possible source of hum, as may occur in a vacuum diode when the impedance between the cathode and earth is high.

2) Germanium diodes can be incorporated direct in the wiring merely by soldering the two connecting wires to the appropriate points; only a few types of vacuum tube can be connected in this manner.

3) Germanium diodes are very small and light (weight about 1.1 grams) (*fig. 1*).

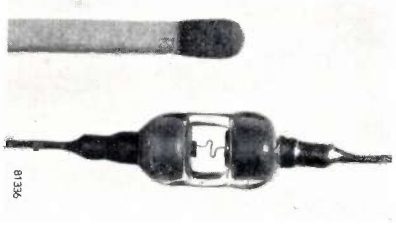


Fig. 1. The small size of point-contact germanium diodes can be judged by comparison with the match head in this photograph. Only two connections are necessary since there is no filament. The diode is mounted without a valve holder.

Features 2) and 3) show to the best advantage in circumstances where the diode is to be accommodated with other components in a relatively small space (e.g. in a coil can).

Apart from these three general advantages, which are direct results of the difference in design of germanium and vacuum diodes [1]), some of the electrical properties of the germanium diode are also superior, viz.:

4) Low shunt capacitance, i.e. about 1 pF, an outstanding feature of point-contact germanium diodes, which may be vitally important at high frequencies.

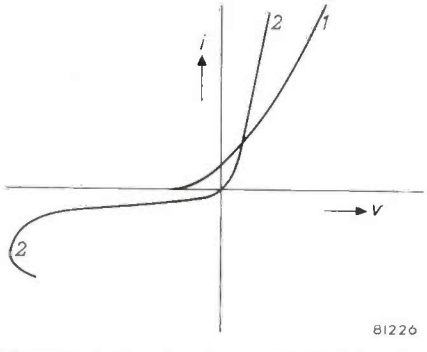


Fig. 2. Characteristics of a vacuum diode (*1*) and a germanium diode (*2*). A vacuum diode in a closed circuit is conductive in the forward direction when no external voltage operates in the circuit, and will even remain conductive in the presence of a small negative voltage (starting-current region). However, it carries no leakage current (current in the reverse direction). On the other hand, a germanium diode has a lower forward resistance and no starting-current region, but does exhibit a leakage current. Also, the differential resistance (dv/di) is zero at a certain negative voltage (vertical tangent to curve).

5) The internal resistance of a germanium diode operated in the forward direction is lower than that of a vacuum diode, even when the latter is specially designed to operate with a low load-resistance.

6) The germanium diode has no starting-current region (*fig. 2*), unlike the vacuum diode, and therefore requires no compensation for starting current in measuring-circuits.

The above summary is impressive and, in fact, these advantages will often turn the scale in favour of the germanium diode. However, this diode has one or two other features which compare unfavourably with those of its vacuum equivalent. They may be defined in the following manner:

a) Whereas a vacuum diode is not conductive in the reverse direction, the germanium diode carries an

[1]) J. C. van Vessum, The theory and construction of germanium diodes, p. 213 of this issue.

appreciable leakage current, particularly at relatively high voltages.

b) A vacuum diode, suitably designed, is able to withstand the application of a high negative voltage between anode and cathode, whereas the shape of the current-voltage characteristic ($i = f(v)$) of the germanium diode is such that, at a certain voltage, the differential resistance, that is, dv/di, is zero (fig. 2). In those germanium diodes which are most efficient in this respect, zero differential-resistance occurs at about −300 V.

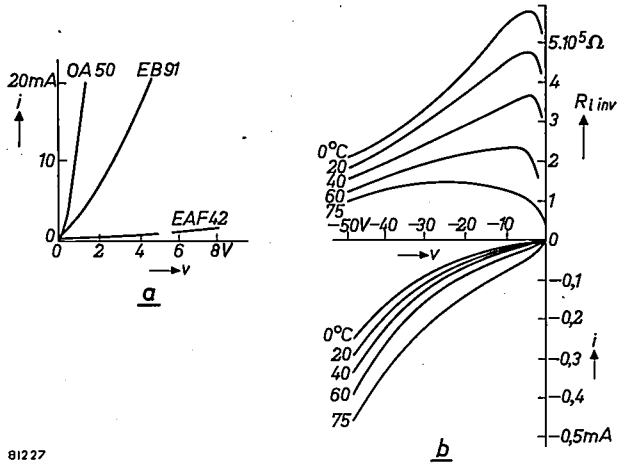


Fig. 3. *a*) Forward characteristics of a general-purpose germanium diode (type OA 50) and of two vacuum diodes, viz one half of an EB 91 double diode, designed to operate with a low load-resistance, and the diode section of a diode-pentode EAF 42, designed to operate with a high load-resistance.
b) The leakage current (i) of a germanium diode type OA 50 versus the applied voltage (v) at different ambient temperatures (bottom part of the diagram). Compare this with (*a*), but note the difference in the scales used in the two diagrams.
The top part of diagram (*b*) shows the ratio $R_{i\,inv} = v/i$, i.e. the leakage resistance, plotted as a function of voltage v.

c) The characteristic of a germanium diode depends, to a considerable extent in the forward direction but especially in the reverse direction, on the temperature. The characteristic is also dependent on the rate of change of voltage. This effect becomes important at high frequencies.

All the examples given in this article refer specifically to point-contact germanium diodes, although in referring to them we shall omit the words "point contact" for the sake of brevity.

Fig. 3a shows the forward characteristics of a germanium diode for general service, a vacuum diode designed to operate with a low load-resistance (e.g. 3 kΩ, as employed in the video detectors of television receivers), and a valve of the type used in ordinary radio receivers, where the load-resistance is usually much higher (i.e. about 0.5 MΩ).

The conductivity of the germanium diode in the reverse direction, unlike the two vacuum diodes, is considerable. Fig. 3*b* (bottom half) shows the reverse current of a germanium diode plotted against the voltage, for various ambient temperatures. The upper half of fig. 3*b* shows the corresponding inverse resistance (i.e. the resistance in the reverse direction).

Diodes are extensively used for the detection of amplitude-modulated signals. The differences between vacuum diodes and germanium diodes will now be discussed by a consideration of their applications. One or two practical examples of diodes designed for special purposes will also be discussed.

The performance of vacuum diodes and germanium diodes in detector circuits

The quality of a detector circuit is governed primarily by two quantities, i.e. the detection efficiency and the equivalent attenuation resistance. These quantities will now be discussed more fully, to ascertain how they are affected by the characteristics of the particular diode employed. The results so obtained may help us to decide in a given case whether a germanium diode should be employed, and, if so, which type is most suitable.

Detection efficiency

Fig. 4a shows a circuit of a type frequently employed in detectors. Here, the diode load comprises a resistance R and a capacitance C in parallel.

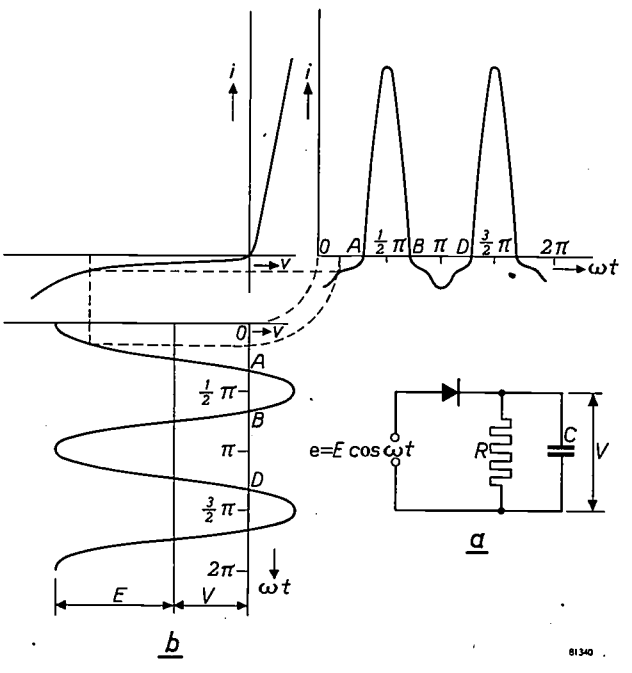


Fig. 4. *a*) Simple circuit for the detection of amplitude-modulated signals.
b) The current in a germanium diode operating in the circuit shown in (*a*), plotted as a function of the time (top right-hand side), as deduced from the signal voltage varying harmonically with time, $e = E \cos \omega t$ (bottom left), and the characteristic of the diode (top left). A leakage current flows in the diode in the interval BD. The D.C. voltage (V) across capacitor C adjusts itself until the ingoing and outgoing charges during a period are exactly equal. The charge-balance so established depends considerably on the discharge current through load-resistance R.

The signal voltage $e = E \cos \omega t$ is supplied by a generator, the internal resistance of which, for simplicity, is assumed to be zero. In reality, E varies (modulation); however, this variation is immaterial to the present discussion, and we may therefore consider E as a constant. Moreover, we shall assume that the time constant (RC) of the load is large compared with the period of the signal voltage; since the ripple on the capacitor voltage will then be small, we may also suppose this voltage (V), to be constant. Accordingly, the voltage across the diode, varying with time, is given by $e - V$. Fig. 4*b* shows the current in a diode, derived as usual from the current-voltage characteristic, for a given arbitrary value of V. As will be seen from this diagram, the charging of the capacitor takes place during a time interval represented by AB. The capacitor discharges a constant current of magnitude V/R through the load resistance; however, part of its charge is also extracted, during another interval BD, through the germanium diode. Of course, the voltage V stabilizes at such a value that the total charge entering the capacitor in one period is equal to that leaving it (charge-balance). Any increase in the leakage through the diode necessitates an increase in the charging-interval AB, which drives the diode further into the region of positive current and so reduces the D.C. voltage V produced across the capacitor. The maximum value which the voltage V can attain is the peak-value E of the applied A.C. voltage; this value would be reached if both the diode inverse resistance and the load resistance were infinitely high. The ratio V/E is commonly known as the detection efficiency, η_{det}. In the case of a germanium diode with an idealized characteristic, that is, two straight lines joining at the origin (*fig.* 5*a*), the detection efficiency is readily calculated by equating the ingoing and outgoing charges of the capacitor during one period. It is simpler to use the reciprocal resistances (conductances) in the calculation, instead of the resistances themselves; G then represents the conductance of the load resistance, and G_i and $G_{i\,\text{inv}}$ the conductances of the diode in the forward and reverse directions respectively. The calculation is straightforward and is not given here. It proves that η_{det} is a function of $(G + G_{i\,\text{inv}})/(G_i - G_{i\,\text{inv}})$. Since $G_{i\,\text{inv}}$ is invariably $\ll G_i$, $G_{i\,\text{inv}}$ may be neglected in the denominator. The calculation, the result of which will be seen in fig. 5*b* produces a qualitatively correct interpretation of the effect of the various associated quantities on the detection efficiency. (The straight-line approximation to the diode characteristic is too rough for use in practical calculations.) If G be high (that is, R low) in relation to the leakage $G_{i\,\text{inv}}$, the effect of leakage will be negligible. The detection efficiency increases as G decreases, but at the same time the detrimental effect of leakage becomes more and more significant.

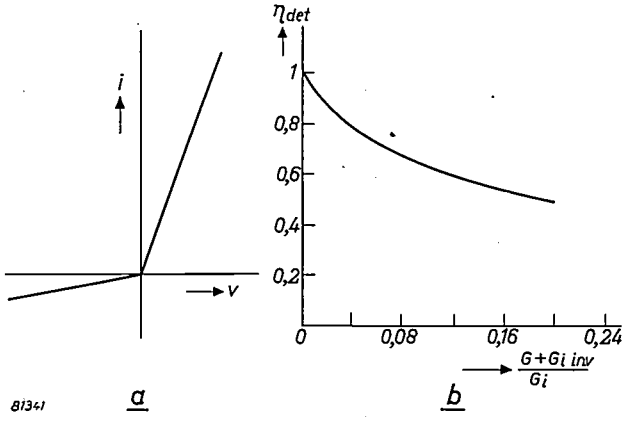


Fig. 5. *a*) Idealized characteristic of a germanium diode, that is, two straight lines meeting at the origin.
b) The detection efficiency η_{det} as calculated for a circuit such as that in fig. 4*a* containing a diode with an idealized characteristic as shown in fig. 5*a*. It is seen that in this case η_{det} is governed solely by the quantity plotted on the abscissa, that is, $(G + G_{i\,\text{inv}})/G_i$, where $G = 1/R$, $G_{i\,\text{inv}} = 1/R_{i\,\text{inv}}$ and $G_i = 1/R_i$ (R being the load-resistance (see fig. 4*a*), $R_{i\,\text{inv}}$ the leakage-resistance of the diode and R_i its forward resistance).

As regards the efficiency, then, a germanium diode cannot possibly compete with any vacuum diode, since in the latter, where $G_{i\,\text{inv}} = 0$, η_{det} approaches unity when G approaches zero (see fig. 5*b*), whereas a germanium diode, with $G_{i\,\text{inv}} \neq 0$, cannot approach this level of efficiency.

Practical tests confirm this theory. Let us consider *fig.* 6*a* showing the D.C. voltage V, produced across the capacitor, versus the peak value (E) of the signal voltage for the diodes whose characteristics are shown in fig. 3. The curves shown in fig. 6*a* refer to

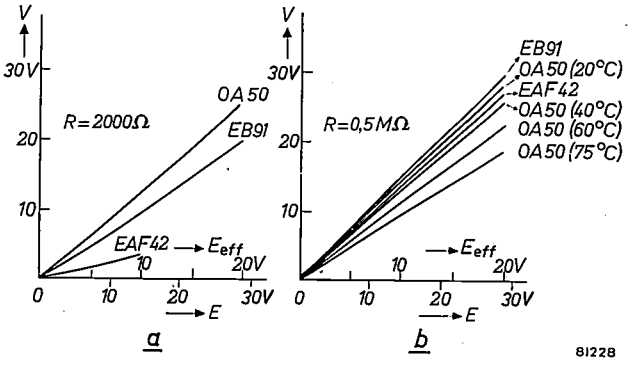


Fig. 6. *a*) The D.C. voltage V produced across the capacitor of a detector circuit as shown in fig. 4*a*, plotted against the peak value E of the applied signal voltage (the r.m.s. value E_{eff} of this voltage has also been plotted on the abscissa of the diagram). In the case considered here, the load-resistance is relatively low, i.e. 2 kΩ. These curves apply to the particular diodes specified in the diagram; the associated characteristics are shown in fig. 3.
b) As (*a*), but with a considerably higher load-resistance, i.e. 0.5 MΩ. In this case, the curve referring to the germanium diode (OA 50) is very dependent on the ambient temperature.

a low load-resistance, i.e. 2 kΩ. Their slight curvature arises from a deviation of the characteristics from the idealized case (straight lines meeting at the origin); G_i and $G_{i\,\mathrm{inv}}$ are not constant, but depend upon the voltage across the diode. Owing to this curvature, the detection efficiency V/E increases with the signal voltage. It will be seen that by virtue of the low load-resistance employed, the favourable forward characteristic of the germanium diode (high G_i) is reflected in a detection efficiency superior to those of the vacuum diodes.

Here, then, the leakage of the germanium diode does not impair its efficiency, as may be deduced from the fact that the curve appropriate to this diode is independent of the ambient temperature. Although a rise in temperature increases the leakage considerably, it has no appreciable effect on the forward resistance. As follows from theory outlined above, the rate of increase of the detection efficiency with the load resistance is higher in vacuum diodes than in germanium diodes. Curves plotted at a very much higher load resistance, i.e. 0.5 MΩ, are shown in fig. 6*b*. The leakage in the germanium diode then plays an important part, as is evident from the fact that the detection efficiency decreases with rising ambient temperature. There is no appreciable difference in detection efficiency between the three diodes considered at temperatures up to 40 °C.

It should be noted that the germanium diode type OA 50 is adopted as an example to illustrate the effect of a high load-resistance. In other diodes, e.g. types OA 51, OA 61 and OA 71, the leakage is very much smaller, and the detection efficiency associated with a high load-resistance is therefore higher than in the above case.

Another point illustrated by fig. 6 which should be mentioned in passing, is that the diode part of the EAF 42 must not be operated with a low load-resistance; it is, in fact, designed to operate with a load of the order of 0.5 MΩ.

Equivalent attenuation resistance

An important factor in many circuits, apart from the detection efficiency, is the power extracted from the voltage source. In so far as the power absorbed depends on the detector circuit (it also depends, of course, on the magnitude of the signal voltage), it is usually defined in terms of the so-called equivalent attenuation resistance (R_d). This is the resistance which, if substituted for the detector circuit, would draw the same amount of power from the voltage source. If P be the power so extracted and E the peak voltage supplied by the particular source, the equivalent attenuation resistance R_d is defined by the equation:

$$\frac{E^2}{2R_d} = P. \quad \ldots\ldots\ldots (1)$$

Given a circuit as shown in fig. 4*a* with both the load-resistance (R) and the diode leakage-resistance ($R_{i\,\mathrm{inv}}$) high in relation to the diode forward resistance (R_i), the approximate value of R_d is readily evaluated. In this case (to be precise, when R and $R_{i\,\mathrm{inv}}$ are infinitely high), the D.C. voltage produced across the capacitor (V) is E (with $\eta_{\mathrm{dét}} = 1$); the power dissipated in R is then E^2/R. Accordingly, a small current flows during the very short forward period (see fig. 4*b*); however, we shall ignore the power dissipated by this current in the forward resistance. The voltage across $R_{i\,\mathrm{inv}}$ is now, during virtually the whole period, given by $E(-1 + \cos\omega t)$, so that the power dissipated in $R_{i\,\mathrm{inv}}$ is $\frac{3}{2} E^2/R_{i\,\mathrm{inv}}$. Hence $P \approx E^2(1/R + 3/2\, R_{i\,\mathrm{inv}})$, and it follows, with the aid of (1), that:

$$1/R_d \approx 2/R + 3/R_{i\,\mathrm{inv}}. \quad \ldots\ldots (2)$$

Accordingly, the equivalent attenuation resistance of a vacuum diode ($R_{i\,\mathrm{inv}} = \infty$) is roughly equal to half the load-resistance, whereas in the case of a germanium diode ($R_{i\,\mathrm{inv}} \neq \infty$) the resistance R_d is equal to half the load-resistance and a third of the diode leakage-resistance in parallel.

Fig. 7 shows the calculated equivalent attenuation conductance G_d ($= 1/R_d$) versus the load conductance G ($= 1/R$), for different values of the leakage conductance $G_{i\,\mathrm{inv}}$ ($= 1/R_{i\,\mathrm{inv}}$), in the case of an idealized diode characteristic consisting of straight lines (constant forward and leakage resistances). The approximation involved in the above argument means that the tangents to the starting-points of the curves, seen at the bottom left-hand side of the diagram, are used instead of the curves themselves.

It is seen that the effect of increasing load and of increasing leakage upon the attenuation is not quite as serious as formula (2) suggests; however, this does not detract from the practical validity of this formula.

A difficulty in applying the formula arises from the fact that, as can be seen from the characteristics of the germanium diodes, the leakage resistance is governed by the voltage. Hence the average leakage-resistance appropriate to a given peak voltage must be estimated from the characteristic.

An increase in the signal voltage relative to zero will at first produce an increase in the average leakage resistance; but if the signal voltage increases

far enough, it will diminish again (fig. 3*b*). Accordingly, a high signal voltage and a high ambient temperature, both of which reduce the leakage resistance, impair the efficiency of a germanium diode.

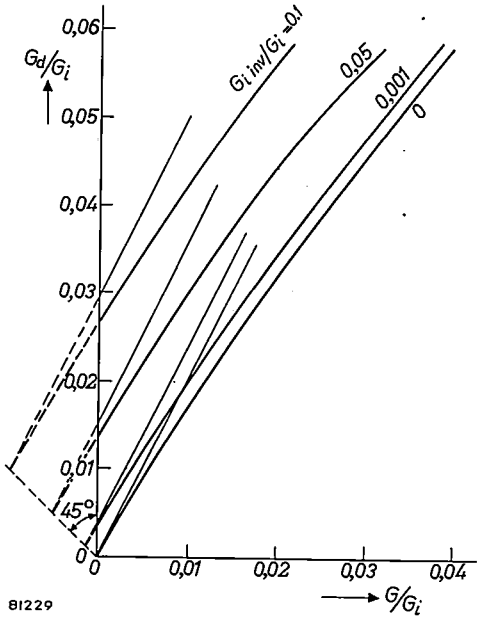


Fig. 7. Calculated attenuation conductance ($G_d = 1/R_d$) versus the load conductance ($G = 1/R$), for various values of the leakage conductance ($G_{i\,inv} = 1/R_{i\,inv}$), in the case of a diode with the characteristic shown in fig. 5*a*. As in fig. 5*b*, all the quantities involved are expressed in terms of G_i ($= 1/R_i$), that is, the forward conductance of the diode. To obtain the curve appropriate to a given value of $G_{i\,inv}/G_i$, transfer the $G_{i\,inv}/G_i = 0$ curve a distance equivalent to the particular value of $G_{i\,inv}/G_i$, first to the left, and then the same distance upwards.

From (2), if the load-resistance be reduced, the extra attenuation arising from leakage will become less important; in fact, given a sufficiently drastic reduction, it will be possible to substitute a germanium diode for a vacuum diode with virtually no change in attenuation. Practical experience has shown that the difference in attenuation becomes negligible at a load-resistance of some ten-thousands of ohms. Of course, the precise load-resistance to produce this effect depends upon $R_{i\,inv}$, which, in turn, depends upon the particular type of germanium diode, and upon the ambient temperature and the applied A.C. voltage. Now, given a load-resistance of the above-mentioned order of magnitude, the detection efficiency of the various types of germanium diode, and that of the vacuum diode type EB 91, will all be in the region of 100%; hence there will be virtually no difference in performance between any of them. However, any further reduction of R will affect both the attenuation and the detection efficiency unfavourably. Whereas the equivalent attenuation resistance of any diode will then remain equal to half the load-resistance, the higher forward resistance of the germanium diode, as we have already seen, will give it a detection efficiency superior to that of the vacuum diode.

The above argument only holds good for a vacuum diode whose forward resistance is not unduly high. Given a vacuum diode having a high forward resistance, e.g. the EAF 42, the situation will be more complex because, amongst other reasons, such a valve operates in a range perceptibly affected by deviations from formula (2) (see fig. 7, bearing in mind that G_i is now low).

Provided that the signal voltage is low ($< 0{\cdot}5\ V_{eff}$) and the ambient temperature is not unduly high, it is usually possible to procure a better detection efficiency and less attenuation with germanium diodes than with vacuum diodes. The reason for this is that, under these conditions, vacuum diodes operate in the starting-current region; here the detection efficiency and the equivalent attenuation resistance are both low, and the relation $R_d = \frac{1}{2}R$ no longer holds good in this region. However, the behaviour of vacuum diodes under these conditions will not be pursued here.

In the region between 0 and about 10 μV, there is no difference in slope between the forward and reverse characteristics of a germanium diode; which then behaves like an ordinary resistor, that is, without rectification.

That the germanium diode is particularly suitable for use in circuits where the load resistance is fairly low and the signal voltage and ambient temperature not unduly high can be deduced from the above argument. However, the other advantages of the germanium diode as mentioned on the opening page of this article frequently turn the scale in its favour even in cases where the electrical properties of this diode do not match those of the vacuum diode.

Performance at high frequencies

The efficiency of a germanium diode as a detector at high frequencies (> about 40 Mc/s) cannot be properly predicted from the static characteristic. We have already seen something of the nature of the effects ("hole storage") precluding such a prediction in the preceding article [1]).

We shall now illustrate, with the aid of one or two practical examples, how these effects modify the performance of a germanium diode in a practical circuit. The values referred to are the results of tests on a TV video detector circuit containing a 3.9 kΩ load-resistor shunted by a capacitor of roughly 20 pF. Given a germanium diode type OA 50, and a signal of r.m.s. voltage 5 V, the measured detection efficiency of this circuit is 0.62 at 30 Mc/s and 0.605 at 70 Mc/s.

The first thing that we notice about these values is that the detect efficiency is quite low at both frequencies. This is partly attributable to the fact that, particularly at 30 Mc/s, the time constant of the load is not very long compared to one cycle; it is also attributable to the self-capacitance of the diode, owing to which an appreciable part of the H.F. voltage occurs across the load (20 pF). The difference in the detection efficiency as between the two frequencies is only small but the difference in the attenuation caused by the detector circuit in the circuit immediately preceding it is considerable: the equivalent attenuation resistance is 2400 Ω at 30 Mc/s and 1450 Ω at 70 Mc/s. Since the D.C.voltage across the load-resistance and therefore the power dissipated in it is very much the same in both cases, the relatively greater attenuation at 70 Mc/s must be attributed to extra losses in the germanium diode. In a TV receiver, such strong attenuation detracts from the gain of the amplifier stage preceding the detector; moreover, owing to the variation in attenuation between individual diodes of the same type, it is responsible for variations in the amplification and in the frequency characteristic of production receivers.

It is also worth mentioning that the decline in performance at relatively high frequencies is all the more noticeable according as the load-resistance is increased.

Examples illustrating the use of diodes for special purposes

As we may conclude from the above, the germanium diode may be substituted for the vacuum diode in a very large number of cases. At the same time, the leakage exhibited by the germanium diode, must often be taken into account in the design of a particular circuit. With this reservation, however, there is freedom of choice from several types, according to the particular application envisaged; these include types of low leakage and high maximum reverse voltage (*fig. 8*).

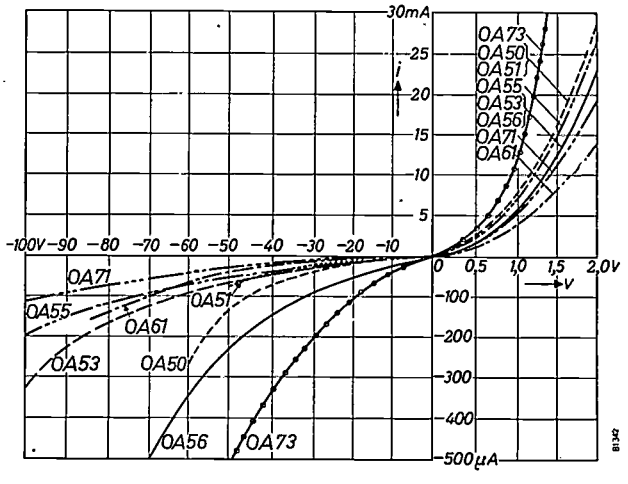


Fig. 8. The characteristics of various types of germanium diode.

Video detection

Fig. 9 is the circuit diagram of a video detector employing a germanium diode, i.e. type OA 60 or type OA 70. As we have already seen, these diodes are specially designed for use in such a circuit. In the case here considered, the operating frequency is 24 Mc/s and the signal voltage in the I.F. circuit (L_2-C_2) is 5 V r.m.s.; the detection efficiency is then about 70% that is, the D.C. voltage generated across the 3·9 kΩ load-resistor (R) is about 5 V. The equivalent resistance of the detector circuit is about 3 kΩ.

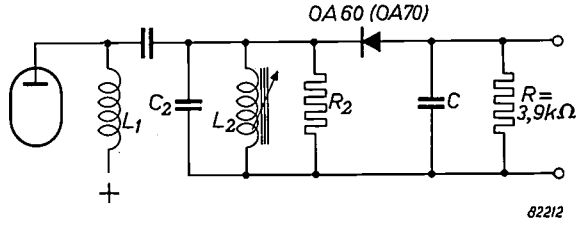


Fig. 9. Video circuit containing a germanium diode type OA 60 or OA 70.

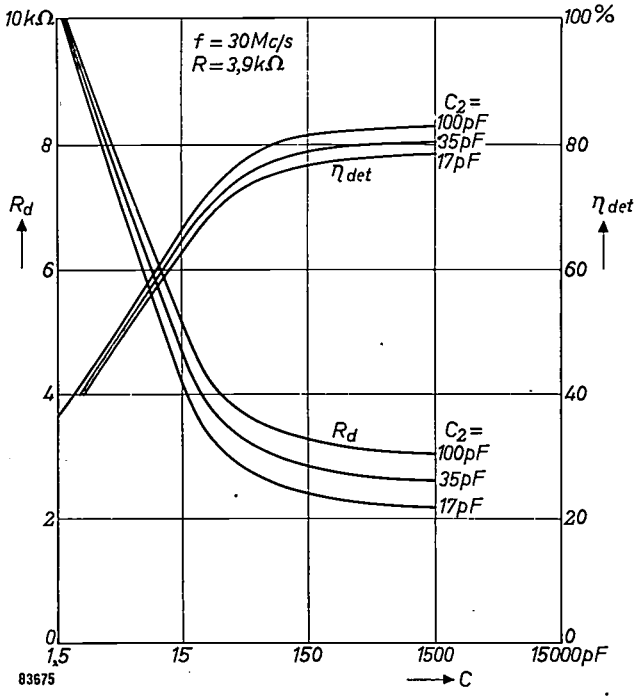


Fig. 10. Effect of the load-capacitance C on the detection efficiency η_{det} and equivalent attenuation resistance (R_d) of the video detector circuit shown in fig. 9. These quantities also depend upon the tuning capacitance C_2; the curves for several values of this parameter are given. The curves are valid for a carrier-wave frequency of 30 Mc/s and a load-resistance R of 3.9 kΩ.

The detection efficiency and the attenuation are also affected by the total tuning capacitance (C_2) of the circuit (see *fig. 10*), which, in the diagram shown in fig. 9, comprises the self-capacitance of the coils and the valve and wiring capacitances (totalling about 17 pF in all).

D.C. restoration

Fig. 11a shows a circuit employed to fix the black level in the signal applied to the picture tube of a television receiver. This circuit includes a germanium diode, type OA 61 or type OA 71, specially designed for the purpose and exhibiting only a small amount of leakage even when operated with a high reverse voltage.

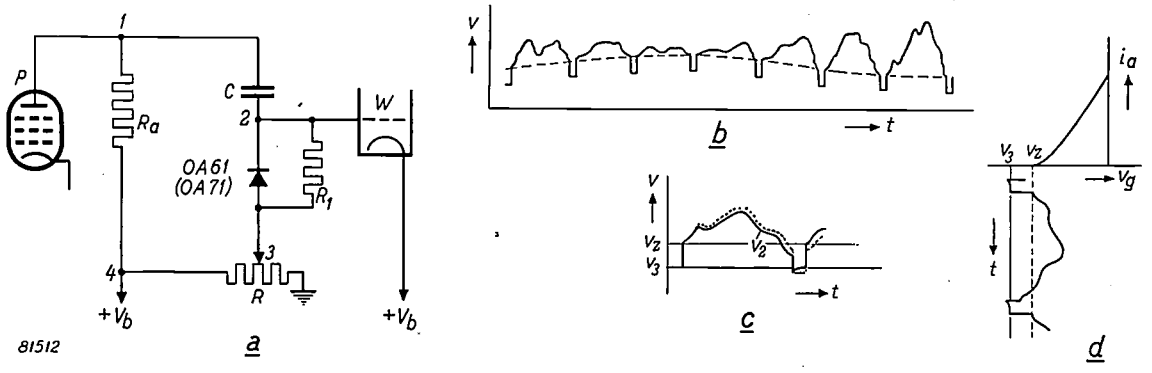


Fig. 11. *a*) D.C. restorer circuit employed to fix the black level of the signal applied to the picture tube W of a television receiver.
b) Video signal with line-synchronizing pulses, as applied to the grid of the picture tube when the diode shown in (*a*) is omitted. In practice, the variation of the sync. pulse peaks is much more gradual than this diagram suggests.
c) Video signal on the grid of the picture tube (full line). Each line-period starts at a fixed value (v_z) corresponding to the pre-adjusted black level of the picture tube. The dotted line represents the signal as it would be if not corrected by the diode and the resistor R_1.
d) The video signal referred to the i_a-v_g characteristic of the picture tube. With the aid of potentiometer R(see (*a*)), v_3 is so adjusted that v_z (the black level) coincides roughly with the cut-off point of the tube-characteristic.

The action of such a circuit will be explained here for those unfamiliar with TV circuit technique. Fig. 11*b* shows roughly what the shape of the video signal on the grid of the picture tube would be in the absence of a diode. The upper ends of the sync. pulses, which correspond to black in the scene televised, should remain at a constant level in order that black in the scene will always give rise to the same degree of blackness on the picture tube. However, it will be seen that these peaks rise and fall with the average picture-content of the video signal preceding them. Accordingly, the task of the circuit shown in fig. 11*a* is to fix the black level at the start of each line period, thus correcting any variation in this level introduced during the preceding line period, that is, preventing any *cumulative* variation. Because the depth of the synchronizing pulses is constant, it is sufficient to maintain the bases of these pulses at a fixed level. This is achieved as follows.

As long as the potential (v_2) of the bottom plate of capacitor C remains above the potential at point 3 (v_3), the diode remains non-conductive. Owing to the high resistance of R_1, the charge and potential difference of the capacitor C can vary only gradually; hence v_2 must respond almost exactly to any rapid variations in v_1 (potential at point 1).

However, as soon as v_2 drops below v_3, as it does during the application of a synchronizing pulse (fig. 11*c*), the diode becomes conductive. A charging current then flows to C so that, at the end of the sync. pulse, v_2 (which is also the control-grid potential of the picture tube W) is raised to the fixed value represented by v_3. Potential v_1 then increases again, thus initiating the next line period, during which v_2 again responds almost exactly to the variations of v_1. Owing to the fact that a small proportion of the charge of C leaks away through R_1 during the line period, voltage v_2 again drops below v_3 at the next sync pulse; however, this leakage is made good and a new line period starts from v_3 (fig. 11*c*). The necessity of ensuring that v_2 will drop below v_3 during each synchronising pulse imposes a maximum limit on the size of R_1.

The direct current flowing in potentiometer R causes v_3 to assume a certain value below the cathode potential; hence this potentiometer can be used to vary v_3 and so adjust the black level of the picture tube to a particular fixed value (fig. 11*d*).

If much of the charge of capacitor C leaks away through the diode in the interval between two successive sync. pulses, the grid voltage (v_2) of the picture tube (W) will already have fallen well below the pre-adjusted black level (v_z) by the time the second sync. pulse arrives (fig. 11*c*). One detrimental effect of this premature drop in v_z is that the side of the picture at which the raster-lines end becomes darker than the side where these lines start. There is, however, still another, more serious disadvantage. With such a drop, the diode becomes conductive, and therefore the charging of C commences before the sync. pulse is fully developed. Instead of flowing in R_a, as under ordinary conditions, the greater part of the anode current of valve P then

takes the line of least resistance through C, thus reducing the amplification of the valve abruptly and considerably. As long as the charging current continues to flow — that is, during a considerable part, if not the whole, of the pulse period (since the considerable amount of charge at first released must now be made good) — the pulse is unable to develop to its full height and is therefore distorted and attenuated. This effect may attain such proportions as to upset the synchronization of the picture.

This explains why it is necessary in this application to employ a diode exhibiting only a small leakage at the high reverse voltages involved.

Summary. Among the obvious advantages of germanium diodes compared with their vacuum equivalents, are their relatively small size and weight and the fact that they contain no filament. As regards the normal functions of a diode, that is, rectification and detection, the leakage current of the germanium diode and its superior forward characteristic, constitute the most striking differences between it and the vacuum diode. The effect of these differences upon the detection efficiency and equivalent attenuation resistance in a detector circuit is investigated. It is shown that germanium diodes can be employed in a very large number of cases and that, provided that the load-resistance is low and the signal voltage and ambient temperature are not unduly high, the performance of such a diode is even better than that of a vacuum diode. Special germanium diodes have been developed for certain purposes, e.g. to operate at high frequencies (video detection in a television receiver) and for use as D.C. restorers for maintaining the black level in the picture tube of a TV receiver.

SEDIMENTATION OF FLUORESCENT SCREENS IN CATHODE RAY TUBES

by F. de BOER and H. EMMENS.

535.371.07:621.385.832

The ever-increasing use of cathode ray tubes for television and other purposes creates the need for a sufficiently reproducible method of forming the fluorescent screen in the tube. The problems thereby involved are largely of a colloid-chemical nature. A detailed study of the effects occurring on sedimentation of the screen is possible only if the chemical composition of the layer is known. This composition can be determined with the aid of radio-active tracers.

Introduction

The fluorescent screens of picture-tubes and other types of cathode ray tube are composed of a mixture of substances. The fluorescent substance itself, the so-called phosphor, of course predominates, but for the deposition of the screen it is necessary to include a number of components to ensure a sufficiently adhesive and finely grained layer. These admixtures affect the characteristics of the screen, such as the secondary emission factor. If this is too low, heavy negative charging of the (insulated) screen will occur during use, to the detriment of the luminous efficiency and fidelity of reproduction [1]). On the other hand, the content of non-fluorescent substances must not be too high, as this also has an unfavourable effect on the fluorescent efficiency. It is therefore necessary to select and apportion the additional constituents with care.

The method of preparation of fluorescent mixtures nowadays employed is as follows. The fluorescent grains are suspended in water containing waterglass (a colloidal solution of potassium silicate in water with excess silica). A gelatinising agent is also added. This suspension is poured into the tube and, after a certain time, the powder settles to form a uniform, closely adhering layer, the solvent being then poured off. *Fig. 1* shows how this is done in manufacture.

The use of a gelatinising agent is necessary to prevent the layer of sediment from sliding to one side and also to prevent individual granules from becoming detached, when the tube is tilted in the decanting process. Moreover, in the dry condition, during exhaustion of the tube as well as during use, effective adhesion of the granules must be guaranteed; the presence of loose granules in a tube operated at high voltages must be avoided at all costs.

Until a few years ago the gelatinising medium used was potassium sulphate, but a disadvantage of this "sulphate" method is that it takes some time to produce a sufficiently adhesive layer, viz. 1 to 2 hours. A more rapidly reacting medium was therefore sought. Such a medium was found among certain soluble calcium or barium salts, of which barium nitrate, for example, produces a firmly

[1]) See for example J. de Gier, A. C. Kleisma and J. Peper, Secondary emission from the screen of a picture tube, Philips tech. Rev. **16**, 26-32, 1954/55.

takes the line of least resistance through C, thus reducing the amplification of the valve abruptly and considerably. As long as the charging current continues to flow — that is, during a considerable part, if not the whole, of the pulse period (since the considerable amount of charge at first released must now be made good) — the pulse is unable to develop to its full height and is therefore distorted and attenuated. This effect may attain such proportions as to upset the synchronization of the picture.

This explains why it is necessary in this application to employ a diode exhibiting only a small leakage at the high reverse voltages involved.

Summary. Among the obvious advantages of germanium diodes compared with their vacuum equivalents, are their relatively small size and weight and the fact that they contain no filament. As regards the normal functions of a diode, that is, rectification and detection, the leakage current of the germanium diode and its superior forward characteristic, constitute the most striking differences between it and the vacuum diode. The effect of these differences upon the detection efficiency and equivalent attenuation resistance in a detector circuit is investigated. It is shown that germanium diodes can be employed in a very large number of cases and that, provided that the load-resistance is low and the signal voltage and ambient temperature are not unduly high, the performance of such a diode is even better than that of a vacuum diode. Special germanium diodes have been developed for certain purposes, e.g. to operate at high frequencies (video detection in a television receiver) and for use as D.C. restorers for maintaining the black level in the picture tube of a TV receiver.

SEDIMENTATION OF FLUORESCENT SCREENS IN CATHODE RAY TUBES

by F. de BOER and H. EMMENS.

535.371.07:621.385.832

The ever-increasing use of cathode ray tubes for television and other purposes creates the need for a sufficiently reproducible method of forming the fluorescent screen in the tube. The problems thereby involved are largely of a colloid-chemical nature. A detailed study of the effects occurring on sedimentation of the screen is possible only if the chemical composition of the layer is known. This composition can be determined with the aid of radio-active tracers.

Introduction

The fluorescent screens of picture-tubes and other types of cathode ray tube are composed of a mixture of substances. The fluorescent substance itself, the so-called phosphor, of course predominates, but for the deposition of the screen it is necessary to include a number of components to ensure a sufficiently adhesive and finely grained layer. These admixtures affect the characteristics of the screen, such as the secondary emission factor. If this is too low, heavy negative charging of the (insulated) screen will occur during use, to the detriment of the luminous efficiency and fidelity of reproduction [1]). On the other hand, the content of non-fluorescent substances must not be too high, as this also has an unfavourable effect on the fluorescent efficiency. It is therefore necessary to select and apportion the additional constituents with care.

The method of preparation of fluorescent mixtures nowadays employed is as follows. The fluorescent grains are suspended in water containing waterglass (a colloidal solution of potassium silicate in water with excess silica). A gelatinising agent is also added. This suspension is poured into the tube and, after a certain time, the powder settles to form a uniform, closely adhering layer, the solvent being then poured off. *Fig. 1* shows how this is done in manufacture.

The use of a gelatinising agent is necessary to prevent the layer of sediment from sliding to one side and also to prevent individual granules from becoming detached, when the tube is tilted in the decanting process. Moreover, in the dry condition, during exhaustion of the tube as well as during use, effective adhesion of the granules must be guaranteed; the presence of loose granules in a tube operated at high voltages must be avoided at all costs.

Until a few years ago the gelatinising medium used was potassium sulphate, but a disadvantage of this "sulphate" method is that it takes some time to produce a sufficiently adhesive layer, viz. 1 to 2 hours. A more rapidly reacting medium was therefore sought. Such a medium was found among certain soluble calcium or barium salts, of which barium nitrate, for example, produces a firmly

[1]) See for example J. de Gier, A. C. Kleisma and J. Peper, Secondary emission from the screen of a picture tube, Philips tech. Rev. **16**, 26-32, 1954/55.

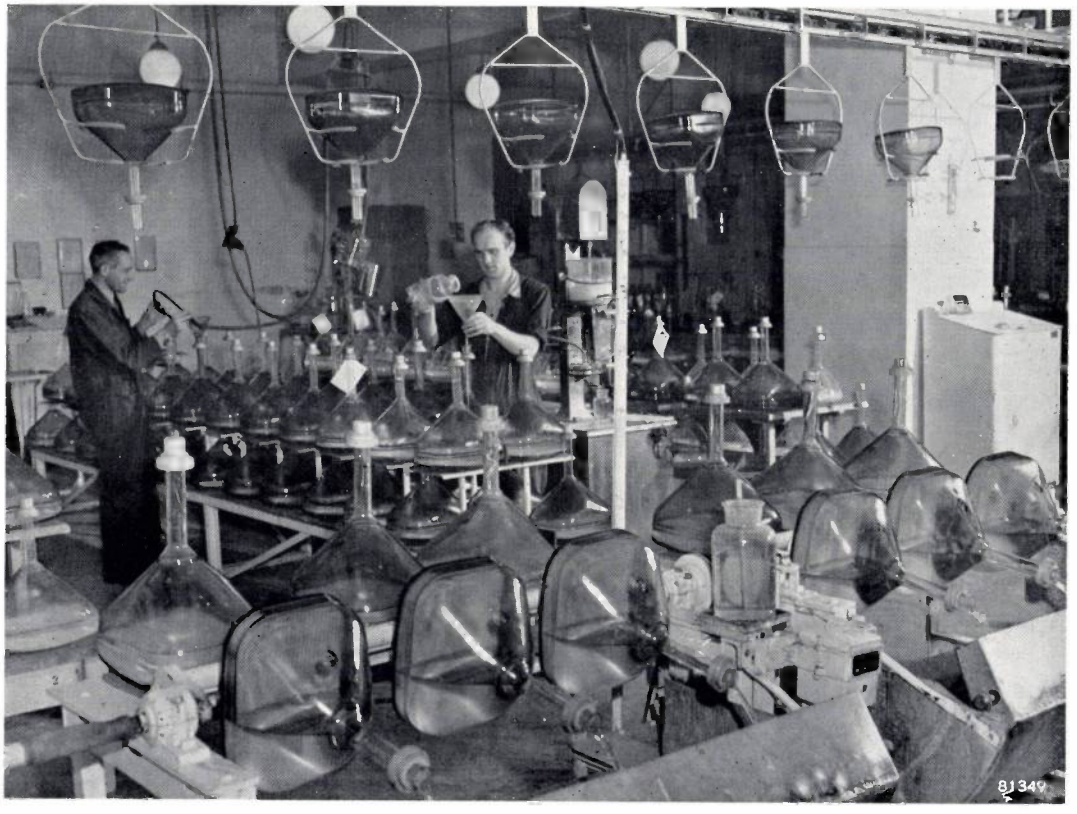


Fig. 1. Photograph of the manufacturing process for the application of the screens to television picture tubes. The tubes enter the settling room on an overhead conveyor. Background: filling with the liquid containing the phosphor and binding agents in suspension. Foreground: decanting the liquid.

attached layer in roughly 15 minutes. This represents a considerable advantage in mass production, but the use of barium salts for this purpose nevertheless has the disadvantage that the secondary emission factor of the deposited screen is lower than that of screens made by the sulphate method. Particularly in modern tubes, operated with very high acceleration potentials — at which the secondary emission is in any case fairly low — this is an important consideration.

It will be obvious that the adhesive properties of the settled screen depend upon the composition of the solution. Screens prepared in accordance with the barium method yield the curve depicted in *fig. 2*. The various concentrations must be selected within the area to the right of the curve *BAC* in order to ensure a sufficiently adhesive layer. We shall refer to this diagram again later.

In order further to clarify the processes that take place while the layer is being formed, and to observe the relationship between the composition of the fluorescent screen and the secondary emission factor an investigation has been made into the potassium, barium and also the SiO_2 content of fluorescent screens prepared by the two different methods. This investigation, in which use was made of radio isotopes, will now be described briefly. First, however, it is worth considering the picture that has been built up of the actual mechanism of formation of the layer.

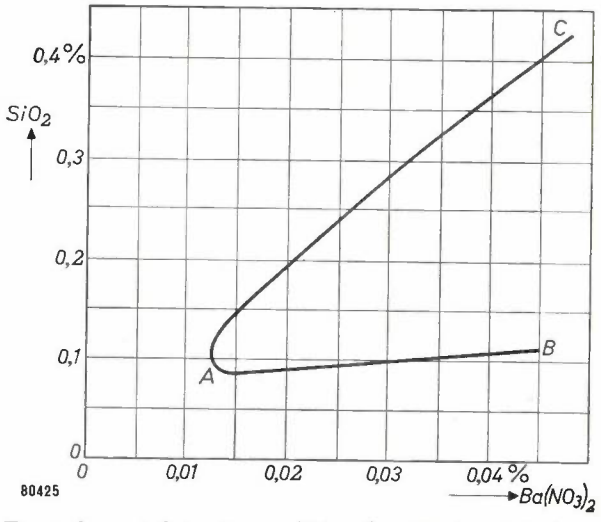


Fig. 2. Curve defining the conditions for adhesion of the fluorescent screen of a cathode ray tube. Vertical axis: concentration (% by wt) of waterglass, i.e. SiO_2, in the sedimentation liquid; horizontal axis: concentration of gelatinising medium (barium nitrate). In the area enclosed within the curve *BAC* effective adhesion of the screen is obtained.

Mechanism of formation of the layer

In the suspension, the granules of the fluorescent powder are coated with a slowly forming layer of silicate, and during sedimentation these layers come into contact with each other and coalesce into a gel, resulting in considerable mutual adhesion between the granules, and between them and the glass tube. In the absence of a gelatinising agent, hardly any gelling can take place, the adhesion is inadequate and the ultimate layer unstable.

Potassium sulphate owes its gelatinising properties solely to the electrostatic effect of the potassium ions, which partially neutralises the mutual electrostatic repulsion of the silica acid groups and thus promotes gelation. The potassium ions do not themselves function as constructional elements in the resultant gel and are therefore capable, in principle, of securing the adhesion of a practically unlimited number of granules. However, the gelatinising process takes a relatively long time for its completion (viz 1 to 2 hours, dependent on the potassium content) and hence much longer than the actual sedimentation.

Under otherwise similar conditions (equal acidity) barium salts produce much more rapid gelatinisation. (in about 20 minutes), but in the resultant gel the barium ions probably occur in the form of a compound. Once the gelatinisation is completed, no more barium ions will be available in the suspension, and granules which have by then not become attached to each other will not do so with the passage of time. The size of the granules in layers settled by the barium method are therefore, for a given barium concentration, limited.

Even without any fluorescent powder at all the presence of barium salts in the potassium silicate results in a gel, as seen by increasing turbidity of the liquid.

Determination of the composition of the layer

The composition of screens produced by the sulphate method can be easily determined by drying the settled layer after decanting the liquor, and leaching with distilled water. The phosphor granules do not dissolve, but the Si_2O-K_2O compound binding them together (briefly called binders in the following) do; and the solution can be analysed.

In the case of screens made with barium nitrate, however, such chemical analysis does not yield the desired results. Barium and potassium (from the water glass) are then both present and it is not a simple matter to separate the very small quantities concerned. A solution to this difficulty has been found in the use of radioactive tracers, viz. Ba^{133} (an artificial barium isotope of atomic weight 133), and K^{42} (potassium isotope, atomic weight 42). Both isotopes were prepared in the cyclotron at Amsterdam [2]). Ba^{133} has a half-life of 38.8 hours and it emits both soft γ radiation, with an energy of 0.3 MeV, and β radiation (conversion electrons). Neither radiation is very dangerous provided that the necessary precautions are taken in effecting the analysis. The half-life of K^{42} is shorter than that of Ba^{133}, viz. 12.4 hours, and the radiation emitted (γ) is harder and less innocuous; the energy is 1.5 MeV. As the time elapsing between preparation of the isotopes and their arrival at Eindhoven was about 16 hours, i.e. of the same order as the half-lives just mentioned, it was necessary to carry out the investigation immediately on receipt of the isotopes.

The barium in the form of the nitrate was received in bottles, packed in lead and consisting of a solution of 100 mg nitrate in 100 cc water. Immediately on receipt the solution was diluted with 500 cc distilled water, the necessary precautions being observed. Owing to absorption of the electrons in the water the isotope was thus rendered safe enough to carry out the various operations by hand, provided that this was done reasonably quickly.

The radioactive potassium arrived as an aqueous solution of KCl and was similarly diluted. This solution could also be handled quite safely in the ordinary way for a short period of time.

To a known quantity of the solution to be used for the screens, and containing the normal barium nitrate, 50 cc of the radioactive solution was added. Because of the chemical identity of all barium atoms, regardless of their weight, a homogeneous mixture of radioactive and non-radioactive barium results: hence the radioactivity of a portion of the mixture is a measure of the total quantity of barium in that portion. After sedimentation the liquid was decanted from the tubes by mechanical means. In this way 99 % of the radioactive substances was disposed of down the sink where, owing to the degree of dilution and the short half-life, they could do no harm locally. The residual radiations within the tube were also harmless. When dried, the deposited fluorescent powder was scraped off and spread out on an aluminium dish and the barium content was ascertained by measurement of the radiation intensity with a Geiger counter. For control purposes, a count was first taken from a preparation containing the same quantity of

---

[2]) A. H. W. Aten, and J. Halberstadt. The production of radio-isotopes, Philips tech. Rev. **16**, 1-12, 1954/55.

fluorescent powder and a *known* quantity of Ba^{133} of the same age as that used in the analysis.

To determine the quantity of potassium in the deposited layer, a similar procedure was followed, except that in this case 50 cc of the radioactive potassium solution was initially added to the normal barium nitrate solution.

The necessary correction was of course made for the background counts (cosmic radiation, etc).

The SiO_2 content of the powder was determined by normal chemical methods.

Results of the investigation

The table below gives some of the results relating to the proportions of BaO and SiO_2 in the binder as a function of the concentration of barium nitrate and potassium silicate in the liquid. From this it is seen that the molecular ratio $BaO\text{-}SiO_2$ in the binder is practically independent of the composition of the liquid, which agrees with the view expressed above that we are not here concerned with an evaporated solution yielding a mixture, but that the barium and SiO_2, are present in the binder in the form of a sort of compound. Another argument in favour of this conclusion is that the quantity of binder is roughly proportional to the Ba content of the solution. On the other hand, a very strong binding, of the nature of an adsorption, could explain these facts.

In *fig. 3* the $K_2O\text{-}SiO_2$ ratio in the binder is shown plotted against the concentration of potassium silicate in the solution. Curve *1* refers to screens made by the sulphate method, and curve *2* to the barium nitrate method. Clearly there is no question of the $K_2O\text{-}SiO_2$ ratio in the binder being constant.

These conclusions are fully in accordance with what has been said above regarding the behaviour of the gelatinising media in the solution. The diagram in fig. 2, representing the conditions for good adhesion of the settled screen can now be explained as follows. The line AB represents the (linear) relationship between the quantity of binder deposited and the barium concentration of the solution. To the left of A, the barium concentration is too low to produce a sufficiently adhesive layer, whilst below the line AB the SiO_2 content is too low. The time needed to ensure a properly adhesive layer increases with the concentration of potassium (derived from the water-glass), and, above the line AC this time is so long that the layer is no longer stable. The slope of the curve AC implies that adhesion is improved as the barium content of the solution is increased.

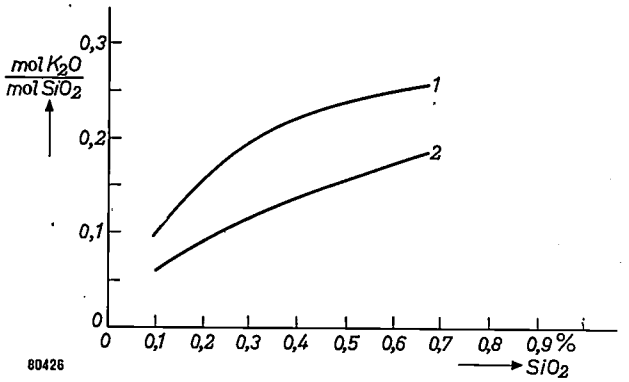


Fig. 3. Relationship between the molecular ratio $K_2O\text{-}SiO_2$ in the binder, and the silicate concentration (% by weight) in the solution. *1*) screens made by the sulphate method, *2*) screen made by the barium method.

Some idea of the thickness of the adsorbed layer of binder around the phosphor granules can be obtained from the relationship between the absolute quantity of binder settled and the phosphor concentration of the suspension. This is shown in *fig. 4*; the relationship is practically linear. The intersection

Fig. 4. Quantity of deposited binder plotted against quantity of phosphor in the suspension. The intersection of the curve on the vertical axis represents the quantity of binder deposited on the glass. The slope is a measure of the quantity of binder adsorbed per granule of phosphor.

Table. Analysis of the molecular ratio $BaO\text{-}SiO_2$ in sedimented layers.

Percentage by weight $Ba(NO_3)_2$ in the liquid	Percentage by weight SiO_2 in the liquid	Molec. ratio $BaO\text{-}SiO_2$ in liquid	Molec. ratio $BaO\text{-}SiO_2$ in binder	Quantity of binder deposited		
				BaO	SiO_2	Total
0.02 %	0.25 %	1:32	1:16	1.5 mg	9.4 mg	10.9 mg
0.02	0.38	1:48	1:18	1.2	8.6	9.8
0.02	0.53	1:64	1:16	1.3	8.0	9.3
0.01	0.12	1:32	1:19	0.7	4.8	5.5
0.01	0.19	1:48	1:19	0.7	4.8	5.5
0.01	0.26	1:64	1:22	0.7	5.9	6.6

of the curve on the vertical axis is a measure of the quantity of binder deposited on the glass screen. If the grain size of the phosphor be known it is possible to calculate from the slope of the line the quantity of binder per granule and, from this, the thickness of the adsorbed layer. A thickness of the order of 100 Å (10^{-6} cm) was found, and this agrees with the results of tests to ascertain the acceleration voltage that has to be applied to the tube to produce perceptible fluorescence. The acceleration voltage determines the depth of penetration of the electrons into the layer of binder. It is found that the penetration depth appropriate to the observed minimum acceleration voltage is also of the order of 100 Å.

It is important, especially for high-voltage cathode-ray tubes, that the potassium concentration in the screen should be as high as possible since, as explained above and as shown in *fig. 5*, this is accompanied by a high secondary emission factor of the screen. As indicated by the results plotted in fig. 3 this can be ensured by making the concentration of potassium silicate in the liquid as high as possible. This does not alter the $BaO\text{-}SiO_2$ ratio of the screen, but the time taken to secure satisfactory adhesion, is increased.

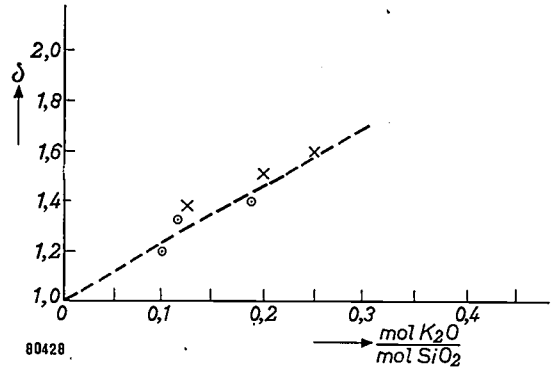
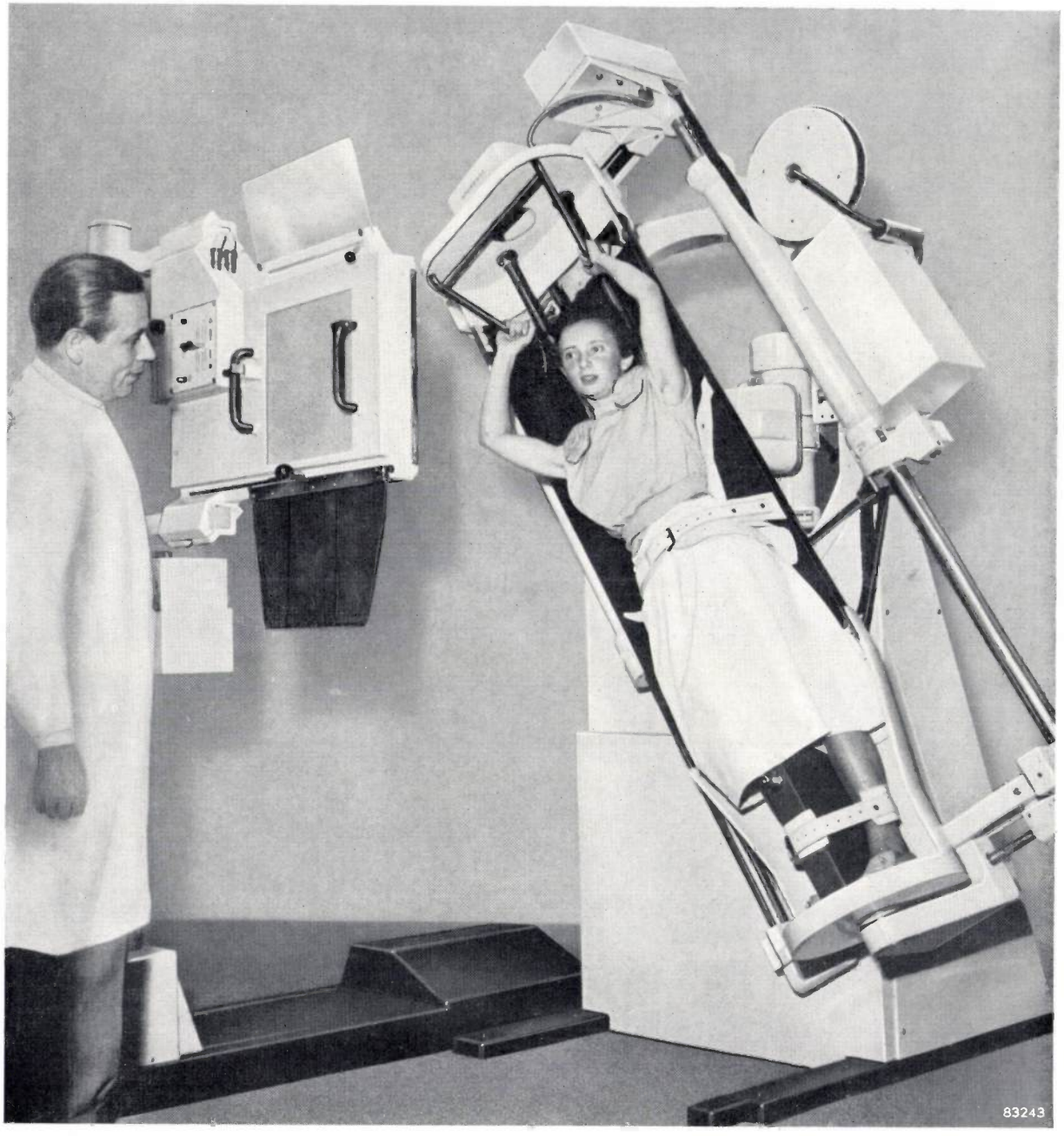


Fig. 5. Secondary emission factor δ of the fluorescent screen plotted against the molecular ratio $K_2O\text{-}SiO_2$ in the binder. The points denoted by crosses relate to screens prepared by the sulphate method; those shown as circles relate to the barium method.

As will be seen from fig. 3, the highest potassium concentration in the screen is obtained by employing the sulphate method; this concentration is of course reduced if part of the potassium be replaced by barium. On the other hand, as already pointed out, the addition of barium promotes rapid adhesion, so that it is necessary to aim at a composition for the suspension that will include enough barium to ensure reasonably rapid adhesion, but not so much that the secondary emission factor of the screen is reduced beyond the permissible limit.

Summary. The fluorescent screen in TV and other types of cathode ray tube is produced by sedimentation from a suspension. This article deals with the processes that take place during this desimentation. The effect of gelatisining agents on the adhesive properties of the screen is found to be very pronounced. Until recently potassium sulphate was used as gelatinising agent, but at the present time preference is given to barium salts, which promote more rapid adhesion. The chemical analysis of screens made from various solutions leads to some understanding of the behaviour of the gelatinising agents. For the analysis of the Ba and K in screens settled with barium salts, use was made of tracers, i.e. radio-isotopes of these elements, which also made possible an investigation into the effect of the composition on the secondary emission factor of the screen.

"MÜLLER" UGX APPARATUS FOR X-RAY DIAGNOSTICS



Apart from the X-ray tube and generator and the equipment for observing the X-ray picture, a diagnostic apparatus must also be provided with mechanical means for placing the patient in any position necessary for the examination. The usual type of universal diagnostic apparatus provides for the tilting of the patient about a horizontal axis perpendicular to the X-ray beam, and the X-ray tube and fluorescent screen or camera must follow this movement.

With the new apparatus shown, this conventional arrangement has been abandoned in favour of a fixed X-ray source and screen. (The protective screening is then also fixed and the radiologist does not need to move in order to continuously observe the X-ray picture). Instead of tilting the patient about the above-mentioned axis, this apparatus can turn the patient about a horizontal axis parallel to the direction of the X-ray beam and about his longitudinal axis. In this way, quite new positions of the patient and new irradiation directions are possible, which may be of great value for the examination of certain organs.

The apparatus, which has been developed by H. Verse and K. Weigel of Hamburg, was first demonstrated at the 7th International Congress of Radiology [1]) in Copenhagen, 1953. The apparatus will be more fully described in this Review in due course.

---

[1]) H. Verse and K. Weigel, Ein gerätetechnische Betrachtung zur Röntgendiagnostik, Fortschr. Röntgenstrahlen **80**, 520-524, 1954, No. 4.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can de obtained free of charge upon application to the Administration of the Philips Research Laboratory, Eindhoven, Netherlands.

2105*: J. C. Francken: Electron optics of the image iconoscope (Thesis, Delft, 1953).

Chapter I deals with the so-called cathode lens, an electrostatic lens which differs from ordinary immersion systems in that the electrons leave the object with nearly zero velocity. The image iconoscope, in which the image is formed by a cathode lens placed in a rotationally symmetric magnetic field, is then briefly discussed. In Ch. II the electrostatic field is discussed. General solutions of the rotationally symmetric Laplace equation are applied to a case which resembles that of the image iconoscope. For measurement of the field a resistance network was used. This method is analyzed and improved. In Ch. III the magnetic field of the deflection coil is dealt with. The measuring method of Van Mentz and Le Poole was used, and is examined as to accuracy. In Ch. IV the general and the paraxial equations for the trajectories are given, and the conditions specified under which the latter may be used. In Ch. V solutions of these equations in the form of expansions are found which lead to a general solution for the electron trajectories in the type of electron lens considered. In Ch. VI the paraxial trajectories in the Philips image iconoscope 5840/00 are calculated. A comparison is made between tubes with and without an "anode mesh". The computing methods used are discussed at some length and examined as to their accuracy. The mechanism of the image formation is explained with the aid of the computed electron trajectories. The definition of cardinal points is adapted to the special character of this type of electron lens. In Ch. VII experiments are described which provide a relation between the shape of the imaging coil and the magnification and rotation of the image. These experiments lead to a type of coil which permits a continuously variable magnification without rotation of the image. Ch. VIII and IX deal with aberrations, ion spot and diamond distortion. Finally, in Ch. X, the application of variable magnification in television cameras is discussed (see also Philips tech. Rev. **14**, 327-335, 1952/53).

2106: J. F. van Wieringen: Influence du traitement mécanique sur la résonance paramagnétique du manganèse dans les poudres de sulfure de zinc (Physica **19**, 397-400, 1953). (Influence of mechanical treatment on the paramagnetic resonance of manganese in zinc sulphide powders: in French)

If wurtzite, the hexagonal modification of ZnS, is compressed at room temperature, it changes into sphalerite, the cubic modification. This has been observed by X-ray diffraction. The gradual transition manifests itself also in the paramagnetic resonance spectrum of small quantities of Mn dissolved in ZnS. This is because the paramagnetic resonance spectrum of divalent manganese is very sensitive to the presence or absence of non-cubic components of the surrounding crystalline field. The spectrum as a function of Mn concentration is shown. In non-compressed wurtzite powders the complete spectrum is resolved at higher concentrations (0.05%) than had been reported before (0.001%) for compressed powders.

2107: K. F. Niessen: Condition for vanishing spontaneous magnetization below the Curie temperature (Physica **19**, 445-450, 1953).

For a special spinel structure in which the spontaneous magnetization at a temperature T_0 below the Curie temperature (T_c) becomes zero, a method has been derived to determine the temperature T_0. A condition is derived for the extreme case $T_0 = T_c$ which can be applied to experiments for finding the ratio of mutual interactions of magnetic ions.

2108: H. G. van Bueren: A model for demonstrating dislocations in crystals (Brit. J. appl. Phys. **4**, 144-145, 1953).

A simple dynamic model, in which the interatomic forces are simulated partially by magnetic forces and partially by the elastic forces exerted by small springs, is described. With it the movement of edge and screw dislocations through a crystal lattice can easily be demonstrated.

2109: D. Polder and J. Smit: Resonance phenomena in ferrites (Rev. mod. Phys. **25**, 89-90, 1953).

With the help of a model, consisting of an ellipsoid of ferrite with Bloch-walls perpendicular to one of the major axes, it has been demonstrated that considerable resonance losses can occur in ferrite throughout a very wide frequency band, viz. from the natural resonance frequency predicted by Snoek

up to the frequency of the Larmor-precession in a magnetic field of intensity of $4\pi M_s$, where M_s represents the saturation moment of the ferrite.

2110: H. P. J. Wijn, M. Gevers and C. M. van der Burgt: Note on the high frequency dispersion in nickel zinc ferrites (Rev. mod. Phys. **25**, 91-92, 1953)

In connection with a publication by Rado and others regarding two dispersion regions in ferrites, the behaviour of the imaginary component μ'' of the initial permeability ($\mu_i = \mu' - j\mu''$) is investigated in the frequency range 3Mc/s — 3000 Mc/s, in nickel-zinc ferrites of various compositions (18-50% NiO, 32-0% ZnO). No more than one resonance maximum is found. It is concluded from the behaviour of μ_i as a function of the frequency during tensile stresses (up to 2.7 kg/mm^2), that the dispersion in these ferrites is linked up with a rotation process of the spin vectors.

2111: H. P. J. Wijn and H. van der Heide: A Richter type after-effect in ferrites containing ferrous and ferric ions (Rev. mod. Phys. **25**, 98-99, 1953)

With manganese-zinc ferrite (28 MnO, 19 ZnO, 3 FeO, 50 Fe_2O_3) for each frequency f, the value of tan δ as a function of the temperature T shows a maximum at a temperature T_{max}. The relationship between $1/T_{max}$ and f is linear, which suggest an activation energy (0.11 eV). This effect is not found when a nickel-zinc ferrite is sintered at 1250 °C, but it is present when the sintering process takes place at 1525 °C, when 0.5% of Fe^{2+} ions are formed. Here too, the activation energy is approx. 0.1 eV. This leads to the conclusion that the residual losses in ferrites containing both ferrous and ferric ions are due to electron diffusion.

2112: G. W. Rathenau: Saturation and magnetization of hexagonal iron oxide compounds (Rev. mod. Phys. **25**, 297-301, 1953)

Review of results formerly published elsewhere (see Philips tech. Rev. **14**, 1952/53 and these Abstracts No. 2059). The value of the saturation magnetization at absolute zero of $BaO.6Fe_2O_3$ and related compounds can be explained as due to non-compensated anti-ferromagnetism. The Bloch wall formation in small particles is discussed. One is led to the assumption that in these materials Bloch walls are nucleated at imperfections. In specimens containing randomly oriented large crystals, Bloch wall formation becomes appreciable at a positive field strength of the order $4\pi I_s$. By orienting the crystals in a magnetic field $(BH)_{max}$ values of 3×10^6 gauss oersted have been obtained. The critical diameter for wall formation changes with temperature. An excess of walls formed at a different temperature from the temperature of measurement may persist in a metastable equilibrium.

2113: E. W. Gorter and J. A. Schulkes: Reversal of spontaneous magnetization as a function of temperature in Li-Fe-Cr spinels (Phys. Rev. **90**, 487-488, 1953)

Néel's theory of non-compensated anti-ferromagnetism predicts that in some materials the spontaneous magnetization should change its sign below the Curie temperature, since it is the difference between the anti-parallel magnetizations of the two sublattices, each of which may vary differently with temperature. This behaviour has been established for a series of Li-Fe-Cr spinels.

2114: J. S. van Wieringen: Anomalous behaviour of the g-factor of Li-Fe-Cr-spinels as a function of temperature (Phys. Rev. **90**, 488, 1953)

The g-factor of the spinel mentioned in abstract No. 2113, which has the value 2 at 0°K, rapidly decreases above 200°K. Starting from high temperatures (450°K), at which g is likewise about 2, g increases as the temperature is reduced and approaches infinity at $T = 337$°K. A qualitative explanation of this behaviour is derived from Kittel's g-formula and from the change of the spontaneous magnetisation as a function of temperature.

2115: J. M. Stevels: Le verre considéré comme polymère (Verres er Réfractaires **7**, 91-104, 1953)

Certain properties of silicate-glasses do not depend on the kind of network-modifying ions, but mainly on the number of bridging oxygen ions per tetrahedron, Y, a number which also indicates the number of contact points of each tetrahedron: Y, therefore, is a measure of the internal coherence of the network. The physical properties of the glass are very considerably influenced by this internal coherence of its network: glasses whose Y-value lies between 4 and 3 are machnically strong and chemically resistant; at values between 3 and 2 the glass becomes less and less resistant; if $Y = 2$ chains of unlimited length occur, held together by the network-modifying ions, thus constituting an extremely weak lattice, highly susceptible to thermal and mechanical influences. If Y is smaller than 2,

the length of the chains is no longer unlimited; silicates whose chains possess a high degree of symmetry may crystallize in spite of their great length. The phosphates, on the other hand, having a non-symmetrical structure, may still yield vitreous substances at Y-numbers as low as 1.6. As regards silicates, however, a small disturbance of the symmetry may suffice to prevent crystallisation: similar situations are frequently met in the chemistry of organic polymers. These considerations are confirmed by the examination of the dielectric losses of the glass in a H.F.electric field. At very low temperatures "deformation losses" of the network are found to occur, due to the sudden transition into another position of certain portions of the network chains. As the Y-value of the glass decreases, the deformation losses as a rule become greater, whilst the maximum of these losses as a function of the temperature is shifted to increasingly high temperatures.

2116: W. J. Oosterkamp: General considerations regarding the dosimetry of roentgen and gamma radiation (Appl. sci. Res. **B3**, 100-118, 1953 and addendum Appl. Sci Res. **B3**, 477 1954).

A distinction is made between the quantities "dose" (ionising ability, measured in roentgen), and "absorbed dose" (energy imparted by the radiation measured in rads or ergs per gramme). The methods by which both quantities can be measured are analysed. The measurement of dose at photon energies above 3 MeV has not yet been realised. The correlation between dose and absorbed dose is discussed. In air-equivalent or nearly air-equivalent material at moderate photon energies, dose is a fair measure of absorbed dose; in non air-equivalent tissue, the differences in mass absorption coefficient between this tissue and air should be taken into account; at discontinuities in the atomic composition, the increased generation of secondary electrons in materials with higher atomic number will also cause an increased ionization in neighbouring tissues with lower atomic number; at photon energies above 1 MeV there is an increasing discrepancy between dose and absorbed dose at the same place. (In the first-mentioned paper the author used the terms "irradiation" instead of "dose", and "dose" in place of "absorbed dose". The terms used in the second-mentioned paper and in this abstract are those since adopted by the International Commission on Radiological Units.)

2117: B. D. H. Tellegen: Synthesis of $2n$-poles by networks containing the minimum number of elements (J. Math. Phys. **32**, 1-18, 1953)

Brune's method for the synthesis of two-poles is extended to the synthesis of $2n$-poles. This leads to networks with the minimum number of reactances and the minimum number of elements.

2118: J. I. de Jong and J. de Jonge: Kinetics of the hydroxymethylation of phenols in dilute aqueous solution (Rec. Trav. chim. Pays-Bas **72**, 497-509, 1953)

The hydroxymethylation of phenols has been investigated in the pH range 1-11 and between 70-130°C in dilute aqueous solution. The reaction appears to be bimolecular. Below about $pH = 4$ the rates were found to be proportional to the concentration of the hydrogen ions, while a proportionality to the concentration of the hydroxyl ions was observed in more alkaline solutions. An influence of the concentration of buffers on the rates has not been observed. The presence of small amounts of triethanolamine does not influence the rate of the reaction. The mechanism of the reaction is discussed.

2119: J. D. Fast: Low-hydrogen welding rods (Welding J. **32**, 516-520, 1953).

See Philips tech. Rev. **14**, 96-101, 1952/53.

2120: J. I. de Jong: A determination of methylol groups in condensates of urea and formaldehyde (Rec. Trav. chim. Pays-Bas **72**, 652-654, 1953)

Methylol groups in condensates of urea and formaldehyde may be determined using an alkaline solution of potassium cyanide. Excess cyanide is back titrated with mercuric nitrate.

2121: N. W. H. Addink: Quantitative spectrochemical analysis by means of the constant-temperature D.C. carbon arc (Spectrochim. Acta **5**, 495-499, 1953)

Survey of the method developed by the author, of using complete evaporation in the D.C. carbon arc and projection of the spectrum on a standard paper density scale (s.p.d. scale). Some results are compared with those of conventional chemical analysis. See also Philips tech. Rev. **12**, 337-348, 1950/51.

Vol. 16 No. 9, pp. 241-272 MARCH 1955

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

THE TWIN-CONE MOVING-COIL LOUDSPEAKER

by J. J. SCHURINK.

621.395.623.742

In amplitude modulation radio reception, the upper limit of the frequencies reproduced lies at about 4500 c/s. On the other hand, in F.M. radio reception as well as in the reproduction of recordings on disc or tape, the higher tones up to 18,000 c/s can be reproduced quite well. In order to do justice to this extra range, the response of the loudspeaker must be extended by at least one octave. A very satisfactory way of doing this is by means of the twin cone moving-coil loudspeaker, a description of which is given below.

Introduction

The range of frequencies to which the human ear is sensitive extends from about 20 to 18,000 c/s. For ideal reproduction, therefore, frequencies should be reproduced at their correct strength throughout this range. Up till now, however, this standard of ideal reproduction has rarely been reached in radio sets and gramophones. Neither the lowest nor the highest frequencies are reproduced really properly. In this article we shall concern ourselves chiefly with improvements in the treble response.

Until recently there was little demand for good response to the higher audio frequencies for, although amplitude-modulated radio transmissions contain high audible frequencies, a large part of the sound spectrum is cut off in the receiver (above about 4500 c/s) in order to minimize interference by other transmitters operating on adjacent wavelengths. In gramophone reproduction, moreover, the older records (consisting of shellac mixed with a filler) were capable of modulation almost up to the limit of audibility, but it was necessary to attenuate or even cut off a considerable part of the higher frequencies in order to reduce needle hiss and distortion.

Recent developments in frequency modulation and in methods of record production have changed this situation. In F.M. radio transmissions, the carrier frequency is so high (about 100 Mc/s) that there is no longer crowding of the frequency band, so that the higher audio frequencies can be received without interference. The use of new materials (plastics) for gramophone records has appreciably reduced surface noise, and this, too, makes it possible to reproduce much higher audible frequencies [1]). High tones can also be recorded and reproduced by means of magnetic tape, especially now that improved tapes are now available which reduce the previously troublesome modulation noise [2]).

Under these new conditions, the frequency response of the loudspeaker now becomes a major factor in determining the width of sound spectrum actually reproduced. The upper limit of most present loudspeakers is in the region of 8000—9000 c/s, and this limit has been high enough to ensure that the attenuation of the higher frequencies hitherto occurring in radio and gramophone reproduction was not actually accentuated by the loudspeaker itself. The noise that occurs in a certain limited range of frequencies (the so-called "coloured" noise) is thereby also avoided.

In view of the above-mentioned improvements in radio and gramophone reproduction it is now necessary to raise the upper limit of the loudspeaker

[1]) See L. Alons, New developments in the gramophone world, Philips tech. Rev. **13**, 134-144, 1951/52.

[2]) D. A. Snel, Magnetic sound recording equipment, Philips tech. Rev. **14**, 181-190, 1952/53; W. K. Westmijze, Principles of the magnetic recording and reproduction of sound, Philips tech. Rev. **15**, 84-96, 1953/54.

response to 16 000—18 000 c/s. The limit of auditory sensitivity in persons under 30 years of age is about 18 000 c/s; in older persons it is lower (16 000 c/s at 40, and so on down to less than 10 000 c/s). It is not known precisely to what degree the frequency range of reproduced music might be usefully extended beyond the audible spectrum, but an extension from 9000 to 18 000 c/s already represents another octave.

Moving-coil loudspeakers

The moving-coil cone loudspeaker is the most widely used of all the different types developed so far, because of its many excellent properties: efficiency, simple construction, reliability, consistency in manufacture, and long life (as long as 20 years).

The paper cone of this type of loudspeaker has freedom of movement axially within the limits of the flexibility of the suspension, which is by means of corrugations in the edge of the cone and a flexible centring ring near the apex. The speech coil is attached to the apex and is situated in a radial magnetic field (*Fig. 1*).

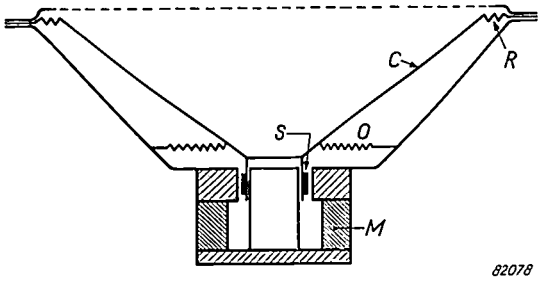
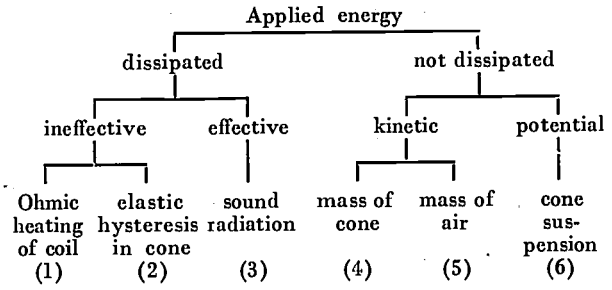


Fig. 1. Cross-section of a conventional moving-coil loudspeaker. *C* cone, *R* corrugated rim, *O* centring spider, *S* speech coil, *M* magnet.

Fairly large forces are required to move the cone rapidly and, hence, also a fairly strong signal current in the speech coil. As this coil necessarily has a finite, though small resistance, ohmic losses (1) occur in the coil. A small amount of energy (2) is dissipated in the cone suspension as a result of the repeated changes in shape. The rest of the applied energy (3) is dissipated in the form of radiation, i.e. sound energy [3]). There is in addition, an exchange of the kinetic energy between the moving cone (4) and the surrounding air (5), and the potential energy (6) in the suspension, which undergoes elastic deformation. This energy is not dissipated.

All these elements are shown in the following energy balance diagram.



In order to assess the quality of different loudspeakers we examine the frequency response curve to determine the efficiency η (radiated power divided by input power) as a function of the frequency f. η is usually only a few per cent [4]).

There are four fairly sharply-defined frequencies which are particularly important in the response curve, viz. the resonant frequency f_r, the frequency f_1 at which the wavelength is comparable to the dimension of the baffle, the frequency f_2 at which the wavelength is comparable to the dimensions of the paper cone, and, lastly the frequency f_3 above which the different parts of the cone are no longer in phase, and complex modes of vibration occur. To give a rough idea of values, f_r might be 50 c/s, $f_1 \simeq 300$ c/s, $f_2 \simeq 600$ c/s and $f_3 \simeq 1000$ c/s.

Table I. Dependence of the amplitude x_m and peak velocity v_m of the cone, and the radiated power P on the frequency, for constant speech coil current. P_d (d for dipole) refers to a loudspeaker without baffle, P_w to the speaker with infinite baffle, and P_k to the speaker with baffle or cabinet of finite dimensions.

	$f<f_r$	$f>f_r$		
v_m	$\propto f$	$\propto f^{-1}$		
x_m	const.	$\propto f^{-2}$		
		$f_r<f<f_2$		$f>f_2$
P_d	$\propto f^6$	$\propto f^2$		$\propto f^{-2}$
P_w	$\propto f^4$	const.		$\propto f^{-2}$
		$f_r<f<f_1$	$f_1<f<f_2$	
P_k	$\propto f^6$	$\propto f^2$	const.	$\propto f^{-2}$

Table I above shows the (theoretical) frequency dependence of the efficiency η, or (what is roughly the same thing), the power P radiated, with a constant alternating current applied to the speech coil The way in which the amplitude of the deflection (x_m) and the peak velocity (v_m) vary with the frequency is also shown.

[3]) A. Th. van Urk and R. Vermeulen, The radiation of sound, Philips tech. Rev. 4, 213-222, 1939.

[4]) J. de Boer, The efficiency of loudspeakers, Philips tech. Rev. 4, 301-307, 1939.

The equation of motion of the cone

For a current i passing through the speech coil, a force $K = Bli$ operates on the coil in the direction of the axis (l denotes the length of wire in the speech coil and B the induction in the air-gap). If the current i is an alternating current of frequency f ($= \omega/2\pi$), the cone will vibrate. The amplitude v_m of this vibration, for a given value of i, reaches a maximum at the resonant frequency f_r, which is determined mainly by the stiffness S of the cone suspension and the total mass M (the latter is the sum of two components, the mass M_c of the cone and coil, and that of the air that moves with the cone, Hence:

$$\omega_r = 2\pi f_r \approx (S/M)^{\frac{1}{2}}. \quad \ldots\ldots\ldots \quad (1)$$

The movement of the coil sets up in the surrounding air an alternating pressure p which in turn exerts a force on the cone equal to $K_p = \pi R^2 p$, where R is the radius of the circular edge. The ratio of this force to the velocity v of the cone is known as the mechanical impedance Z. If we represent the speech current by $i = i_m \exp(j\omega t)$, then Z has the complex value

$$Z = Z_r + j\,Z_i. \quad \ldots\ldots\ldots\ldots \quad (2)$$

where Z_r and Z_i are functions [5]) of kR, and $k = \omega/c = 2\pi f/c = 2\pi/\lambda$ (c = velocity of sound, λ = wave length).

In the region where $kR<1$, for a cone with an infinite baffle, Z_r is proportional to f^2 and, for a cone without baffle, to f^4; in both cases Z_i is proportional to f. In the region $kR>1$, for both the two cases, Z_r is practically constant and Z_i is proportional to f^{-1}.

The transition between $kR<1$ and $kR>1$ occurs at a frequency $f_2 = c/2\pi R$. As this is near to the frequency f_3 above which the vibrations of all parts of the cone are no longer in phase (a phenomenon which we shall discuss in more detail later), there is little object in applying the above theory in the region where $kR>1$. We shall therefore confine ourselves mainly to the region where $kR<1$ (and hence $f<f_2$). This region can be subdivided into two parts: $f<f_r$, and $f_r<f<f_2$.

Apart from the external force K and the air reaction K_p acting on the cone, the mass of which is M_c, there is also the force $-Sx$ due to the stiffness of the cone (where x denotes its deflection), the frictional force of the suspension $-hv$ and the inertial force $-M_c\ddot{x} = -M_c\dot{v}$.

Since all the factors involved are proportional to $\exp(j\omega t)$, and writing $\ddot{x} = \dot{v} = j\omega v$, $v = \dot{x} = j\omega x$, it follows by equating the sum of all the forces to zero, that:

$$v = \frac{K}{j\omega M_c + h + \frac{S}{j\omega} + Z}, \quad \ldots\ldots \quad (3)$$

where $Z = Z_r + jZ_i$. The term jZ_i can be added to $j\omega M_c$. The term Z_i/ω (which is constant when $f<f_2$) represents the equivalent mass of the displaced air. The real part Z_r of Z is of the same nature as h, the internal friction: both represent dissipated energy. Z_r represents the radiated energy i.e. the useful sound energy radiated. The radiated power is:

$$P = Z_r\overline{v^2} = Z_r\, v_{rms}{}^2 = \tfrac{1}{2} Z_r\, v_m{}^2. \quad \ldots\ldots \quad (4)$$

Let us now assume that $K = K_m \exp(j\omega t)$ and that $K_m =$ constant. Since, when $f<f_r$, Z_i is proportional to f, and Z_r to a power of f, we see that in the region $f<f_r$ the most significant contribution to v is the stiffness term $S/j\omega$, so that in this region, the peak velocity v_m is proportional to f.

Above the resonant frequency, that is where $f_r<f<f_2$, the term $j(\omega M_c + Z_i)$ is dominant, this being proportional to f. In this range, therefore, v_m is proportional to f^{-1}. When $f \approx f_r$, v_m reaches a maximum.

It follows from the above that in the region where $f<f_r$, the amplitude of the cone ($x_m = v_m/\omega$) is practically constant and further, that when $f_r<f<f_2$, x_m is proportional to f^{-2}.

The value of P as a function of f varies according to whether the loudspeaker radiates freely (dipole), or whether it is placed in an infinite baffle. In the former case, when $f<f_2$, P varies as $\omega^4 v_m{}^2$, so that $P \propto f^6$ where $f<f_r$ and $P \propto f^2$ where $f_r<f<f_2$.

For a loudspeaker with infinite baffle, where $f<f_2$, P is proportional to $\omega^2 v_m{}^2$; hence when $f<f_r$, $P \propto f^4$ and when $f_r<f<f_2$, $P \approx$ constant. The power radiated at a large distance from the loudspeaker ($r \gg R$) per unit area is equal to $p_r{}^2/\varrho c$, where p_r is the pressure at a distance r, and ϱ the density of the air. The total power is found by integration of the term $p_r{}^2/\varrho c$, over a sphere of radius r. Hence the quantity $p_r{}^2$ when integrated over the solid angle 4π (i.e. $\int r^2 p_r{}^2 d\Omega$) has the same frequency dependence as the power P, as derived above.

When the loudspeaker is mounted on a baffle of finite dimensions it behaves more or less as if the baffle were infinitely large at the higher frequencies, whilst in the lower frequencies it behaves as a dipole radiator. The boundary between the two conditions occurs roughly at f_1, the frequency at which the length L of the air path around the baffle is equal to $\frac{1}{2}\lambda$, i.e.

$$f_1 = \frac{c}{2L}. \quad \ldots\ldots\ldots \quad (5)$$

It is possible to construct the electrical analogue of equation (3) (*fig. 2*). In this, the force $K = Bli_m.\exp(j\omega t)$ represents the "alternating voltage" which produces the "current" in the network. In the diagram, $R_2 = h$, $R_3 = Z_r$, $L_4 = M_c$, $L_5 = Z_i/\omega$, $C_6 = 1/S$. (The indices correspond to the various factors shown schematically in the energy-balance diagram on p. 290.) [6]).

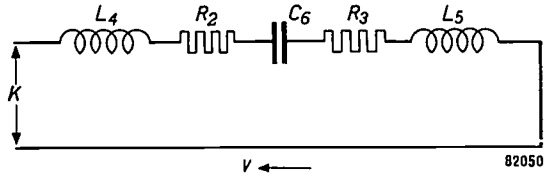


Fig. 2. Electrical analogue of a moving-coil loudspeaker.

The frequency response curve in practice

Having now discussed the theoretical frequency dependence of the various factors relevant to the

[5]) See article referred to in footnote [3]), in which Z_r and Z_i are shown plotted against k for a number of different cases.

[6]) For acoustical analogues, see H. F. Olson, Elements of acoustical engineering, van Nostrand, 2nd Ed., New York, 1947 (particularly Ch. VI). Olson employs for the "current" the acoustical quantity $U = Av$, (where $A = \pi R^2$, and U is the volume swept through by the cone per second), and for the "voltage" Bli/A (having the dimension of a pressure). The impedances must then all be divided by a factor A^2.

In the diagram in fig. 2 the ohmic heat developed in the speech coil (the major source of loss) is not expressed. This could be done by shunting the impedance circuit with a resistance R_1 such that K^2/R_1 would be equal to $R_s i^2$, where R_s denotes the resistance of the speech coil.

operation of a loudspeaker, we may now consider the form of the actual response curve and how the desired characteristics can be attained.

First, however, something should be said about the measurement of the energy radiated by the loudspeaker. It is usual to measure the sound pressure at a certain point, say on the axis, at a certain distance from the speaker. If the speaker were capable of radiating the sound equally in all directions, the frequency dependence of the sound pressure plotted logarithmically would be the same as that of the output power, which is everywhere proportional to the square of the sound pressure.

Now, the radiation diagram is dependent on the frequency, the main cone of sound generally becoming narrower as the frequency is increased. Hence the sound pressure curve as measured on the axis lies at a higher level than the power curve at the higher frequencies ($f > f_2$).

The frequency response curves accompanying this article all represent the sound pressure at the axis as a function of frequency, with an alternating current of constant amplitude flowing in the speech coil. We shall not dwell on the method of measurement employed in plotting these curves; it is sufficient to say that a constant current of uniformly increasing frequency is applied to the speech coil, and that the sound pressure as measured by means of a microphone and amplifier is recorded on a paper strip, whose lateral movement is made this way the whole frequency spectrum is covered to correspond to the increase of frequency [7]). In in about 40 seconds. At a certain frequency the speech coil current is momentarily cut off to provide a marker in the form a of dip in the curve, to check that the curve as traced out is correctly placed with respect to the frequency scale. This interruption can be clearly seen in *fig. 3* (at *M*), as well as in figs. 6 and 9. [8])

Although we are primarily concerned in this article with reproduction of the higher frequencies, we shall in passing briefly mention the middle register and lower frequencies as well.

Reproduction of the low tones

The lowest tone that the loudspeaker is capable of reproducing without distortion is determined by the resonant frequency, which in turn depends on the stiffness (of the suspension of the cone at the edge and apex), and on the total mass (cone, coil and air resistance). Although the resonant frequency is governed to some extent by the size of the cone, and still more so by the maximum load for which the speaker is to be designed, it is not very difficult to make the resonant frequency so low that, as far as the loudspeaker itself is concerned, faithful reproduction of the lowest notes occurring in music will be guaranteed. A 12″ loudspeaker with a power handling capacity of 15 W can have a resonant frequency as low as about 20 c/s, and an 8″ speaker rated for 10 W a resonant frequency of 45 c/s.

Below the resonant frequency, the output drops by 12 db (a factor of 16) per octave, for a loudspeaker with "ideal" (infinite) baffle.

In practice, however, the speaker is generally mounted on a baffle of limited size or in a radio cabinet and, under these conditions, an additional drop in output will occur at very much higher frequencies. This decrease, which is about 6 db per octave (a factor of 4 in the acoustic output power) is due to the fact that the air is able to flow round the baffle, from the front to the back of the cone and vice versa. It begins to be noticeable at the frequency f_1.

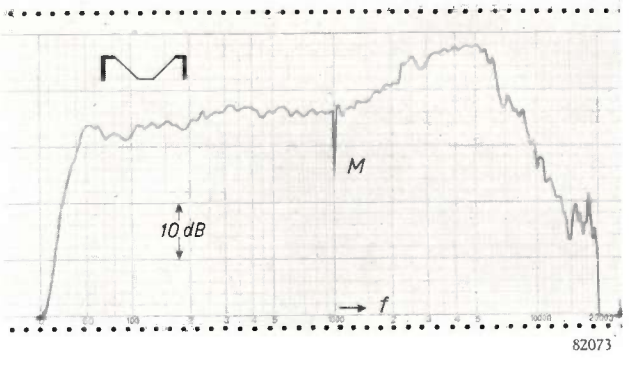


Fig. 3. Response curve of a conventional 8″ loudspeaker, showing on a logarithmic scale the sound pressure p on the axis, for a constant current in the speech coil. The frequency f plotted horizontally also has a logarithmic scale. The marker *M*, obtained by interrupting the speech current (at 1000 c/s) serves to register the curve correctly with respect to the frequency scale.

This drop of 6 db per octave (in the range $f_r < f < f_1$) can be compensated at frequencies which are not too low, say, at 1 octave below f_1, by modifying the response curve of the amplifier. This can be done even in low-power amplifiers, at least for low output levels (room strength), where, of course, it is most needed. A section of the input potentiometer is shunted by a resistor and capacitor in series, this having the effect of attenuating the higher frequencies as compared with the low (automatic bass compensation).

[7]) See also R. Vermeulen, The testing of loudspeakers, Philips techn. Rev. 4, 354-363, 1939.

[8]) In figs 3, 6, 9 and 10 the bold horizontal lines (in fig 10 circles) are spaced at a distance equal to 10 dB, which means that from one line to the next, the pressure p increases by a factor of $\sqrt{10}$, and p^2 by a factor of 10.

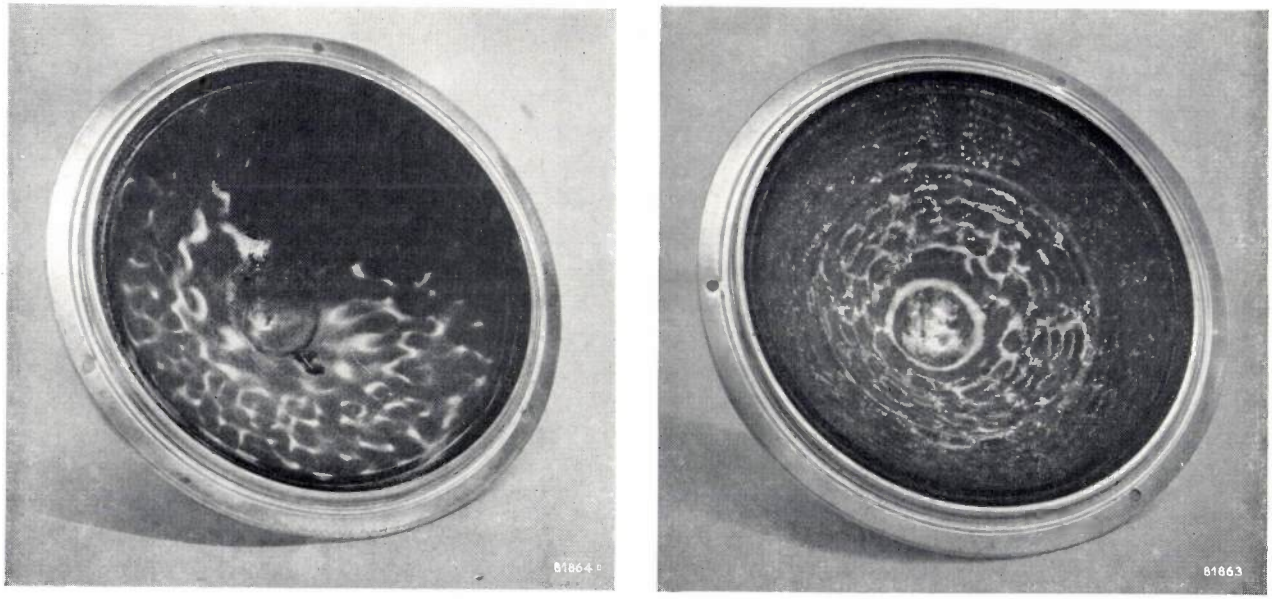
In the case of a loudspeaker mounted in a cabinet or on a small baffle there is little object in striving for a low resonance; it may even be preferable expressly to raise the resonance point, say to one octave below f_1.

Reproduction of the medium frequencies

Above the resonance point (with constant current) the output of a loudspeaker on a large baffle is practically constant with the frequency, at least for $f<f_2$. In the range $f>f_2$, the output theoretically decreases (see table 1). This applies, however, only to that part of the spectrum in which all the parts of the cone vibrate in phase. The upper limit of this range (frequency f_3) usually occurs at 1000 c/s. being somewhat lower in large cones, and higher in small ones.

Reproduction of the higher frequencies

A judicious choice of material for the cone will ensure a vibration mode that yields a gradual and uniform rise in the sound pressure curve in the region beyond 1000 c/s. The rise in response is desirable in order to give satisfactory reproduction up to 6000—8000 c/s. At higher frequencies than this the curve drops sharply because that part of the cone where the vibration is in phase ultimately becomes too small, and also because it is unfavourably placed for diffusion, lying as it does at the deepest point in the cone. As a result, a peak occurs in the region of 6000—8000 c/s which in some of the few publications on this subject is sometimes referred to as the "resonance point of the cone itself". This is incorrect, in that the peak in question is determined by the mass of only a part of the cone



a *b*

Fig. 4. Complex vibration patterns (Chladni figures) in a loudspeaker, *a*) at 4000 c/s, *b*) at 12 000 c/s.

Above 1000 c/s the cone exhibits complex modes of vibration, and only an irregularly defined area round the speech coil is then in phase. The nodal lines can be rendered visible by sprinkling a fine powder over the cone and allowing it to vibrate for some time at the same frequency; vibration patterns are thus obtained like the well-known Chladni figures (*fig. 4*).

In consequence of this behaviour the effective mass of the cone becomes smaller with increased frequency, and the velocities accordingly higher, resulting in a rise in the pressure curve.

and also by the stiffness of the surrounding material.

The rise in the pressure curve with frequency which is needed to compensate for the decrease in output that accompanies a rise in frequency should be about 6 db per octave ($p \propto f$). It is difficult to accomplish this by mechanical means. As a rule about 4 db is accepted, the difference being made good by a rise in the electrical characteristic of the amplifier.

If the loudspeaker is to be used with a conventional type of amplifier, which for a constant input signal delivers an output voltage independent of

frequency (i.e. an amplifier with negative voltage feedback), it must also be borne in mind that the current in the speech coil will be dependent on the impedance of this coil. In most loudspeakers the impedance rises steeply with the frequency. Philips manufacture so-called "constant impedance" loudspeakers: the impedance at 20 kc/s is at most 40% higher than that at 1000 c/s. With "rising impedance" loudspeakers, the impedance at 20 kc/s can easily be 5 times as high as that at 1000 c/s. In *fig. 5* constant and rising curves are compared.

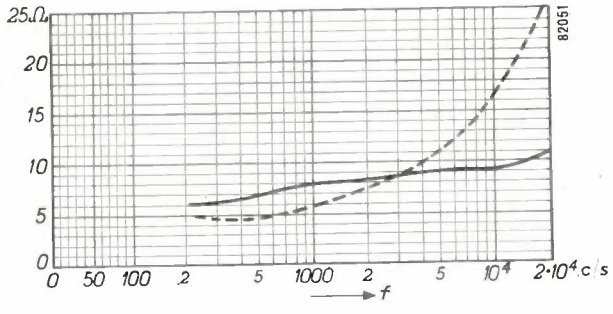


Fig. 5. Speech coil impedance as a function of f in a loudspeaker with "constant" impedance (continuous line) and, for comparison, the curve for a conventional loudspeaker with rising impedance (broken line).

Distortion

Non-linear distortion has its origin in two main causes. One of these is the fact that the restoring force of the cone is not proportional to its displacement (i.e. the stiffness of the suspension is not constant, but increases with the displacement). The other source of non-linear distortion can be traced to the fact that the force acting on the coil depends not only on the current, but also on the position of the coil in the magnetic field (because the latter is not quite homogeneous).

To obtain the optimum ratio between the current in the speech coil and the force acting on it, the ratio of the air-gap depth to the wound length of the speech coil must be carefully chosen; the form of the field in the air-gap is also important. Proportionality between the restoring force and the displacement is ensured by a judicious choice of material for the cone and a suitable stiffness ratio between the two suspension elements.

Non-linear distortion is manifested in various ways. If it is present below 1000 c/s, a sinusoidal signal applied to the loudspeaker will exhibit overtones (higher harmonics). If the rest of the spectrum is taken into consideration as well, intermodulation distortion becomes important; this takes the form of spurious frequencies $f_a + f_b$ and $f_a - f_b$, when both the frequencies f_a and f_b occur simultaneously in the input signal. The difference frequency $f_a - f_b$ can be particularly troublesome if f_a and f_b are high and $f_a - f_b$ occurs in the middle register.

Intermodulation distortion is also influenced to some extent by the shape of the response curve. The more irregular the curve, with many peaks and dips, the more distortion will be present. If non-linear distortion is to be avoided, moreover, the position of the speech coil with respect to the magnetic field, when the coil is at rest, is important. (The centre of the coil should coincide with the strongest part of the field). Further, as already stated, the amplitude of the cone should not be too great; for this reason a large baffle is desirable, because this ensures a low resonant frequency (effect of the air mass), and small amplitude (effect of damping resistance) [9]).

All Philips loudspeakers are very carefully designed from the point of view of distortion; the more expensive types are of course better in this respect but, used under the proper conditions, they are all of a very high standard.

Extension of the frequency range

When the problem of extending the frequency response by a whole octave first arose, efforts were first made to modify the conventional moving-coil loudspeaker — which had already proved its worth up to 8000 c/s — such that it would also meet the new requirements.

It was necessary not only to increase the sensitivity in the higher frequencies, but also to improve the radiation pattern at those frequencies, so as to ensure a better distribution of sound in the higher tones. With a single cone the sound radiation is confined to an increasingly narrow beam as the frequency becomes higher. One method of achieving a wide diffusion angle is to employ a wide angle cone (remembering, however, that, owing to standardisation, we are limited to certain cone dimensions). A slight increase in the range (up to 12 000 c/s) was obtained by "hardening" certain parts of the cone, but this in itself was not sufficient, particularly as it did nothing to improve the diffusion angle.

---

[9]) The damping at the resonance point is also important in relation to the behaviour of the cone when acted upon by a discontinuous force, e.g. a sinusoidal force which commences or ends abruptly (transient response). Apart from the effect of the cone fabric (peaks in the response curve), the build-up and decay of the vibration is very dependent on the damping.

A combination of two loudspeakers

It is known that the response characteristic in the higher frequencies is improved by reducing the size of the cone and coil, provided that their masses are kept as small as possible, and that a slightly stiffer cone material is used. A small speech coil ensures that the cone is excited more towards the apex, and so improves the mechanical coupling between cone and coil. A small loudspeaker constructed along these lines will quite effectively reproduce frequencies up to 15000 c/s, but, owing to the small dimensions, reproduction of the bass is poor. The obvious solution to the problem is therefore to use a small loudspeaker for frequencies above 8000 c/s in series or in parallel with a conventional loudspeaker for the range up to 8000 c/s.

If at the same time the angle of the cone of the small speaker (or "tweeter") is made as wide as possible, a better radiation pattern will be obtained in the higher frequencies; in this connection, the small diameter of the cone with respect to the wavelength also plays an important part.

However, simple and effective as this solution might appear to be, there are objections to it in practice.

In the first place the efficiency of the tweeter is much lower than that of the larger speaker. The area of the air-gap is necessarily smaller; the induction in the gap is also smaller (owing to the saturation of the soft iron core). The product of induction and air-gap area (a measure of the total effective flux, and hence also of the efficiency), is therefore smaller. Reproduction of the higher frequencies is certainly improved, but the general level is too low, so that, because of the acoustic "masking" effect whereby the ear accomodates itself to the sound level in the middle register, the extension in the frequency spectrum obtained in this way does not adequately fulfil its purpose.

The only means of obviating this would be to reduce the efficiency of the large speaker, but this would in turn involve increasing the electrical output of the amplifier to yield the same acoustic output, which would of course be uneconomic.

In some countries efforts have been made to minimize the extra cost of the higher frequency response by using a simpler type of instrument for the tweeter, such as an electrostatic (condenser) or a crystal type of loudspaker, but there are objections to both, viz. considerable distortion, and sensitivity to climatic conditions. If high fidelity is required in the high frequencies, the use of such speakers is not advisable.

Apart from this, there is another drawback in the use of two loudspeakers, which is quite as important as the others. As the two cones do not vibrate in phase, and the phase shift is dependent on both frequency and direction, interference occurs, as a result of which some parts of the spectrum in the cross-over region between the two speakers are appreciably attenuated, depending upon the listening position (cross-over dip, see *fig. 6a*). This can be avoided by placing the speakers some distance apart; interference minima and maxima are thus reduced and the response curve is more uniform, but the separation must be a matter of some yards and the speakers cannot be housed in the same cabinet.

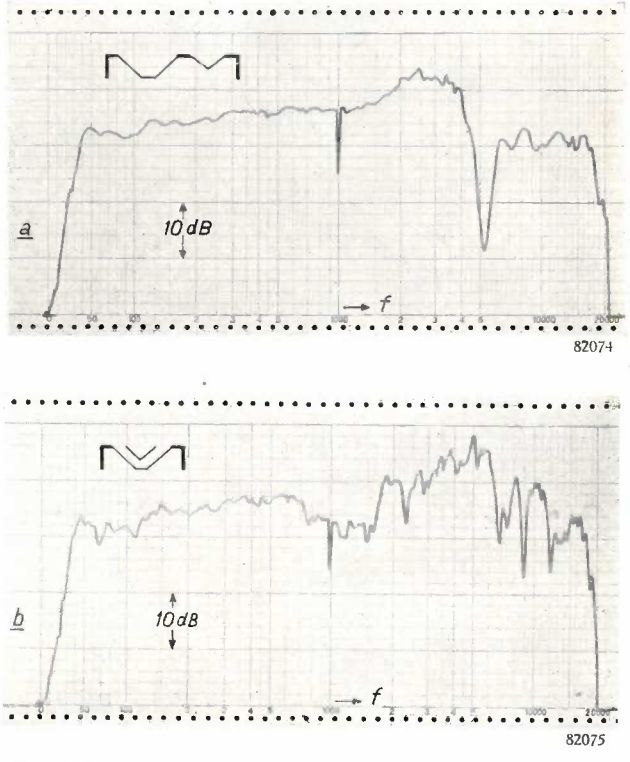


Fig. 6. *a*) Response curve (*p* plotted against *f*, for constant speech coil current) for a tweeter and main loudspeaker mounted side by side in the same cabinet.
b) Response curve with tweeter placed inside the cone of the large speaker.

Nor is it good practice to mount a tweeter in the space within the cone of the larger speaker, as will be seen from the response curve in fig. 6*b*.

The twin cone loudspeaker

A solution that does give satisfactory results is a single loudspeaker with two cones, that is, a small cone inside, and attached to, the large one (*figs. 7* and *8*). Development of this design is based on the assumption that this is the only way, with the two cones driven by the same coil, in which the

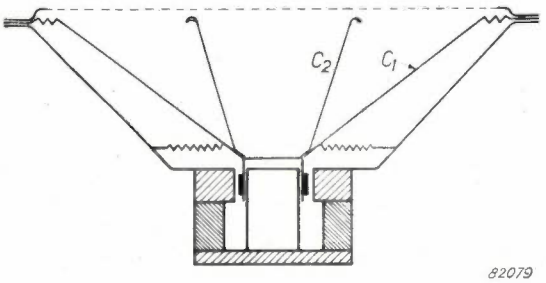


Fig. 7. Cross-section of twin-cone loudspeaker (C_1 outer cone; C_2 inner cone).

problem of the level of sound in the higher frequencies can be satisfactorily solved. Correct proportioning (mass and stiffness) of the inner cone then ensures a response curve that reveals no irregularities in the transition region (*fig. 9a*). It is an advantage from the point of view of continuity of response between the middle register and treble that the mass of the large cone is increased by the addition of the small one. Furthermore, the latter, in the frequencies between 3000 and 8000 c/s (for which the large cone is effective) acts as a diffuser, thus improving the radiation diagrams.

Beyond 8000 c/s the small cone takes over and continues up to 18 000 c/s; the large cone then serves as a reflector to improve the radiation diagram in this range. This is illustrated in *fig. 10*, which shows the radiation patterns of the same loudspeaker with and without inner cone.

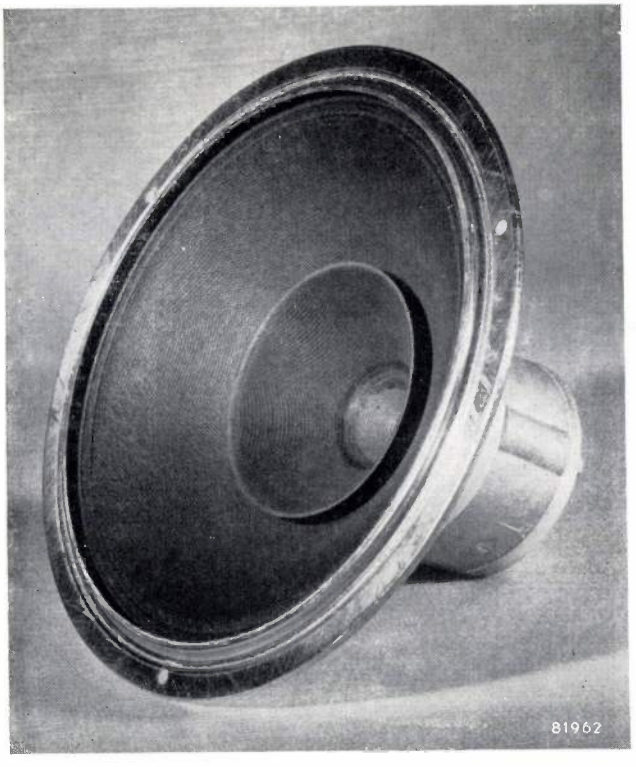


Fig. 8. Photograph of twin cone loudspeaker (Type 9710 M).

The twin cone has only one disadvantage — if it is fair to call it a disadvantage — and this is that the smaller cone cannot be switched out. If, for any reason, a part of the treble is not wanted (as in A.M. radio reception, or in the reproduction of

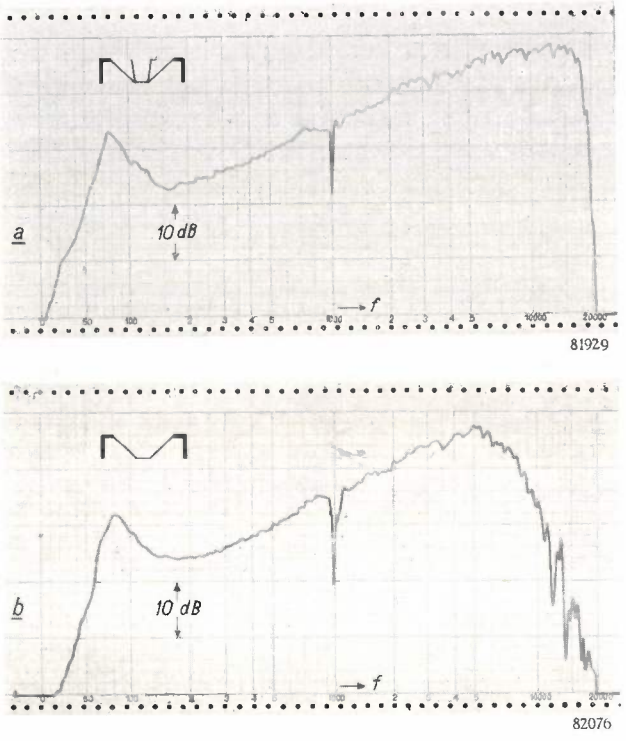


Fig. 9. *a*) Response curve of twin-cone loudspeaker (without baffle). b) Response curve of the outer cone only.

poor quality gramophone records where needle hiss needs to be suppressed), the amplifier will have to incorporate an effective treble control, capable of attenuating or even suppressing the higher frequencies (say, above 4000 c/s). Even in this case, however, the advantage of the diffusive effect of the inner cone on the sound in the range below the cut-off frequency (3000—4000 c/s) is retained.

The manner in which the improvements in the high-frequency response were obtained may be summarized as follows. The complex modes of vibration of the cone at frequencies over 1000 c/s, which improve the response curve up to about 8000 c/s, in itself a favourable feature, gives some trouble at the highest frequencies, as the operative part of the cone is then too small and too deep within the cone. The addition of the small extra cone constitutes, as it were, as extension of the control part of the main cone, which not only increases the radiating area, but also this area (and particularly the edge of the small cone) is brought more to the fore.

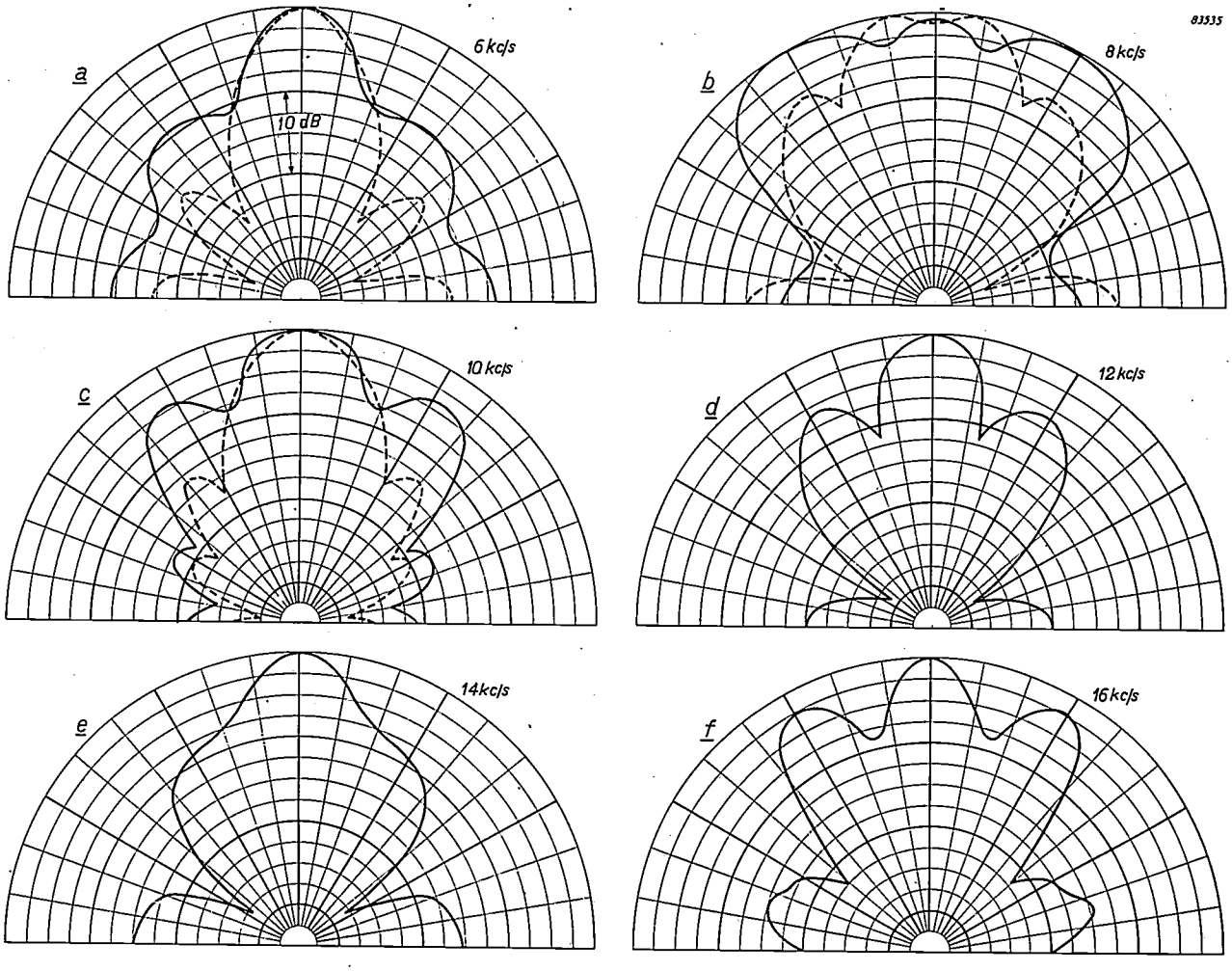


Fig. 10. Radiation polar diagrams of twin cone 8″ loudspeaker (dotted lines refer to outer cone alone), at frequencies of 6, 8, 10, 12, 14 and 16 kc/s. The pressure (at a constant distance) is plotted radially on a logarithmic scale.

Finally it may be noted that in certain cases it may nevertheless be found desirable to employ two loudspeakers, each reproducing a certain range of frequencies. This is usually done to overcome imperfections in the amplifier (electrical intermodulation distortion). A two channel amplifier is then used. As a twin cone loudspeaker connected to the treble channel will also reproduce the lower frequencies sufficiently strongly, the cross-over point may be made fairly low (e.g. 300 c/s), which is desirable because it facilitates the compensation of phase differences. However, if the loudspeakers are mounted too close together in the same cabinet the same objections apply as described for the use of the large speaker with separate tweeter. These objections can be obviated by employing networks to ensure that the phasing of the two channels is the same at the cross-over point. The phase difference can be reduced to less than 5° in this way, thereby eliminating the interference referred to above.

Summary. In view of the improvements in the reproduction of sound offered by F.M. radio, modern gramophones and tape recorders, an extension of the response curve of loudspeakers from about 8000 to some 18 000 c/s is desirable. A brief review of the theory of moving-coil loudspeakers and the radiation characteristics at low and medium frequencies is followed by details of possible methods of achieving the required extension. One very good solution is the twin-cone loudspeaker which gives a response curve at once more uniform and with a higher cut-off (18 000 c/s) than the more usual combination of loudspeaker and separate tweeter. The reflecting action of the large cone, and the diffusing effect of the small one give a considerable improvement in the radiation diagrams, and eliminate the inevitable "cross-over dip" which can otherwise be avoided only at the expense of additional expensive measures. In some cases it is better to employ a two-channel amplifier and to feed the two channels into the loudspeakers separately. Steps must also be taken to avoid phase differences at the cross-over point; this can be done at a relatively low cross-over frequency by using a twin-cone model for one of the speakers.

A PHOTO-MULTIPLIER TUBE FOR SCINTILLATION COUNTING

by R. CHAMPEIX *), H. DORMONT *) and E. MORILLEAU *) 621.383.27:539.16.08

In experiments involving radioactive substances, accurate and sensitive methods are necessary for the measurement of the radiations emitted. Such measurements are often made with the aid of the familiar scintillation properties exhibited by certain crystals, by converting the scintillations into electric currents in a photo-multiplier tube. A photo-multiplier tube specially designed for this purpose is described in the present article.

The radiations emitted by radioactive substances can be measured in a number of ways. The Geiger-Müller counter [1]), for example, is widely used: any radioactive radiation striking such a gas-filled tube initiates momentarily heavy discharges, which send current pulses through the tube. The voltage pulses generated by these current pulses in a suitable resistance are strong enough to be readily detected and recorded ("counted") with the aid of simple electronic instruments [2]).

Despite one or two inherent disadvantages, the Geiger-Müller counter is very widely used on account of the simplicity of the tube itself and its associated circuits. The disadvantages will now be outlined. Firstly, the magnitude of the current pulses produced by incident particles is constant, that is, it is independent of the energies of the individual particles; hence the Geiger-Müller counter does not provide quantitative data as to the energy of the particles detected (which is necessary if it is required to know the energy distribution of the particles or quanta released during a particular radioactive process).

Another drawback of the Geiger-Müller tube is its rather long "dead-time". Each discharge occupies a certain period (about 100 micro-seconds), and the tube does not respond to any particles which happen to enter it during this time: hence such particles remain uncounted. Since this effect becomes more serious as the strength of the incident radiation increases, the increase in the number of discharges per second with the radiation intensity is not linear.

As regards the measurement of weak radiation, on the other hand, the principal disadvantage of the Geiger-Müller tube is low quantum-efficiency; at the most, $1/2$ to 1 % of the quanta striking the tube initiate discharges. Hence the remainder are omitted from the count.

In view of these disadvantages other methods of measuring radioactive radiation have been sought. One solution to the problem lies in the application of the scintillation effect. Scintillation, first noted as a feature of α-particles by Crookes in 1903, may be described in the following manner. Certain substances when exposed to radioactive radiation exhibit very brief flashes of light at one point after another. It has been shown that each of these scintillations is produced by a single incident particle, or quantum, and that the intensity of each scintillation is proportional to the energy of the particle producing it. It was recognised that this gave the possibility of measuring the number and intensity of the different scintillations with an instrument having a very short dead-time, a considerable improvement on the Geiger-Müller counter.

During the initial investigations of the scintillation effect, the flashes were, in fact, observed and counted visually with the aid of a magnifying glass; however, it will be seen that this method is subject to severe limitations. The obvious course is to convert the scintillations into electric pulses by means of a photo-electric cell, but the flashes are so faint as to evoke hardly any response in an ordinary photoelectric cell, owing to the high level of noise generated by the cell and its associated amplifier. However, in 1944 Curran and Baker conceived the idea of counting the scintillations by means of a photo-multiplier tube, that is, a special photo-electric cell whose current is amplified by secondary emission, a process involving very much less noise than the amplification of the output signal of an ordinary photo-electric cell by means of a conventional amplifier circuit.

The subsequent widespread and successful adoption of this idea is sufficient evidence of its merit. However, whilst appreciating that this method does all that is claimed for it in eliminating the disadvantages of the Geiger-Müller tube, we should bear in mind that it involves the use of a photo-multiplier tube which is itself very much more complex than the inherently simple Geiger counter.

*) Laboratoires d'Electronique et de Physique appliquées, Paris.

1) See, for example, Philips tech. Rev. **13**, 282-292, 1951/52.

2) Some examples of counting equipment are given in Philips tech. Rev. **14**, 313-326, and 369-376, 1952/53.

In the present article we shall consider a multiplier tube developed by the "Laboratoires d'Electronique et de Physique appliquées" in Paris. However, before examining the considerations governing the actual design of this tube, we shall discuss briefly certain features of the scintillating material, and give a summary of the advantages of the photo-multiplier tube from the point of view of noise-level [3]).

Choice of the scintillating material

The phosphor material, that is, the material which scintillates when struck by the particles to be counted, should satisfy the following requirements:

1) it should readily absorb radioactive radiation;
2) it should be highly efficient in converting "radio-active" energy into "light" energy;
3) it should absorb no more than a small proportion of the light produced by scintillation, that is, it should be transparent;
4) the duration of its scintillations should be as short as possible with a view to a short dead-time.

Several materials offer adequate combinations of these properties. Although in most cases single crystals are employed, certain polycrystalline materials, plastics and liquids are likewise suitable. Some characteristics of certain single crystal materials are given below. The wave lengths quoted refer to the maximum of a continuous spectrum 1000 to 2000 Å wide.

Material	Wavelength of scintillation in Å	Time-constant in sec.
Sodium iodide (activated with thallium) .	4100	30×10^{-8}
Anthracene	4400	3×10^{-8}
Naphthalene	3500	5×10^{-8}

Under favourable conditions, such materials may be up to 20 % efficient in converting the energy of each absorbed particle into light energy. The percentage of incident radiation absorbed depends upon the nature of the radiation and also, of course, upon the dimensions of the particular crystal. Quanta of γ-radiation are absorbed quite readily (about 40 %); X-radiation may even be completely absorbed by crystals of this type.

[3]) See Philips tech. Rev. 10, 263, 1948/49. A description of older photo-multiplier tube designs will be found in Philips tech. Rev. 3, 134, 1938 (tube with 11 stages and magnetic focusing) and 5, 253-257, 1940. (tube with 4 stages and electrostatic focusing).

Noise characteristics of photo-electric cells and photo-multipliers

A considerable difference exists between the noise characteristics of a photo-electric cell and its associated amplifier as opposed to those of a photo-multiplier tube. This will now be elaborated.

In the case of the combination photo-electric cell + amplifier, the photo-current of the cell, passes through a resistor, and the voltage thus generated is amplified by one or more electron tubes. Apart from the signal voltage, a noise-voltage exists between the ends of the resistor, which arises from the thermal agitation of the electrons in the resistor. This voltage is equivalent to an e.m.f. in series with the resistor. The noise e.m.f. is proportional to the square root of the resistance.

Even in the absence of any photo-current, then, a certain noise-voltage occurs across the resistor. This augments the noise-voltage arising from the noise-component of the photo-current itself. The signal-to-noise ratio of the voltage applied to the amplifier is therefore lower than that of the original photo-current.

However, it is possible to lessen this increase in the relative noise-level by increasing the load-resistance of the photo-electric cell, since, given a relatively high resistance, the voltage produced by the photo-current will increase linearly with the resistance, whereas the noise-voltage of the resistor itself will increase only as the square root of the resistance.

Accordingly, the load-resistance is always made as high as possible; its value if limited, however, by the bandwidth of the particular signal to be amplified: the maximum bandwith of the amplified depends on the stray capacitance in parallel with the load-resistance, and becomes smaller as the load-resistance is increased. On purely theoretical ground, however, it is true to say that, for a sufficiently narrow band width, the signal-to-noise ratio of a photo-electric cell-amplifier combination may be as high as that of the original photo-current.

The noise characteristics of the photo-multiplier, on the other hand, do not depend on the output resistance, since (owing to the initial amplification by secondary emission) the current at the output is so strong that the noise-voltage generated in the resistance itself is negligible by comparison. There is instead, another source of noise as a result of the secondary emission, viz. that which arises because the number of secondary electrons released by each primary electron striking the dynodes fluctuates about a certain mean.

Now, the spectral intensity distribution of this "secondary emission noise" is identical with that of

the original photo-current noise; hence the increase in the relative noise level arising from secondary emission does not depend on the bandwidth of the signal and may therefore be considered constant.

The noise characteristics of the photo-electric cell as compared with those of the photo-multiplier may well suggest that the cell and amplifier combination is the best, since this combination affords at least some measure of control over the noise characteristic through limitation of the bandwidth. In fact, however, precisely the opposite is the case owing to the fact that, with the bandwidths employed in practice, the proportion of the original noise-level attributable to the load-resistance of the photo-electric cell far outweighs that arising from secondary emission in a photo-multiplier tube. Hence the latter enables us to detect scintillations about 100 times fainter than can be discerned with the aid of a photo-electric cell and associated amplifier.

The nature of the "signal" in scintillation counting — viz. current surges in the form of pulses alternating with non-conductive intervals — gives the photo-multiplier a further advantage over the rival combination. Given the possibility of procuring zero "dark current" (output current in the absence of light), then regardless of the amount of secondary emission noise produced, any scintillation, however faint, will produce an output current pulse strong enough, to be discerned without undue difficulty, because of the absence of background noise (the magnitude of such a pulse, will of course depend on the secondary emission noise).

In a photo-electric cell and amplifier combination background noise is always present, owing to the noise-voltage generated by the load-resistance itself, and this prevents the detection of very small signal voltages.

Accordingly, it is seen that the sensitivity to scintillations of the photo-electric cell is inherently limited by the resistance noise, whereas that of the photo-multiplier is governed by the magnitude of the "dark current". Hence it is desirable to reduce the latter as far as possible. Means by which the dark current may be reduced are in fact known and will be described presently.

Properties desirable in a photo-multiplier tube

A photo-multiplier tube employed to measure scintillations should possess the following properties:

1) Adequate sensitivity, that is, enough to ensure that each scintillation will produce in the output circuit a voltage strong enough to be amplified in the following stages without any perceptible effect on the signal-to-noise ratio.
2) Low inherent noise-level.
3) Low resolving time, to ensure that particles striking the tube in quick succession will be counted.
4) Constant ratio between incident light and the resultant photo-current, so that the output current is a true measure of the initial quantity of light, that is, of the energy of the original radioactive particle.

The factors governing these different properties will now be considered individually.

Sensitivity

The possible sensitivity of a photo-multiplier depends of course on its construction. Two different types of electrode system are commonly employed in photo-multiplier tubes.

Firstly, there is the system in which each secondary-electron emitter ("dynode") comprises several plates arranged like the slats of a venetian blind (*fig. 1*). In this system the electrons emitted by the photo-cathode travel to the first dynode, which is maintained at a certain positive potential relative to the cathode. Striking this dynode, they release secondary electrons, which in turn are attracted to the second dynode (this being at a still higher potential), thus releasing more secondary electrons, and so the process continues.

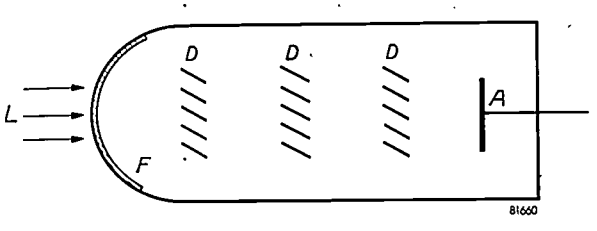


Fig. 1. Diagram showing the basic construction of a photo-multiplier tube in which the dynodes *D* are arranged in the form of a "Venetian blind".
The incident light falls on the photo-sensitive layer *F*, and all the electrons emitted are collected by the anode *A*.

However, in the multiplier tube considered here, the other system has been adopted, for reasons which will be seen later. Here, the photo-electrons are focused on the first (one-piece) dynode (*fig. 2*)

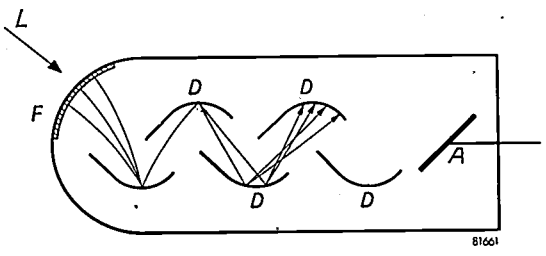


Fig. 2. Diagram of a photo-multiplier tube in which the electrons are focused electrically. The meaning of the letters is the same as in fig. 1.

by a suitable electric field; the secondary electrons thus released are accelerated and deflected by the electric field between the first and second dynodes, proceed to the second dynode, and so on. As in the system previously considered, each dynode is maintained at a higher potential relative to the cathode than the one preceding it, with the anode, the final destination of all the electrons, at the highest potential of all.

The number of electrodes employed to emit secondary electrons depends on the desired current-amplification, and on the secondary emission factor (δ) of the particular dynode-material. The secondary emission factor may be defined as the number of secondary electrons emitted per incident primary electron, as averaged over a large number of primaries. A secondary emission factor of up to 3 or 4 can be obtained with caesium-coated dynodes; ten dynodes in one tube give the required current-amplification.

The number of dynodes required may be calculated as follows. Given a particular type of radioactive radiation, say the α_3 radiation emitted by polonium (the energy of each particle being about 5.3 MeV), falling upon a crystal of anthracene, and assuming the "luminous efficiency" of such a crystal to be 20%, we find that the amount of energy converted into light per incident particle is $0.20 \times 5.3 \times 10^6 \approx 10^6$ eV. Suitable reflecting layers applied to the walls are assumed to focus most of this energy on the photocathode of the multiplier. The wavelength of the light thus generated in the anthracene is 4400 Å, and at this wavelength each light-quantum corresponds to an energy-level of about 2.8 eV; hence the light energy of 10^6 eV comprises about $10^6/2.8 \approx 3.6 \times 10^5$ light-quanta

The efficiency of a photo-cathode operating at this wavelength is about 0.1 electron per light-quantum; therefore one α-particle will release about 3×10^4 primary electrons in the photo-multiplier tube. In a 10-dynode multiplier tube with a secondary emission factor of 3.6, the current amplification is $(3.6)^{10} \approx 4 \times 10^5$, so that the anode current-pulse generated by one incident α-particle comprises 12×10^9 electrons, which corresponds to a charge of $12 \times 10^9 \times 1.6 \times 10^{-19} \approx 2 \times 10^{-9}$ coulombs. If the total capacitance of the anode circuit be 20 pF, such a charge corresponds to a voltage $Q/C = 2 \times 10^{-9}/20 \times 10^{-12} = 100$ Volts. In practice, however, voltage variations as large as this are not permissible, since they would cause defocusing of the electron beam at the last of the dynodes (aggravated by the space charges arising from the heavy currents involved); this might well impair the linear relationship between the primary photo-current and the output voltage.

Inherent noise level

It has already been remarked that a good quality photo-multiplier tube is one having zero or a very small dark-current. Let us now consider the possible causes of electron-emission in the absence of light.

Owing to the current amplification taking place after the photo-cathode and the first dynode, any "parasitic" electrons emitted by these two electrodes give rise to much larger currents in the output circuit than would be produced by similar parasitic emission from the other electrodes in the tube. It is therefore necessary to pay particular attention to the primary electrodes.

Spontaneous electron-emission in the absence of light may be either thermionic, or "cold". Although usually very slight at room temperature, thermionic emission may cause noise owing to the high current-amplification occurring in the photo-multiplier. For the photo-sensitive layer therefore a material is chosen with the highest practicable threshold-potential, so as to minimize the thermionic emission at a given temperature. However, since the same potential must be surmounted by the electrons constituting the photo-current, light-quanta having a fairly high energy-level are required to extract photo-emission from such a material. Hence the threshold-wavelength for photo-electric response is so short that red light will not produce photo-emission. Fortunately, most crystals employed for scintillation counting produce light of short wavelengths: hence this lack of red-sensitivity is no real disadvantage. The photo-sensitive coating employed in the present photo-multiplier tube is a combination of caesium and antimony, having a high threshold potential (1.7 eV); its long-wave threshold for photo-emission is 7000 Å.

Similar considerations apply to the choice of the dynode material, in this case copper coated with caesium oxide.

Moreover, in view of the invariable, although slight, thermionic emission, the area of the photo-sensitive layer and that of the dynodes are made as small as possible having regard to the size of the scintillating crystals used.

Another possible source of parasitic current is so-called "cold" emission; this arises from the direct action of an external electric field upon the surface of the material. Since a strong field, such as may build up around sharp edges or irregularities on the electrodes, is required to produce such emission, smooth surfaces and electrodes with suitably rounded shapes are effective as a means of avoiding it.

An increase in the noise-level may also proceed from causes other than those already described. One of them is leakage current between the different electrodes, caused by impurities on the surface of the glass or the mica electrode-supports. Since the leakage path on which the size of such currents depends tends to vary and so cause crackling in the tube, it is necessary to provide the best possible

insulation for all the electrodes, and particularly for the anode, since the leakage current of the anode passes direct into the external load-impedance.

Again, any electrons drifting at random in the tube may touch the glass and so excite fluorescence. If the light so produced happens to fall on the photo-cathode it may cause undesired photo-emission, thus increasing the fluctuation-level. The electrodes should therefore be so designed as to minimize the possibility that electrons will drift through the tube outside the electrode system proper.

Any residual gases present in the tube are likewise a possible source of noise, since positive ions formed by collision between electrons and the gas atoms travel to the photo-cathode, causing secondary emission and so affecting the primary current. To avoid such interference it is necessary to exhaust the tubes very thoroughly.

In the case of the photomultiplier under consideration, careful attention to these different sources of noise has resulted in a "dark current" of less than 5×10^{-8} amperes, combined with a current-multiplication of 500 000×.

Response time of the photo-multiplier

We have already seen that one of the disadvantages of the Geiger-Müller tube is its long dead-time (about 100 μsec). In the case of the photo-multiplier tube, the duration of a single output-pulse is very much shorter. Assuming for the moment that the flash in the photo-cathode is infinitely short, we find that the factors governing the duration of the anode pulse are as follows:

1) The velocity distribution of the secondary electrons emitted. — Between each pair of consecutive dynodes there is a potential difference of about 100 volts. Given a distance of 5 mm between electrodes, those electrons whose initial velocity at their originating dynode is zero will have a transit time of about 2×10^{-9} sec. However, owing to the fact that in practice the initial velocity varies between different electrons, a certain amount of spread, say about 3.5×10^{-10} sec, occurs in the transit time.
2) The emission-lag of the secondary emission. — As far as can be ascertained, this time-constant is very short, viz. of the order of 10^{-22} sec; hence the effect of any variation in it will be small.
3) The difference in the paths of different electrons. — Owing to the often considerable difference in the paths of individual electrons, the spread in the transit time may be considerable; indeed it may become as large as the value of the transit time quoted under 1), that is, about 2×10^{-9} sec. From this point of view, the "venetian blind" electrode arrangement (fig. 1) is unfavourable, since it enables some of the electrons to by-pass a particular electrode and so fosters considerable differences in transit time. For this reason it is not employed in the present photo-multiplier.

The above assumption that the original flash in the scintillating crystal is infinitely short is not, of course, the case in practice; in fact, the duration of the flash in the "fastest" crystals is about 10^{-8} sec. Hence the widening of the pulse in the photo-multiplier as a result of the effects referred to (1, 2, 3, above) does not cause undue disturbance. However, it should be remembered that the time-constant of the output-circuit of the tube also has some effect.

Fig. 3 shows a load resistor (R) in parallel with the anode-capacitance (C). Now, on the one hand the time constant of this RC-combination should be long compared with the duration of a pulse, to get as large a pulse as possible. On the other hand,

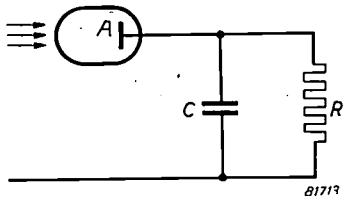


Fig. 3. Anode-circuit of a photo-multiplier tube (with anode A), comprising a load-resistor R and the stray capacitance C.

however, the capacitance itself should discharge as quickly as possible after each pulse, if rapidly succeeding pulses are to be resolved; from this point of view, a short time-constant is required. The two requirements are thus opposed. However, the time-constant of the output circuit need not be shorter than the response time of the crystal. For example, 25,000 ohms for the anode-resistance and about 20 pF for the overall self-capacitance, which gives $RC = 0.2$ μsec., is quite satisfactory in practice. This is very much shorter than the dead-time of about 100 μsec for Geiger-Müller tubes.

Stability of the amplification

To enable qualitative measurements to be carried out, "incident" light-flashes of the same intensity must produce pulses of the same size in the anode-circuit of the photo-multiplier tube. Spread in the size of such pulses may arise from:

1) variation in sensitivity between different points on the photo-cathode ("spread in position");

2) variation in the amplification owing to variation in the secondary emission ("spread in time"). This effect is stronger in the case of the "Venetian blind" design shown in fig. 1 than in that of the electrode-arrangement seen in fig. 2, again because some of the electrons may by-pass one or more electrodes.

Construction of the photo-multiplier

Fig. 4 shows the construction of the photo-multiplier tube. It will be seen from this diagram that the area of the photo-cathode is small, in fact, as

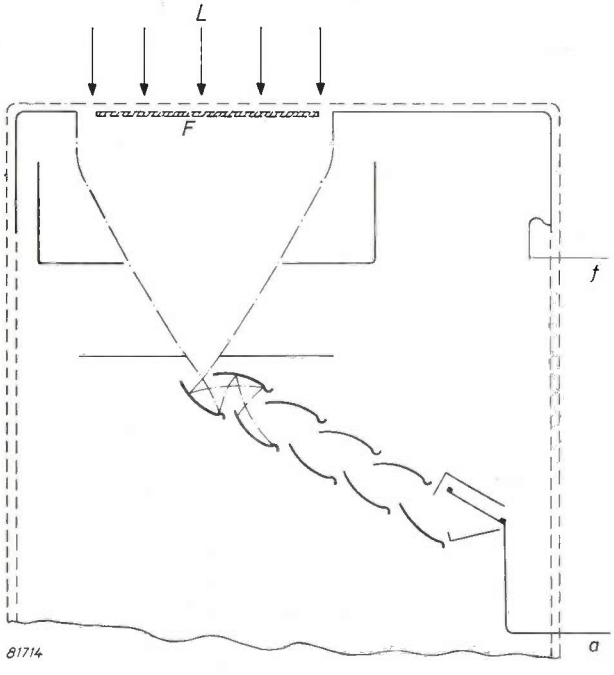


Fig. 4. Diagram showing the construction of the photo-multiplier tube described in this article. The incident light *L* falls on the photo-cathode *F*. The letters *f* and *a* indicate the lead-in wires of the photo-cathode and the anode respectively.

small as it can be for the crystals ordinarily employed; in the present tube it is 5 sq.cm. A Wehnelt cylinder is employed to focus the primary electrons on the first dynode, whose potential relative to the photo-cathode is +100 Volts. It is also seen from fig. 4 that the ends of the plates forming the dynode are curved; this is to prevent any local formation of strong fields which might cause "cold" emission.

The photo-cathode itself is a compound of antimony and caesium, i.e. $SbCs_3$, attaining maximum sensitivity at a wavelength of 4800 Å. That part of the glass envelope which carries the cathode is optically flat and polished so as to permit of direct contact between the glass and the scintillating crystals; hence the transmission of light from crystal to photo-cathode is very efficient.

The electron-optical system focusing the electrons on the first dynode is designed carefully to ensure that 90 % of the photo-electrons will reach this dynode; thus the number of electrons drifting at random in the tube is so limited that the fluorescent effects produced by them on the glass are negligible.

The base material of the dynodes is pure copper. Caesium vaporised in the tube during the "forming" of the photo-cathode settles on the copper dynodes, thus reducing part of the copper oxide on the surface of these electrodes. The surface layer so formed, comprising copper, caesium and the oxides of these two elements, has a secondary emission factor $\delta = 4$ when the energy of the incident electrons is 100 V.

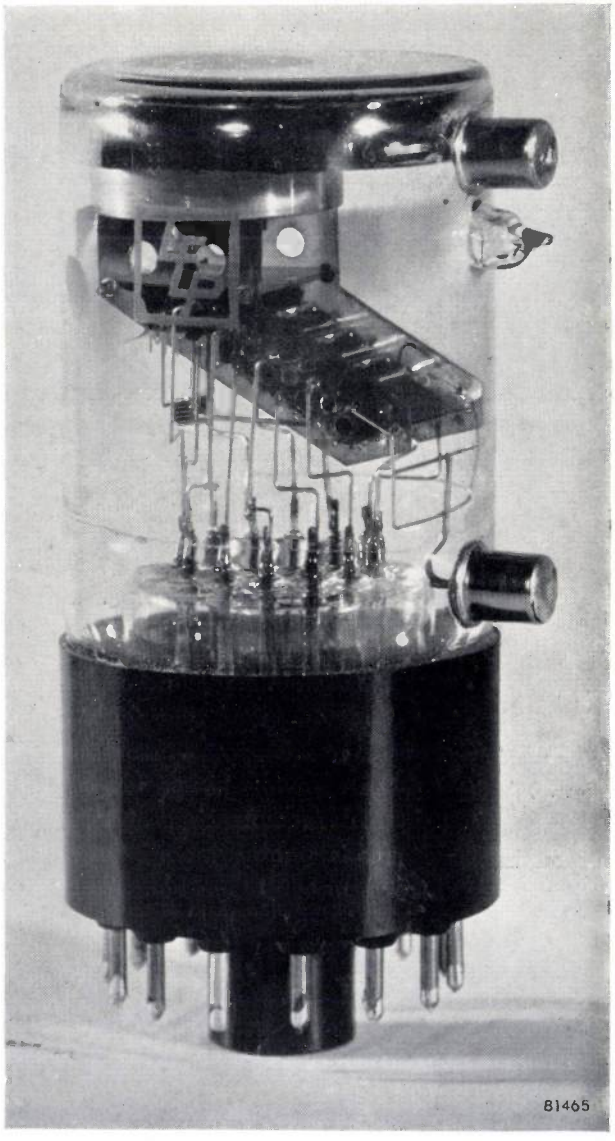


Fig. 5. The photo-multiplier tube. Note the photo-cathode against the flat top of the tube, and beneath it the Wehnelt cylinder, which focuses the electrons on the first dynode. The tube is about 12 cm high.

Through careful degassing of the electrodes and thorough exhaustion of the tubes, the number of gas-ions formed is so reduced as to eliminate all interference from this source; it has even been possible to dispense with the zirconium getter originally employed.

Fig. 5 is a photograph of the photo-multiplier tube.

Suitability for use in spectrometry

As already explained, a scintillation counter should be so designed that it can also be employed to measure the energy distribution of ionizing radiations. For the purpose of spectrometry of ionizing radiations, the output voltage of the multiplier must be a reliable measure of the energy of the incident radiation. This aspect of the quality of a scintillating crystal and photo-multiplier combination may be assessed as follows.

A radioactive substance is used, the radiation from which is known to consist only of α-particles travelling with a sharply defined velocity. A suitable substance is polonium, which emits α-radiation of

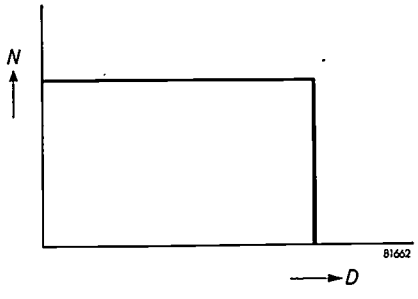


Fig. 6. Diagram showing the number of pulses per second N plotted against the threshold voltage D of the counter, in the case of monoenergetic radiation and ideal measuring equipment.

energy 5.3×10^6 eV. This radiation is allowed to fall upon a scintillating crystal (anthracene for example) fixed to the photo-multiplier. The output-pulses of the photo-multiplier are fed to a counter responding only to pulses whose size exceeds a given (adjustable) threshold value.

If the number of pulses counted per second at different threshold voltages is plotted against the threshold voltage, then with purely monoenergetic radiation and an ideal measuring instrument, the diagram so obtained should be rectangular (*fig. 6*), since all the pulses should then be of the same size and the number of pulses recorded should drop abruptly to zero as soon as the threshold voltage of the counter is made greater than a certain value.

Fig. 7 shows an experimental curve for a photo-multiplier tube not specially designed to measure radioactive radiation, and *fig. 8* shows the results of similar measurements for the present tube. The latter still exhibits a perceptible, through relatively

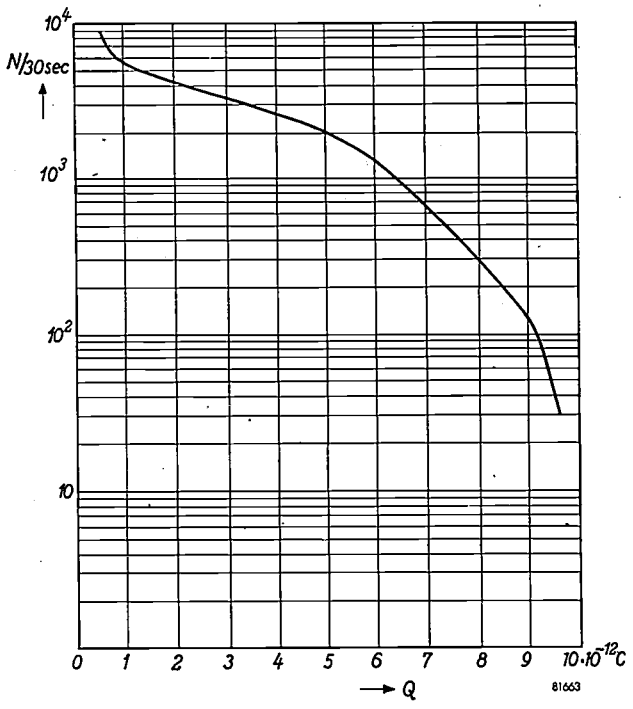


Fig. 7. Number of pulses N per 30 seconds versus the output charge Q per pulse in the case of a photo-multiplier not specially designed for radio-spectrometry.

slight rounding of the rectangular shape: this may be attributed to any of the following causes:

1) non-monoenergetic radiation;
2) scintillation of the crystal not always exactly proportional to the energy of the incident particle;

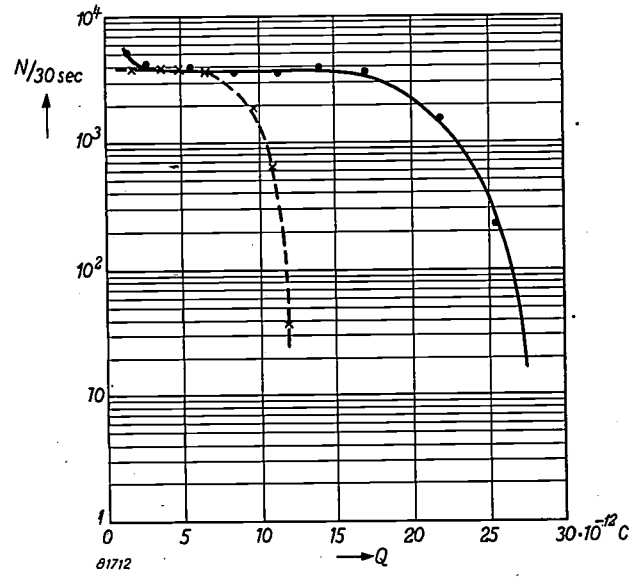


Fig. 8. Curves corresponding to those in fig. 7, for the case of the photo-multiplier described in this article, for two different velocities of the incident α-particles.

3) lack of homogeneity in the photo-sensitive layer;
4) effect of fluctuations in the photo-multiplier;
5) "threshold" selectivity of the counter not infinitely sharp.

Other uses of the photo-multiplier tube

Apart from radio-spectrometry, a sensitive scintillation counter can be employed also for various other purposes associated with experiments involving radioactive substances, viz. the detection of indicators in tracer work (particularly in biology and medicine), the examination of welded seams with the aid of radioactive radiation, geological investigations based on observations of such radiation, and so on.

The high sensitivity of the counter is important also in that it enables such a tube to be employed as a means of measuring very small quantities of light, as in astronomy, or in connection with a spectrograph having a very high resolving power. In fact, the scintillation counter and associated photo-multiplier can be employed wherever a high degree of light-sensitivity and quantitative measurement of intensity are required.

Summary. A scintillating crystal in conjunction with a photo-multiplier tube for investigations of radioactive radiations gives a measuring system of high sensitivity (low noise-level), short dead-time (10^{-7} sec.), and of linear response (output-current pulse $\propto$ to energy of incident particle). These properties of the system make it suitable for the spectrometry of ionizing radiations. Some of the properties necessary in the scintillating material are described, and the superiority of the photo-multiplier tube to the combination photo-electric cell + amplifier as regards noise-characteristics is explained. A photo-multiplier tube developed by the Laboratoires d'Electronique et de Physique appliquées" in Paris is described. This tube contains ten stages of secondary emission, giving an overall current-amplification of 500 000 times. To preserve a low inherent noise-level, it is necessary to reduce all parasitic emission (e.g. thermionic or "cold" cmission) in the tube as far as possible. It is shown that the resolving time of the scintillation counter-photo-multiplier combination is governed mainly by the time-constants of the scintillating crystal and the output circnit. The construction of the tube is described: it is shown that the tube is suitable for spectrometry.

ENTROPY IN SCIENCE AND TECHNOLOGY

I. THE CONCEPT OF ENTROPY

by J. D. FAST.

536.75

The author, who is not unknown to our readers by virtue of his many articles on metallurgy, has made a thorough study of the concept of entropy and written a widely-read book on this subject. It is his intention, in a series of articles in this Review, to consider the significance of the concept of entropy in various fields of science and technology. Owing to the abstract way in which entropy is dealt with in classical thermodynamics, it is a less familiar concept than that of energy, though no less important. The author devotes particular attention to the statistical background of entropy, in an attempt to make it as readily understandable as the concept of energy. It is surprising how many widely divergent problems may be resolved with the aid of the concept of entropy. A general survey like this may contribute in some measure towards a general synthesis, all the more urgently needed as specialization in the exact sciences becomes more acute.

The present article, the first of the series, is devoted to the essential meaning of entropy. The subsequent articles will be mainly concerned with examples of the application of the entropy concept.

Introduction

All phenomena in nature are subject to the laws of thermodynamics. These well-established laws enable us not only to calculate the maximum possible efficiency of engines and to predict the direction and the maximum yield of chemical reactions, but they are also of fundamental importance in almost every field of science and technology.

The first and second laws may be formulated in many different ways. At first sight the various formulations seem to bear little or no relation to each other, but essentially they are equivalent. When applied to an isolated system, i.e. a system without interaction with the outside world, they may, for example be worded as:

First law: The total energy of an isolated system is constant.

Second law: The entropy of an isolated system tends towards a maximum.

The first law of thermodynamics

The first law of thermodynamics, which is sometimes called the law of conservation of energy, finds its origin in the empirical knowledge that heat and mechanical work are both forms of energy and that the one may be converted into the other.

If a system is not isolated (e.g. a quantity of gas in a cylinder under a movable piston), then an amount of heat dQ may be added to it, or an amount of work dW may be done on it. According to the first law the whole of this added energy must appear in the system as an increase in its internal energy U, i.e.,

$$dU = dQ + dW. \quad . \quad . \quad . \quad . \quad . \quad (I,1)$$

From the point of view of classical thermodynamics — i.e. independent of the state of agregation of matter or the precise physical form of the energy — the concept of internal energy gains a significance only by virtue of this mathematical definition. If the atomic state of agregation of matter is considered, the internal energy of a system is the sum of the kinetic and potential energies of all the elementary particles of which the system consists. The internal energy depends solely on the *thermodynamic state* of the system, i.e. on its pressure, temperature, volume, chemical composition, structure, etc. The history of the system does not influence its value.

For this reason U is called a *thermodynamic function*. W and Q are not thermodynamic functions, since according to equation (I, 1) the same change in internal energy dU can be brought about either by supplying only heat, or by only doing work on the system. It is therefore possible to speak of the internal energy of a system, but not of the quantity of work or the quantity of heat of that system. In other words: dW and dQ are only infinitesimal quantities of work and heat, and not differentials of thermodynamic functions.

The second law of thermodynamics

Although W is not a thermodynamic function and dW is not a differential, the latter can generally be expressed as the product of an *intensive* property of the system and the differential of an *extensive* property of the system. The meaning of these terms is given by the fact that a system in equilibrium can always be divided into two equal parts such that those thermodynamic properties which are

extensive (e.g. volume) are halved, while those which are intensive (e.g. pressure) remain unchanged. The work done on a gas by compressing it can, for example, be expressed as:

$$dW = -p\,dV,$$

in which p, the pressure of the gas is the intensive property of the system, and V, its volume the extensive property. In so far as p refers to the internal pressure, this formula applies only to a reversible change of volume, i.e. such a change that the external pressure always differs only infinitesimally from the internal pressure. Analogously, dQ may be expressed as:

$$dQ = T dS,$$

in which T, the temperature of the system, is the intensive property and S, the entropy, the extensive property. This formula, too, applies only if the heat is supplied in a reversible manner, i.e. supplied from a source whose temperature is only infinitesimally higher than that of the system. We may thus write:

$$dS = \frac{dQ_{rev}}{T} \quad . \quad . \quad . \quad . \quad . \quad . \quad (I,2)$$

This formula, apart from providing the definition of the thermodynamic function S (and, strictly speaking, also of the absolute temperature T), also represents the mathematical formulation of the second law of thermodynamics.

For readers without a previous knowledge of thermodynamics the foregoing will still leave the concept of entropy completely obscure. One also feels the lack of any connection between the given formulation and the above-mentioned tendency of the entropy towards a maximum. To bring some light into this darkness it may be useful to leave the path of pure thermodynamics and to consider the atomic aspect of the matter. Only after explaining the atomic aspect of the concept of entropy, shall we return to the thermodynamical definition (I, 2), and demonstrate how this should be modified in order to express the tendency of entropy towards a maximum.

Irreversible processes

The second law of thermodynamics finds its origin in the experience that all spontaneously occurring processes take place in one direction only and are, therefore, irreversible.

If the previously considered isolated system consisted, say, of a vessel containing neon and helium under such conditions that they may be considered perfect gases, then the first law would permit any imaginable distribution of the gas molecules in the available space: pressure and temperature differences may exist between different parts of the mixture without affecting in any way the internal energy of the system. Experience tells us, however, that irrespective of the initial state, the ultimate state of the system left to itself (the "equilibrium state") is always one in which the gases have mixed homogeneously and in which pressure and temperature are uniform. After reaching this ultimate state, the system will never spontaneously return to one of its previous states.

How can we explain this tendency towards homogeneous mixing? May we say that the helium and the neon atoms have a certain "preference" for homogeneous distribution and, if so, on what is this preference founded?

In order to deal with these questions, let us imagine that a small number of white and red billiard balls (e.g. 50 + 50) are substituted for the helium and neon atoms. As the initial state we select a given, regular distribution of the balls in the vessel (*fig.1*), and the thermal motion of the atoms is simulated by

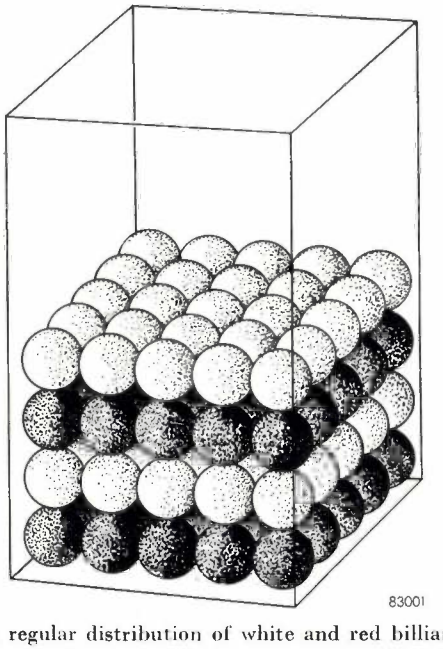


Fig. 1. A regular distribution of white and red billiard balls.

thoroughly shaking the vessel for some time. We know from experience that after shaking, the orderly initial state will never be found again; a random distribution of the white and red balls will always be found. Yet we must take it as axiomatic that each separate random distribution is just as probable or improbable as the initial distribution. In actual fact, however, and here we hit the core of the problem, there are so many more possible random distributions than regular ones that virtually only random distributions will be found

after shaking. This point may be illustrated with the aid of a simple calculation.

A statistical calculation

If all 100 balls could be distinguished from one another (e.g. because they were numbered), there would be $100 \times 99 \times 98 \times \ldots \times 2 \times 1 = 100!$ different ways of arranging them in the 100 available spaces in the vessel.

Since, however, the 50 red balls are, in fact, indistinguishable (not numbered), any interchanging of 2 red balls leaves the distribution unaltered, so that the number of possible arrangements that can be distinguished by eye (m) is far smaller. This number, nevertheless, is still enormous. Taking into account the fact that the 50 white balls are also mutually indistinguishable, it is given by [1])

$$m = \frac{100!}{50!\,50!} = 0.08 \times 2^{100} = 1.01 \times 10^{29}.$$

The chosen number of balls is too small for applying the roughest approximation of Stirling's formula [2])

$$\ln N! \approx N \ln N - N \quad . \quad . \quad . \quad . \quad (I,3)$$

since then the result for m would be

$$m = 2^{100} = 1.26 \times 10^{30},$$

a value which is too large by more than a factor of 12. It will be readily appreciated that the number of 2^{100} includes not only all possible distributions of 50 white and 50 red balls, but also all distributions of a total of 100 balls irrespective of the ratio of red to white, i.e. all distributions of 49 white and 51 red balls, of 48 white and 52 red balls, etc. Without the condition of the 50/50 ratio, we have the following situation. The first place to be occupied by a ball provides a choice of two possibilities (white or red); for occupying two places, each of the two colours may be combined with each of the two colours, so that here one has 2^2 possibilities (w-w, w-r, r-w and r-r), and so on.

According to the result $m = 10^{29}$, an average of 10^{29} shakings are necessary to obtain one given distribution of the balls. If each shaking action is made to last one minute, this means that on the average one would have to shake for 10^{23} years in order to obtain one given distribution. This constitutes a sufficient explanation of the empirical knowledge that the chance of obtaining by shaking a perfectly regular distribution, e.g. an arrangement of the white and the red balls in separate layers or one in which each white ball is exclusively surrounded by red ones and vice versa, is practically nil in view of the comparatively small number of regular arrangements compared to the number g of the irregular distributions ($g \approx m$).

The foregoing considerations were concerned with the small number of 100 balls. Returning to our gas mixture and assuming that it consists of 0.5 gram-atom of Ne and 0.5 gram-atom of He, we arrive at a total number $N_0 = 0.6 \times 10^{24}$ atoms. A calculation analogous to the one carried out for 100 "atoms", now gives for m the value

$$m = \frac{N_0!}{\{(1/2)\,N_0)!\}^2} = 2^{N_0 - 40} \approx 10^{(2 \times 10^{23}) - 12}.$$

The number 40 in the exponent of 2 is negligible with respect to $N_0 = 6 \times 10^{23}$; moreover, the value of Avogadro's number is known to so few decimals that from a physical point of view there is no point in distinguishing between N_0 and $N_0 - 40$. We may thus write:

$$m \approx 2^{N_0}$$

and this is, according to the foregoing, nothing but the total number of distributions of N_0 atoms of two types. In other words: at large values of N and with the ratio 1 : 1 there occurs so sharp a maximum in the curve of the number of distributions as a function of the mixing ratio that there is practically no difference whether we take into account the number belonging to this maximum or the total number of distributions. It is of some importance that the approximation (I, 3) also gives the result $m = 2^{N_0}$. Hence this formula is a perfectly satisfactory approximation when applied to the numbers of atoms normally dealt with in practice.

Such a staggering number as 2^{N_0} is, of course, quite beyond human comprehension; the number of orderly distributions of the atoms is negligibly small when compared to this total number m.

Macro and micro-states

As already stated, we are bound to assume that with our shaking experiment all $m = 10^{29}$ distributions of the balls (all "micro-states") possess an equal probability w,

$$w = \frac{1}{m} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (I,4)$$

The various micro states may be assembled into groups ("macro-states") each of which is character-

[1]) Exact values of $N!$ for integers up to $N = 100$ are given in Barlow's Tables, E. and F. N. Spon, London. Fairly exact values are given by Stirling's formula in the approximation $N! = \frac{N^N}{e^N} \sqrt{2\pi N}$.

[2]) Throughout this article we shall use ln in place of the more cumbersome $\log_e$, to denote natural logarithms.

ized by a certain extent of disorder. A regular arrangement as shown in fig. 1, is only possible in one way, although one might regard it as being possible in two ways (horizontal layers w-r-w-r or r-w-r-w). The same applies to a three-dimensional checkerboard pattern in which each red ball has only white and each white ball has only red balls as its nearest neighbours. An "imperfection" may be introduced into any orderly distribution by interchanging a red and a white ball. Since the position of each of the 50 white balls can be interchanged with that of each of the 50 red balls, this particular macro-state with one imperfection comprises 2500, micro-states. If the number of imperfections is increased to two or more, then we obtain macrostates comprising considerably more micro-states. The probability w of each macro-state is determined by the number of micro-states or *a priori* equally-probable arrangements g inherent to this state:

$$w = \frac{g}{m}. \quad \ldots\ldots\ldots \quad (I,5)$$

The significance of formula (I, 5) may be illustrated by a very simple example. When throwing dice, the probability of the macro-state"even ", comprising the three micro-states 2, 4 and 6, for a single die, is given by

$$w\,(\text{even}) = \frac{3}{6} = \frac{1}{2},$$

obviously with the provision that all 6 faces of the die have an equal probability of coming on top, i.e. that the die is properly made.

Returning to the atomic case, the various atom configurations or micro-states may be combined in groups, again designated macro-states, the probability of each being determined by formula (I, 5). The choice of the groups is dependent upon the properties of the system under consideration.

If the foregoing is applied to our mixture of ideal gases, it will be clear that at a given moment it is in a given micro-state. Due to the motion of the gas molecules this micro-state is continuously changing. It is justifiable to assume that in the course of time the system passes through all spatial distributions (micro-states) that are possible within the scope of the available volume. There is, however, one particular MACRO-state [3]), in which the gases, for as far as can be ascertained by macroscopic measuring equipment, are homogeneously mixed and of uniform density. This MACRO-state comprises such an enormously greater number of micro-states than all other macrostates put together that after any interval, even if short, it is always found to be present to the exclusion of all the other macro-states. This is called the state of equilibrium, because the system always returns to it of its own accord, irrespective of its initial distribution. This state of equilibrium is also the state of maximum entropy.

Fluctuation phenomena

The answer to the questions put under the heading *Irreversible processes*, concerning the tendency towards forming a homogeneous mixture, should apparently be that the reason for this spontaneously occurring state is just the fact that there exists *no* preference for any particular *micro*-state. In the state of equilibrium the system passes continuously from one micro-state into another but as a rule they are macroscopically indistinguishable. Only under very special conditions can fluctuations around the state of maximum entropy be observable.

The blue colour of the clear sky, for instance, reveals the occurrence of local fluctuations in the density of the air, whilst also the well-known Brownian movement in a colloidal suspension is due to the irregular thermal agitation of the molecules of the medium.

A similar fluctuation phenomenon occurs in a conductor due to the thermal agitation of the electrons. Thus extremely small alternating voltages arise spontaneously between the ends of a resistor. The arithmetic mean of this voltage averaged over a considerable period of time is of course zero, but this is not the case with its r.m.s. value. This phenomenon is termed thermal noise, because after sufficient amplification these alternating voltages can be heard as noise through a loudspeaker. These spontaneous voltage fluctuations may be resolved into components with various frequencies. In 1928 Nyquist demonstrated that, with the exception of the very high frequencies, all frequencies are uniformly represented in the fluctuation spectrum. He further showed that the effect of the fluctuations in an electrical network can be computed by assuming in series with each resistor an imaginary electromotive force E such that $\overline{E^2} = 4kTr\Delta\nu$. In this, k is Boltzmann's constant, T the absolute temperature, r the resistance and $\Delta\nu$ the frequency range (bandwidth) occurring under the given conditions. For a network consisting of a resistance r and a parallel capacitance C, this relat-

[3]) As implied above, a macro-state is taken, in this article, to mean any group of micro-states. Sometimes, however, it is used to describe a thermodynamical state which is characterized by a small number of macroscopic quantities, such as temperature, pressure, volume etc. Such MACRO-states, which as a rule comprise many macro-states, will henceforth be designated by MACRO in capitals. The MACRO-states can be physically distinguished; the macro-states generally cannot.

ion gives a mean square voltage $\overline{V^2} = kT/C$. This is in accordance with the theorem of equipartition of energy: $^1/_2 C\overline{V^2} = {}^1/_2\, kT$.

On the basis of the above-mentioned equivalent circuit it can readily be demonstrated that the maximum power of the noise arising across a resistor is $kT\Delta\nu$. This applies not only to a normal resistor but also to an aerial or a cable in which no end reflections occur. A noise voltage is, therefore superimposed upon every signal voltage. The fluctuating character of this noise renders it impossible to observe the finer details of the signal voltage; it puts a fundamental restriction on the amount of "information" that may be transmitted by electric signals in given circumstances. We shall return to this in the last article of this series.

Quantum states

In the foregoing we have mainly concerned ourselves with the number of micro-states in relation to the mixing of two kinds of particles. Of even greater importance are the micro-states corresponding to the thermal energies of the particles.

Consider a system of identical atoms, in the form of a crystal, or say, as a gas confined to a certain volume. In the crystal each particle is allotted a volume of the order of 10^{-23} cm^3 in which it can execute its thermal oscillations; in the gas, however, each particle can move throughout the entire gas volume of, e.g., 10^2 or 10^3 cm^3.

Modern physics teaches us that a particle confined within a restricted space, can only exist in certain, discrete quantum states. Corresponding to each quantum state are a specific energy level of the particle, and a specific wave function, the latter being related to the probability of finding the relevant particle in the different regions of the available space. One of the fundamental problems of statistical thermodynamics is that of determining the distribution of a system of N identical particles among the various quantum states of the system at a given value U of the total energy. The determination is based on the hypothesis that all micro-states, by analogy with the example of the billiard balls, have an equal *a priori* probability. This hypothesis has been confirmed by the successes of statistical thermodynamics.

In order to demonstrate the various distribution possibilities among the available quantum states we shall consider a greatly simplified model of a solid, known as an Einstein solid, in which the atoms execute their thermal vibrations virtually independently of one another. Since a certain interaction is necessary to attain the thermal equilibrium, there is assumed to exist a negligibly small interatomic coupling, enabling the atoms to exchange their energies.

For the present we shall overlook the fact that an atom in a solid has three vibrational degrees of freedom. In our model this number is reduced to one, i.e. we are concerned with an idealized solid in which the atoms behave as linear harmonic oscillators vibrating about fixed centres. These centres are arranged in space according to the points of a crystal lattice. According to quantum mechanics the energy levels of these localized oscillators are spaced equally from one another, i.e. in addition to their lowest energy ε_0 they may take up amounts of energy $\varepsilon_1 = h\nu$, $\varepsilon_2 = 2\,h\nu$, etc., where ν represents the frequency of their fundamental vibration and h is Planck's constant. Each energy level corresponds to one given quantum state of a particle.

We shall first consider a very small number of oscillators viz. 25, represented by one of the four horizontal layers of balls of one colour in fig. 1. At a temperature of absolute zero, all oscillators are in the state of energy ε_0, i.e. at the lowest energy level. When the temperature is raised by supplying energy to the system, particles are raised from the ground state to higher quantum states, i.e. to higher energy levels. The essential point in our consideration is again the number of different ways in which the energy can be distributed. If we supply a total of 25 energy quanta $h\nu$ to the 25 oscillators, either by applying heat dQ or by exerting work dW upon the system, we wish to know the number of possible ways in which the energy $\mathrm{d}U = 25\,h\nu$ can be distributed among the oscillators. As with the shaking experiment we shall ignore for the moment the fact that the system under consideration is too small for a profitable application of statistical-thermodynamics. For the time being our aim is only to demonstrate the method of counting. If each of the oscillators receives one quantum, then the interchanging of two atoms does not create a new state. In other words, the uniform distribution of the energy can only be realized in one way; it represents one micro-state and will, consequently, occur very rarely.

A less regular distribution shows an entirely different picture. An example of a distribution of this type is schematically represented in *fig. 2*. The six oscillators in the positions *C*3, *C*4, *C*5, *D*1, *D*2, *D*3 have each absorbed one quantum, the four oscillators in *D*4, *D*5, *E*1, *E*2 each two quanta, the two oscillators in *E*3, *E*4 each three quanta and the oscillator in *E*5 five quanta. The remaining twelve oscillators have not taken up any energy.

Because each oscillator has its own position in the "lattice", any interchanging of two oscillators having different quantum numbers will create a new micro-state. (This is not the case in a gas, in which each particle has access to the whole gas volume.)

	A	B	C	D	E
1	0	0	0	1	2
2	0	0	0	1	2
3	0	0	1	1	3
4	0	0	1	2	3
5	0	0	1	2	5

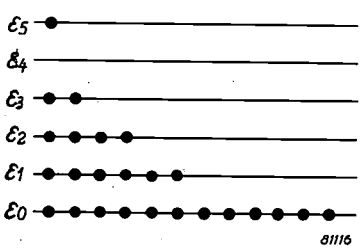
8115

Fig. 2. Schematic representation of a certain distribution of 25 energy quanta among 25 oscillators. Each small square corresponds to one oscillator. The number in it shows the number of quanta per oscillator. The numbers shown can be distributed among the squares in approximately 10^{12} different arrangements; each arrangement corresponds to one micro-state. The total number of possible distributions of 25 quanta among 25 oscillators is substantially greater, viz. approximately 6×10^{13}.

The number of distributions in which any six oscillators have absorbed one quantum each, any four oscillators two each, any two oscillators three each, and any one oscillator five quanta is found from

$$g = \frac{25!}{12!\,6!\,4!\,2!\,1!} = 9.4 \times 10^{11} \approx 10^{12}.$$

Instead of speaking of a distribution of 25 quanta among 25 oscillators as above, we may just as well state that the 25 oscillators have been distributed among the available energy levels in such a way that 12 are at the lowest level, 6 at level 1, etc. We thus obtain the diagram in *fig. 3*. In view of the



8116

Fig. 3. Macro-state comprising the 10^{12} micro-states, one of which is shown in fig. 2.

fact that the spatial positions *A*1, *A*2, etc. of the oscillators cannot be derived from this diagram, this schematic representation corresponds to a macro-state comprising the 10^{12} micro-states discussed above, one of which was shown in fig. 2.

The most probable macro-state and the total number of micro-states

The general expression for the number of micro-states forming a given macro-state of our idealized solid is given, according to the foregoing, by

$$g = \frac{N!}{n_0!\,n_1!\,n_2!\ldots}, \quad \ldots\ldots \text{(I,6)}$$

where N is the number of oscillators, and the numbers $n_0, n_1, n_2, \ldots$ represent the populations of the corresponding energy levels, i.e. the number of oscillators with energies $\varepsilon_0, \varepsilon_1, \varepsilon_2 \ldots$ (cf. fig. 3, in which $n_0 = 12$, $n_1 = 6$, etc.). Only those sets of populations are allowed which satisfy the auxiliary conditions:

$$\Sigma n_i = N, \quad \ldots\ldots \text{(I,7)}$$

$$\Sigma n_i \varepsilon_i = U, \quad \ldots\ldots \text{(I,8)}$$

in which $U = qh\nu$ is the total energy supplied over and above the zero-point energy (q = number of quanta absorbed).

The most probable macro-state, according to (I, 5), is the one containing the largest number of micro-states. Without entering into the calculation, we mention the fact that this maximum in g occurs (observing the auxiliary conditions (I, 7 and I, 8)), if the populations $n_0, n_1, n_2 \ldots$ form a descending geometrical series. For the numerical example under consideration, fig. 3 shows one of the most probable macro-states. (The *most* probable distribution for $N = 25$, $q = 25$, has the set of populations: $n_0 = 11$, $n_1 = 7$, $n_2 = 4$, $n_3 = 2$, $n_4 = 1$.) Different macro-states are characterized by different n_i-series. The total number of micro-states m is determined by the sum of all expressions of the form (I, 6) which satisfy the auxiliary conditions (I, 7) and (I, 8). In our example, m may also be evaluated in a far simpler way by directly counting the number of micro-states which form the MACRO-state defined by the number of oscillators N and the total energy U ($= qh\nu$). We may imagine that the q quanta and $N - 1$ of the N oscillators are arranged in arbitrary order along a straight line. At the right-hand end we place the last, the N^{th} oscillator. If the quanta between two oscillators are considered as belonging to the oscillator to the right, then each sequence represents a complete distribution, since the last position is always occupied by an oscillator.

The total number of micro-states is thus given by the total number of different arrangements possible on a straight line. Noting that the oscillator last placed in position has no ambiguity of location and that the interchange of two oscillators or of two quanta leaves the sequence unaltered, the total number of possible arrangements is given by:

$$m = \frac{(q + N - 1)!}{q!\,(N-1)!}. \quad \ldots\ldots \text{(I, 9)}$$

If this formula is applied to our numerical example, we find:

$$m = \frac{(25+24)!}{25!\,24!} = 6.3 \times 10^{13}.$$

With the two problems considered above, viz. the mixing of two kinds of particles and the distribution of quanta among localized oscillators it is permissible, if the number of particles or the number of quanta and oscillators is very large, to substitute the total number of micro-states for the number of micro-states of the most probable arrangement. This is justifiable for the same reason that it was permissible in the former example to replace 2^{N_0-40} by 2^{N_0} notwithstanding the fact that the latter is 2^{40} times greater. This reasoning is further reinforced by the fact — to be explained below — that we are not so much interested in g and m themselves as in the natural logarithms of these quantities, and while g_{max}/m becomes smaller and smaller with an increasing number of particles, the value of $(\ln g_{max})/(\ln m)$ approaches closer and closer to unity.

We shall demonstrate this by calculating these ratios for three different sets of populations of the three lowest levels, viz. for the following numbers of oscillators N and quanta q:

	1	2	3
N	111	1110	11100
q	12	120	1200

The most probable distributions for these three cases are given by the populations:

	1	2	3
n_0	100	1000	10000
n_1	10	100	1000
n_2	1	10	100

With the aid of (I, 6) and (I, 9), and using more accurate factorials [4]) than could be obtained via (I, 3), we can now obtain:

N	g_{max}	m	g_{max}/m	$(\ln g_{max})/(\ln m)$
111	5.2×10^{15}	1.3×10^{16}	4×10^{-1}	0.975
1110	1.0×10^{168}	2.5×10^{169}	4×10^{-2}	0.992
11100	3.6×10^{1699}	8.7×10^{1705}	4×10^{-7}	0.996

Hence we see that as the number of particles increases, the value of $(\ln g_{max})/\ln m)$ approaches unity. Real crystals usually contain at least 10^{20}, and in most cases from 10^{21} to 10^{24} atoms. For such numbers of oscillators m is enormously greater than g_{max}, whilst at the same time the difference between $\ln g_{max}$ and $\ln m$ is negligible.

The statistical (atomic) definition of entropy

The state of greater entropy towards which a system strives according to the second law, is, as has been demonstrated in the foregoing, the MACRO state with the largest number of equally probable arrangements, i.e. the most probable state. This brings us to the statistical definition of the concept of entropy:

$$S = k \ln g, \quad \ldots\ldots\ldots \text{(I,10)}$$

in which g is the number of micro-states from formula (I, 5) and k is Boltzmann's constant, derived from

$$k = R/N_0 \ldots\ldots\ldots \text{(I,11)}$$

(R = gas constant; N_0 = Avogadro's number).

As explained in detail above, one may in many cases substitute the total number of micro states m for the number of micro-states g of the most probable arrangement. One then obtains the alternative statistical definition of S:

$$S = k \ln m \ldots\ldots\ldots \text{(I,12)}$$

The statistical evaluation of the entropy thus amounts to counting numbers of micro-states. As an example of such a procedure we shall calculate the increase in entropy which accompanies isothermal expansion of a perfect gas. If we attempted this starting from the quantum states of the molecules, the derivation would become rather complicated, due to the fact that in quantum mechanics each micro-state bears a relation both to the energy and to the distribution of the molecules in space (see above), and the intervals between energy levels will decrease due to the enlargement of the volume. In the case in question it is permissible and more convenient to use the classical method of considering separately the number of possible molecular distributions in the available space and their distribution among the various velocities. So far as the velocity distribution is concerned, the entropy does not change on isothermal expansion. The increase in entropy to be calculated is entirely that due to the larger space that becomes available to the molecules. Let us imagine the volume containing the gas to be divided into a very large number of unit cubes, so small that the large majority of the cubes are empty, whilst a small fraction of them contain one gas molecule each. Due to the thermal motion, the arrangement of occupied and unoccupied cubes is continuously changing. If there are z cubes

[4]) Here we use the more accurate approximation $n! = \frac{n^n}{e^n}\sqrt{2\pi n}$.

and n molecules, then there are n occupied and $(z-n)$ unoccupied cubes. Because the mutual interchanging of two empty cubes as well as that of two occupied cubes leaves the distribution unaltered, the total number of micro-states is given by

$$m = \frac{z!}{n!\,(z-n)!}, \quad \ldots \quad \text{(I,13)}$$

or, using formula (I, 3):

$$\ln m = z \ln z - n \ln n - (z-n) \ln (z-n).$$

Using the approximation $\ln (1 - n/z) = -n/z$ for $n/z \ll 1$, we arrive at:

$$\ln m = n \ln \frac{z}{n} + n. \quad \ldots \quad \text{(I,14)}$$

If the volume is increased by a factor $r = V_2/V_1$, then rz has to be substituted for z in the formula. The increase of the entropy is thus, according to (I, 12):

$$\Delta S = k \ln m_2 - k \ln m_1 = kn \ln \frac{rz}{n} - kn \ln \frac{z}{n}$$

$$= kn \ln r = kn \ln \frac{V_2}{V_1}.$$

For 1 gram-molecule of gas, upon reference to (I, 11), we arrive at:

$$\Delta S = R \ln \frac{V_2}{V_1}. \quad \ldots \quad \text{(I,15)}$$

The two definitions of entropy

With the formulae (I, 2) and (I, 10) or (I, 12) we have given two definitions of entropy that seem at first sight unrelated and even appear to lead to contradictory conclusions. If, for instance, we double the volume of a gram-molecule of a perfect gas by allowing the vessel in which it is contained to communicate with an equally large evacuated vessel, then the entropy will increase according to (I, 15), although there is no flow of heat into or out of the system. At first sight one might be tempted to think that application of formula (I, 2) would lead to an entropy change of zero.

On further consideration, however, one sees that formula (I, 2) cannot be simply applied to this typically irreversible process. This formula applies only to reversible processes, and in order to calculate the change in entropy, a reversible process must be found that leads from an identical initial state to the same final state. Such a process is as follows. The gas is contained in a cylinder having a piston that can move without any friction. In constant temperature surroundings, the expansion is made to take place in such a way that the back pressure on the piston is at all times an infinitesimally small fraction less than the gas pressure. In these circumstances the process can be regarded as being reversible, because the system is in equilibrium in any stage of the process, so that an infinitesimally small change in the back pressure is sufficient to reverse the process. During the reversible expansion the perfect gas performs work on its surroundings, given by

$$\int dW = -\int p\,dV.$$

(Work done on a system and heat applied to a system are designated as positive). Since the internal energy of the gas is not changed by the isothermal expansion, according to (I, 1) a quantity of heat will be absorbed from the surroundings, given by

$$\int dQ = +\int p\,dV.$$

According to (I, 2) the change in entropy is given by

$$\Delta S = \int_{V_1}^{V_2} \frac{dQ}{T} = \int_{V_1}^{V_2} \frac{p\,dV}{T} = \int_{V_1}^{V_2} R \frac{dV}{V},$$

i.e., $$\Delta S = R \ln \frac{V_2}{V_1},$$

which agrees with the result (I, 15).

The same change of entropy is bound to occur with the *irreversible* expansion from V_1 to V_2, in view of the fact that S is a thermodynamic function. During this irreversible change, however, no heat is exchanged with the surroundings, so that

$$dS > \frac{dQ}{T},$$

and this has general application to all irreversible processes. The second law of thermodynamics can, therefore, be expressed in a form more general than (I, 2) as:

$$dS \geqq \frac{dQ}{T}, \quad \ldots \quad \text{(I,2')}$$

in which the symbol $=$ applies to a reversible change of state, and the symbol $>$ applies to an irreversible change. For an isolated system $dQ = 0$ is always valid, and hence, according to (I, 2), $dS \geqslant 0$. Formula (I, 2'), is therefore the mathematical formulation of the written form of the second law as in the introduction, viz. the entropy of an isolated system strives towards a maximum. It was the analogy between the classical thermodynamical picture of an isolated system striving towards maximum entropy, and the atomic picture of the system striving towards the state with the maximum number of equally probable arrangements g, that led, in the 19th century to the assumption (Boltzmann) that a relationship should

exist between S and g. This relationship could not be otherwise than of a logarithmic nature, since the entropy is an additive variable, whereas the number of equally probable arrangements is a multiplicative variable, as we have already seen. That S is an additive variable is a direct conclusion from the fact that the total amount of heat necessary in order to raise the temperature of the system $A + B$, in a reversible manner from T to $T + dT$, is the sum of the quantities of heat required to raise the temperatures of A and B separately to the same extent (cf. formula (I, 2)). The value of the proportionality constant k in (I, 10) and (I, 12) could then be directly derived from the application of this formula to a perfect gas, as described above.

A different formula for the entropy

According to formula (I, 6) the entropy of a system of N oscillators with energy level populations n_i can also be written as:

$$S = k \ln g = k\,[N \ln N - \Sigma\, n_i \ln n_i]\,.$$

Since $N = \Sigma\ n_i$ we may write:

$$S = -k\,\Sigma\, n_i \ln (n_i/N)$$

and for the entropy per oscillator:

$$s = -k\,\Sigma\, p_i \ln p_i, \quad \ldots\ldots\ldots \quad \text{(I,16)}$$

in which the fractions $p_i = n_i/N$ represent the fractions of the total number of oscillators at different energy levels i.

In some books on thermodynamics the last formula is chosen as the statistical definition of entropy. Outside the field of statistical thermodynamics an entropy formula similar to (I, 16) is used in information theory. The values p_i then relate to the probability of occurrence of certain possible events. We shall return to this subject in the last article (IV) of this series.

Justification of the first and second laws

Up to now we have paid hardly any attention to the historical path leading to the mathematical statements

$$\mathrm{d}U = \mathrm{d}Q + \mathrm{d}W, \quad \ldots\ldots \quad \text{(I,1)}$$

and

$$T\mathrm{d}S \geqq \mathrm{d}Q \quad \ldots\ldots\ldots \quad \text{(I,2')}$$

of the first and second laws. From a logical point of view it is perhaps most satisfactory to regard these relationships as postulates, and find their justification in the fact that all conclusions derived from them are confirmed by experiment. Indeed, if a single experiment were to be devised whose results contradicted one of the two laws, the whole admirable structure of thermodynamics would collapse. The concept of absolute temperature T introduced in (I, 2') finds its justification in an analogous way: it proved to be identical to the temperature scale derived experimentally by measurement with a gas thermometer.

The statistical definitions (I, 10) and (I, 12) of entropy are justified by the fact that in all cases studied up to now they are found to lead to the same results as those derived from the thermodynamical definition (I, 2). This has been demonstrated with one example, viz. that of a perfect gas expanded isothermally from V_1 to V_2.

Free energy

In order to enlarge on the significance of the concept of entropy we have in the mixing experiments so far considered only the perfect gas state, i.e. a state of matter in which there is negligible interaction between the atoms. In this condition the state of equilibrium always corresponds to a disordered distribution of the various types of atoms. This is not necessarily the case if forces of attraction are present between the atoms, or if the distribution of the atoms is influenced by external forces or fields. If, in our case of the 100 billiard balls, the 50 white balls suffered strong mutual attraction (e.g. by magnets incorporated inside each of the balls), then after a shaking experiment one would, in most cases, find a distribution such that the system is separated into two phases, one containing almost exclusively white balls and the other almost exclusively red balls. In this case the *disordered* distribution of the balls corresponds to a state of greater energy. The striving towards maximum entropy can, therefore, in certain cases be apparently counteracted by another tendency, viz. the striving towards a minimum energy. To formulate a quantitative relationship, the formulae (I, 1) and (I, 2') are combined to give

$$\mathrm{d}U - T\mathrm{d}S \leqslant \mathrm{d}W, \quad \ldots\ldots \quad \text{(I,17)}$$

or, for a constant value of T:

$$\mathrm{d}(U - TS)_T \leqslant \mathrm{d}W. \quad \ldots\ldots \quad \text{(I,17a)}$$

If in the course of the change of state not only the temperature remains constant, but also the volume, (and any other parameter whose change would lead to the performance of external work), then $\mathrm{d}W = 0$, so that (I, 17a) for an irreversible process can be written as

$$\mathrm{d}(U - TS)_{T,V} < 0\,. \quad \ldots\ldots \quad \text{(I,18)}$$

Hence, where we are concerned not with an

isolated system, but with a system in which the temperature T and the volume V are kept constant, then the tendency towards maximum entropy is replaced by the tendency of the function $(U-TS)$ towards a minimum. This thermodynamic function is called the *Helmholtz free energy* or the free energy at constant volume, and is indicated by the symbol F.

If instead of temperature and volume, temperature and pressure are kept constant, then even if the process is irreversible, work is done by the system, viz. the expansion work $-\mathrm{d}W = p\mathrm{d}V$ (p representing not the internal, but the external pressure). Referring back to (I, 17a), it will be seen at once that here another thermodynamic function, viz. the function $(U - TS + pV)$ tends towards a minimum. This function is called the *Gibbs free energy*, or alternatively the free energy at constant pressure, the free enthalpy, or the thermodynamic potential, and is usually indicated by the symbol G. We thus obtain for irreversible processes, depending on the auxiliary conditions, the two relations

$$(\mathrm{d}F)_{T,V} < 0 \quad \text{and} \quad (\mathrm{d}G)_{T,p} < 0 . \quad . \quad \text{(I,19)}$$

Finally formula (I, 17) provides the following relations applying to an irreversible change of state:

$$(\mathrm{d}U)_{S,V} < 0 \quad \text{and} \quad (\mathrm{d}S)_{U,V} > 0 \quad . \quad . \quad \text{(I,20)}$$

subject, again, to the condition that the volume and any other parameters whose change would lead to the performance of external work are kept constant.

The relation (I, 18), viz. $\mathrm{d}(U - TS)_{T,V} < 0$ may be more or less arbitrarily divided into two parts, $(\mathrm{d}U)_{T,V} < 0$ and $(\mathrm{d}S)_{T,V} > 0$, which may be considered as the mathematical statement of the two opposing tendencies mentioned above in this section. If only the latter $(\mathrm{d}S > 0)$ were operative, we would expect to find only those processes and reactions in nature in which the number of equally probable arrangements (the "disorder") increases. If, on the other hand, only the former $(\mathrm{d}U < 0)$ were operative, we would expect the occurrence of only those processes in which the opposite happens, i.e. whereby heat is liberated and, generally speaking, the degree of order increases. To determine the direction of a process it is therefore necessary to take into account the free energy which involves both thermodynamic functions U and S. From (I, 18) it follows that at low temperatures, the tendency towards minimum energy and the corresponding order predominates, whereas at high temperatures (violent shaking of the billiard balls) the tendency towards maximum entropy and the corresponding disorder, prevails.

These conclusions are borne out in practice. At low temperatures the atoms and molecules, under the influence of their mutual attraction, form the ordered, periodic structures we know as crystals. At high temperatures, however, all matter is ultimately transformed into the chaotic state of a gas. The temperature range in which this entropy effect begins to predominate depends on the magnitude of the attractive forces between the gas particles. Even in the gases a certain degree of order is often present in the form of ordered groups of atoms (molecules). At high enough temperatures, however, this order also disappears, and at the surface of the sun (temperature approximately 6000 °C) matter only exists as a gaseous mixture of the atoms of the various elements.

Even this is not complete chaos, however; as the temperature assumes higher and higher values even the order of the electron shells is finally completely destroyed due to thermal ionization. At temperatures of a few million degrees the ionization is complete. This state in which the atoms are entirely split up into naked nuclei and electrons occurs in the interior of the sun and other stars. The only order remaining is that of the nuclei. A complete disintegration of these into protons and neutrons would require even higher temperatures than seem to occur in the hottest stars.

Zero-point entropy

After this brief digression into the field of extremely high temperatures we shall now consider that of extremely low temperatures. According to *Nernst's heat theorem*, the entropy of all systems in a stable or metastable equilibrium tends to zero on approaching the absolute zero of temperature. This means, according to (I, 10) or (I, 12) that at absolute zero a system in equilibrium can exist in only one micro-state. This situation can be readily interpreted in terms of the quantum-mechanical picture: all particles are in their lowest quantum state at absolute zero. In the diagram of fig. 3 the higher levels gradually empty as the temperature decreases, until ultimately all 25 particles are at level ε_0. Nernst's theorem would not apply, however, if the lowest energy level corresponded to two or more quantum states, in other words, if this level was "degenerate". In that case, the particles in the state of equilibrium would be uniformly distributed among these quantum states even at absolute zero. Nernst's theorem thus necessarily includes the postulate of the non-degenerate state of the lowest energy level. The validity of this theorem can be tested experimentslly in many cases.

One of the classical examples is the transformation of white into grey tin. Grey tin is stable below 13 °C (286 °K), white tin is stable above this temperature. Owing to the fact that white tin can be supercooled right down to the lowest temperatures attainable in the laboratory, it has been possible to measure the specific heat c_p of both modifications at low temperatures. The validity of Nernst's theorem can thus be tested as follows. From (I, 2) the entropy of white tin at 286 °K can be found in two ways: directly from the c_p-measurements on white tin, and indirectly from the value of c_p for grey tin and the change in entropy occurring with the transformation of grey into white tin. Both ways should lead to the same result, i.e. the following equation should be valid:

$$\int_0^{286} \frac{c_p(\mathrm{w})\mathrm{d}T}{T} = \int_0^{286} \frac{c_p(\mathrm{g})\mathrm{d}T}{T} + \frac{Q}{286}, \qquad \text{(I,21)}$$

in which $c_p(\mathrm{w})$ and $c_p(\mathrm{g})$ represent the specific heat per gram-atom of white and grey tin respectively, and Q stands for the heat of transformation, i.e. the quantity of heat absorbed during the isothermal and reversible transition of 1 gram-atom of tin from the grey to the white modification. It has been found that (I, 21) is satisfied within the limits of experimental accuracy. Unfortunately the heat of transformation Q is not known with sufficient accuracy to attach very much value to this agreement. Moreover, even if complete agreement were established this would only prove that the *difference* in entropy between the two modifications at 0 °K is equal to zero. The heat theorem is therefore often worded in a somewhat more cautious form, e.g.: at zero absolute temperature all entropy *differences* between the states of a system in internal equilibrium vanish. This formulation has the same significance in practice as that which states that the separate entropy values approach zero, for one can now justifiably *define* the zero point entropy of all substances in stable or metastable equilibrium, to have the value zero. If, for example, in a chemical reaction of the type $A + B \rightleftarrows AB$, the entropy change is zero at 0 °K, then it is logical to assign to A and B as well as AB a zero-point entropy of zero. For the energy, this is not possible, as the extrapolations to $T = 0$ clearly demonstrate that there is no question of the heat of reaction disappearing at zero temperature.

Stronger evidence for the validity of Nernst's theorem is derived from measurements on gases. With the aid of the statistical thermodynamical expression $S = k \ln m$ (or $S = k \ln g$) the entropy of many gases can be evaluated, using information on their molecular rotational and vibrational states derived from their spectra.

On the other hand, assuming the validity of Nernst's theorem, the gas entropy may also be calculated with the aid of the classical formula $\mathrm{d}S = \mathrm{d}Q_{\mathrm{rev}}/T$, employing existing data on the specific heats c of these substances in the solid, liquid and gaseous states and that on heats of transformation, heats of fusion, and heats of evaporation. The entropy of a substance in the gaseous state at temperature T, assuming no transformations occur in the solid state, can be written as:

$$S = \int_0^{T_f} \frac{c_{\mathrm{solid}}}{T}\,\mathrm{d}T + \frac{Q_f}{T_f} + \int_{T_f}^{T_e} \frac{c_{\mathrm{liq.}}}{T}\,\mathrm{d}T + \frac{Q_e}{T_e} + \int_{T_e}^{T} \frac{c_{\mathrm{gas}}}{T}\,\mathrm{d}T,$$

in which T_f and T_e respectively are the melting point and the boiling point, and Q_f and Q_e are the heat of fusion and the heat of evaporation.

This "calorimetric entropy" is thus obtained entirely without reference to the existence of atoms, being based merely on the results of calorimetric measurements; the "statistical entropy" on the other hand is evaluated by methods entirely independent of the *de facto* existence of the liquid and the solid states. It is most satisfying that the two ways as a rule lead to the same result, whilst the few exceptions that have been found can be satisfactorily explained. The nature of these exceptions may be twofold. Some only *appear* to be exceptions, caused by the fact that the measurements of the specific heat were not extended to a sufficiently low temperature; other exceptions are caused by the non-attainment of equilibrium at decreasing temperature. Neither of the two types of exceptions is contradictory to Nernst's theorem, which only claims validity for the absolute zero temperature, and even then only for systems in a state of internal equilibrium.

These apparent exceptions to the agreement between calorimetric and statistical entropy occur if the "lowest energy level" of the particles, upon further scrutiny, is found to consist of a group of energy levels at intervals $\Delta\varepsilon$ which are small compared to kT even at the lowest temperatures of measurement [5]). When this is the case, the particles

[5]) The distribution among the available energy levels is entirely determined by the ratio $\Delta\varepsilon/kT$, where k is Boltzmann's constant.

are still uniformly distributed over the aforementioned group of energy levels, even at this lowest measuring temperature. The states corresponding to these energy levels will, therefore, not be manifest in the specific heat. The gradual emptying of the higher levels of the group will only start when the temperature reaches a value for which kT is of the same order of magnitude as $\Delta\varepsilon$; not before temperatures are reached for which kT is appreciably smaller than $\Delta\varepsilon$ will all particles have settled at the lowest level ε_0. This regrouping will manifest itself by a peak in the specific heat temperature curve (see *fig. 4*). If such a peak has not been established because the measurements have not been extended to a sufficiently low temperature, the value of the calorimetric entropy derived from them will be too low.

The fact that too small a value for the calorimetric entropy is sometimes found, may thus be

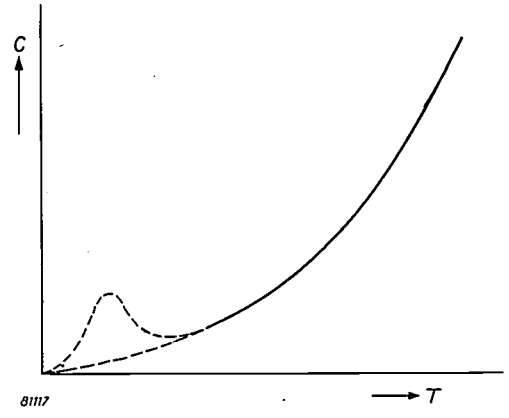


Fig. 4. The calorimetric entropy can be calculated if the specific heat c is known as a function of the temperature T. Extrapolation from the lowest attainable temperatures to absolute zero is of course necessary, but it must be noted that this involves the risk of overlooking a peak in the specific heat curve situated below the temperature range in which measurements can be made.

due to the extrapolation of the specific heat from too high a temperature. The cause may, however, also be a different one, namely the non-attainment of equilibrium, mentioned earlier. An example of such a "frozen-in" distribution of molecules among various energy levels, is solid carbon monoxide (CO) at low temperatures. The value found for the calorimetric entropy is smaller by an amount $R\ln 2 = k\ln 2^{N_0}$ cal/mol.degree than the statistical entropy derived from the CO-spectrum. This discrepancy corresponds to a number of microstates $m = 2^{N_0}$. This immediately suggests that the molecules in the crystal have two possible orientations. The CO-molecules are assumed to have such a small electric moment and to be so nearly symmetrical that the crystal lattice does not show a pronounced preference for the one or the other orientation, CO or OC. As a consequence of this, the two opposed directions of orientation would remain irregularly distributed among the positions of the lattice down to the lowest measurable temperature. It is highly improbable that in solid CO the rotational shift through 180° required to produce a state of equilibrium, is at all possible. In other words, it is not to be expected that an extension of the measurements to lower temperatures will eliminate the discrepancy. If this assumption is true, then the discrepancy between the statistical and the calorimetric entropy can only mean that the energy difference between the two CO-positions at the freezing point is still too small with respect to kT_f to determine a given orientation. Other examples of systems possessing a zero-point entropy due to the fact that the internal equilibrium is not attained at low temperatures, can be found among the many disordered solid solutions of metals. Further examples are the glass-like substances, which may be considered as supercooled liquids.

Nernst's theorem cannot be derived from the two main laws and is therefore often referred to as the "third law of thermodynamics".

In the three subsequent articles we shall demonstrate with the aid of examples that the concept of entropy plays an important part in widely divergent fields of science and technology.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Administration of the Philips Research Laboratory, Eindhoven, Netherlands.

2122: W. J. Oosterkamp: Die Dosierung weicher Röntgenstrahlung, insbesondere bei Kontakttherapie (Strahlentherapie **91**, 591-494). (Dosimetry of soft X-rays, with particular reference to contact therapy; in German.)

A description is given of ionization chambers for the measurement of the heterogeneous, extreme soft radiation emitted by X-ray tubes with a beryllium or mica-beryllium window. Attenuation curves in aluminium and depth dose curves in a water phantom for 10, 15, 20, 30 and 50 kV, measured on a new constant-potential contact therapy apparatus are given.

2123: J. Fransen and W. J. Oosterkamp: A universal X-ray dosemeter (Trans. Instr. Meas. Conf. Stockholm 1952, p. 93-94).

Brief description of ionization chambers and a dynamic electrometer for measuring dose rate and integrated dose for a wide range of radiation qualities and intensities.

2124: W. J. H. Beekman: An X-ray spectrometer for the 200 kV region using a scintillation counter (Trans. Instr. Meas. Conf. Stockholm 1952, pp. 84-87).

An X-ray spectrometer using a scintillation counter is described, constructed for the purpose of estimating the peak voltage of X-ray generators in those cases where conventional methods fail. The principle is that of measuring the shortest wavelength of the X-rays generated and determining the peak voltage by means of the Duane-Hunt relation. This instrument has been used in the case of an unorthodox form of X-ray generator, consisting of a grid-controlled X-ray tube, the anode of which is connected to a D.C. source via a high inductance coil. By applying negative pulses to the grid, the stray capacitance of the tube and coil can be alternately charged by the current through the coil and discharged through the tube.

2125: H. G. van Bueren: The formation of lattice effects during slip (Acta Metallurgica **1**, 464-465, 1953).

Tentative explanation of the formation of elementary structure (slip lines with a length of the order of 10^{-3} cm and 10-50 inter-atomic distances apart), as found by Wilsdorp and others in plastically deformed aluminium. The explanation is based on the effects of the formation of vacancies and interstitials during plastic flow.

2126: A. van Weel: Susceptance valves and reactance valves as phase modulators) J. Brit. Inst. Radio Engrs. **13**, 315-320, 1953).

Triode valves may be used in three different ways to give variable-impedance circuits. The most used form, usually known as a "reactance-valve circuit" behave much more as a variable "susceptance-valve circuit". Of the other two circuits, one is also a "susceptance-valve circuit", while only the third is a "reactance-valve circuit" proper. All three kinds can be used successfully in phase modulator stages, with the feature that, for introducing phase variations of up to 45 degrees, only one valve is necessary. Moreover, this may be realized with mutual-conductance variations of not more than 0.5 mA/V, from which follows both a high sensitivity and a good linearity.

2127: H. C. Hamaker: Beispiele zur Anwendung statistischer Untersuchungsmethoden in der Industrie (Mitt. Math. Statist. **5**, 211-229, 1953).

The application of "variance analysis" to industrial problems is explained with the aid of a suitable example, viz. the measuring of the thickness of the oxide coating deposited on nickel bars by electrophoresis. The results of the analysis are presented in a simple manner which can be understood by persons not specially trained in statistics. The same method is then applied to some other examples, viz. the measuring of the heat of combustion of 9 different coal samples by 9 different laboratories, the measuring of the diameters of 6 different bicycle bearing-balls with 7 different micrometers, and smelling-tests for the selection of an odour-judging panel in the perfume industry. The article also deals in some detail with the concept of interaction and how its existence can be established by statistical analysis.

2128: H. P. J. Wijn: Ferromagnetic domain walls in Ferroxdure (Physica **19**, 555-564, 1953).

In the preparation of $BaO.6Fe_2O_3$ (Ferroxdure) it is possible to distinguish between the contributions to the initial permeability of Bloch-wall displacements and of Weiss-domain rotations. From measurements of these contributions as functions of the frequency, there is evidence of a resonance effect of the Bloch-walls at about 300 Mc/s, and the contribution of the rotations remains independent of the frequency to above 3000 Mc/s. The possibility of Bloch-wall resonance has several times been proposed in the literature. The resonance frequency to be expected from theoretical considerations agrees well with the observed results.

2129: W. Hoogenstraaten: The chemistry of traps in zinc sulphide phosphors (J.Electrochem. Soc. **100**, 356-365, 1953).

Trap characteristics of zinc sulfide phosphors are studied as a function of chemical composition. The simplest phosphor systems are found to have simple, single-peaked glow curves. They contain only activators and coactivators in pure sulfide base materials. Coactivators are defined as impurities necessary to stabilize the activators in the zinc sulfide lattice. They are found to exert a major influence upon trap characteristics. The trap depths are found to be 0.37 electron volt for Cl^-, Br^-, and Al^{3+}, 0.51 eV, for Sc^{3+}, 0.62 eV for Ga^{3+}, and 0.74 eV for In^{3+} as coactivators in ZnS-Cu. Additional glow peaks and traps are produced by oxygen and by the killers cobalt and nickel. The formation of mixed crystals with cadmium sulfide or zinc selenide generally results in a shift of the glow curves toward lower temperatures.

2130: K. ter Haar and J. Bazen: The titration of "Complexone III" with thorium nitrate at $pH = 2.8 - 4.3$ (Anal. chim. Acta **9**, 235-240, 1953).

The reaction between thorium and "Complexone III" (disodium salt of ethylenediaminetetra-acetic acid) at $pH = 2.8 - 4.3$ has been developed to a quantitative method. As indicator alizarin-S is used. In the first place the reaction is suitable to back-titrate an excess of "Complexone III" in the *pH* range mentioned and it is the basis for an Al, Ni and Bi determination still to be published; moreover, it probably presents the possibility of a simple determination of thorium.

2131: A. J. W. M. van Overbeek and F. H. Stieltjes: Bandwidth limitation of junction transistors (Proc. Inst. Radio Engrs. **40**, 1424, 1952).

Proceeding from results obtained by Steele, a fundamental limiting value to the *Q*-factor for wide-band amplification is derived for a junction transistor. The output capacitance is neglected in this analysis.

2132: J. Feddema and W. J. Oosterkamp: Volume doses in diagnostic radiology (from: Modern trends in diagnostic radiology, 2nd series, edited by J. M. M. McLaren, Butterworth, London, p. 35-42, 1953).

The importance of dose measurement in diagnostic radiology is pointed out. Methods are described for arriving at the volume dose. A table is included of data on the average volume doses relating to a number of different diagnostic conditions. A survey is given of the number of exposures to which patients have been subjected during the course of their life. Case histories show no sign of radiation damage to the patients.

2133: J. L. Meijering: On a statement by C. S. Smith concerning an upper limit to the sharing of corners in aggregates (Acta Metall. **1**, 607, 1953).

Contrary to the proposition put forward by C. S. Smith, it is asserted that the number of corners in a crystal aggregate can be greater than 6 times the number of crystals, even if all the crystals are convex polyhedra with flat boundaries.

2134: H. G. van Bueren: Relation between plastic strain and increase of electrical resistivity of metals (Acta Metall. **1**, 607-609, 1953).

The increase of electrical resistivity of metals due to plastic strain at low temperatures is attributed to the formation of vacancies, interstitial atoms and dislocations. On the basis of the considerations treated in Abstract No. 2125, it is demonstrated that the influence of dislocations gives rise to an increase proportional to $\varepsilon^{1/2}$ (ε = elongation) whilst vacancies and interstitial atoms cause an increase proportional to $\varepsilon^{3/2}$. Manintveld's experiments indicate a $^3/_2$-power relationship, which shows that dislocations have little influence, a conclusion which appears also theoretically justified. This helps to explain the influence of annealing on the resistivity.

2135: W. Hoogenstraaten and H. A. Klasens: Some properties of zinc sulfide activated with copper and cobalt (J. Electrochem. Soc. **100**, 366-375, 1953).

Some properties of ZnS-Cu-Co phosphors under 3650 Å excitation are described, viz., the thermal glow, decay and build-up of fluorescence, tempera-

ture dependence, and light sum. Most of the experimental results can be explained by a model in which cobalt levels act both as electron traps with a trap depth of 0.5 electron volt, and as acceptors for holes, ejected thermally from copper centers with an activation energy of 1.1 eV. The possibility of excitation by 3650 Å radiation of electrons from traps to the conduction band is introduced to explain the observed intensity dependence on the light sum.

2136: J. D. Fast: Erzeugung von reinem und absichtlich verunreinigtem Eisen und Untersuchungen an diesen Metallen (Stahl und Eisen **73**, 1484-1496, 1953).

To investigate the causes of various phenomena occurring in steels, a high-vacuum 300 kc/s induction furnace was developed which is suitable for melting very pure iron in quantities up to 2 kg. Some general directions regarding the melting-procedure and the choice of crucible material are given. (See also Philips tech. Rev. **15**, 114-121, 1953/1954). Starting with iron in its purest form, investigations were carried out into the individual and collective influence of carbon (up to 0.04%), oxygen (up to 0.03%), nitrogen (up to 0.02%) and manganese (up to 0.50%), on (*a*) quench ageing, (*b*) strain ageing, (*c*) blue-brittleness, (*d*) grain-boundary brittleness. Regarded from an atomic viewpoint, a common cause is found for a number of phenomena that seem at first sight to be hardly, if at all, interrelated. These are, the greater solubility of carbon and nitrogen in γ-iron, compared to that in α-iron; the greater solubility of nitrogen in both phases compared with that of carbon; the presence of the dissolved carbon and nitrogen atoms in the octahedral interstices of both phases; the Snoek-damping; the formation of martensite; the occurrence of an upper and a lower yield point in the stress-strain curve of mild steel; the strain ageing; the fact that carbon and nitrogen are less soluble in silicon iron than in pure iron; and the preference of the dissolved carbon and nitrogen atoms for the grain boundaries of iron.

2137: J. I. de Jong and J. de Jonge: The chemical composition of some condensates of urea and formaldehyde (Rec. Trav. chim. Pays-Bas **72**, 1027-1036, 1953).

Some condensates of urea and formaldehyde were prepared from solutions of *pH* 2-7, at temperatures of 20-76° C. These products have been analysed with respect to their content of methylene groups, methylol groups and urea groups. The average molecular weights could be estimated. The analytical data are in harmony with the occurrence of methylene bridges between the urea fragments. The condensates will be formed by a stepwise condensation reaction. "Methylene urea" may be a mixture of condensates with an average molecular weight of 300 — 500.

2138: J. S. C. Wessels and E. Havinga: Studies on the Hill reaction, II (Rec. Trav. chim. Pays-Bas **72**, 1076-1082. 1953).

The influence of the presence of oxygen on the Hill reaction has been investigated by redox potential measurements. A very simple reaction scheme, implying primary formation of a specific reductant reacting subsequently with the oxidant added, seems to fit the kinetic and other data of the reaction. Some of the results of investigations on the influence of inhibitors and biochemically important substances are reported and discussed; possible causes for the discrepancies in the literature concerned with the action of inhibitors are indicated.

2139: K. F. Niessen: Ratio of exchange integrals in connection with angles between partial magnetizations in ferrimagnetic spinels (Physica **19**, 1035-1045, 1953)

Allowing the spin direction of magnetic ions in one sublattice of (octohedral) *B*-sites to differ from that in the other sublattice of *B*-sites (as assumed by Yafet and Kittel for a spinel with only one kind of magnetic ions) a mixed crystal spinel containing two kinds of magnetic ions is considered, taking into account the different physical nature of the ions. Here a situation may be realized where the partial magnetizations are neither parallel nor antiparallel but where in one sublattice (say B_{I}) two special spin directions occur for the two kinds of magnetic ions and in the other (B_{II}) another couple of spin directions lying with the former set symmetrical with respect to the single spin direction of ions on the (tetrahedral) *A*-sites. The *A*-*A* interaction is neglected and consequently a subdivision of the *A*-lattice is not taken into account. Compositions of the mixed crystal are possible where such a non-rectilinear case is just possible, i.e. where the mutual deviations of the spin directions on the *B*-sites are very small, their angles with the spin directions on *A*-sites being nearly π. From three such compositions the ratios of exchange integrals can be determined.

Vol. 16 No. 10, pp. 273-308 APRIL 1955

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

LIGHT-BEACONS TO AID LANDING AIRCRAFT

by J. B. de BOER.

628.971.8:629.139.1

During the Melbourne air race in 1934, one of the winning aircraft, the "Uiver" of K.L.M., made a remarkable emergency landing at night on an improvised runway at Albury in Australia. On this occasion, improvised lighting was provided by the headlights of motor cars parked side by side along the runway. Compared with such an achievement, landing an aircraft at night on an airfield equipped with modern aids is a relatively simple matter. Modern lighting equipment, which enables aircraft to land safely in darkness and fog, is essential to the smooth operation of the vastly increased volume of civil and military air transport.

About 15 years ago, an aerodrome was usually a flat, grassy field a mile or so square. Pilots were permitted to land or take off from such a field in any direction they chose. Flying took place only when weather and visibility were good, and night landings were accomplished with the aid of lights arranged, say, along the edge of the airfield, on various obstacles, and so on, and a more or less uniform overall illumination of part of the field (about 600 × 800 yds, with an illumination-level of a few tenths of a foot-candle). This illumination was provided by transportable light sources arranged on the boundary of the area giving a fan-shaped light beam in the landing direction along the surface of the airfield [1) 2)].

Since that time, the problem of aerodrome lighting as an aid to the landing of aircraft has changed radically for two reasons:

a) Aircraft of far greater weight and speed are flying nowadays, and concrete runways from 1000 to more than 2500 yds long and 100 to 200 ft wide are required to enable such aircraft to take off and land in complete safety. It is impossible to provide adequately uniform overall illumination over such an area.

b) It is the aim nowadays to make air-transport less and less dependent upon weather-conditions; hence aircraft must be enabled to land not only day or night when visibility is good but also in bad visibility at night and — most difficult of all — under conditions of mist or fog in the daytime. In these latter circumstances it is impossible to make objects other than light sources visibile at a sufficient distance.

These two facts have led, of necessity, to the replacement of *lighting* of the landing area, by *beaconing*. In fact, beaconing is the basis of all up-to-date systems of "airfield lighting".

However, also the light sources that can reasonably be used for this purpose, are visible in heavy mist at a distance of one or two thousand feet at the utmost. Such a short distance gives the pilot of an approaching aircraft no margin for important manoeuvres; landing in heavy mist has therefore become possible only with the help of a combination of light-beacon and radio-beacon systems (or radar information conveyed to the pilot by radio).

The radio beacons enable the pilot to to be guided so close to the airfield that he can just discern the lights, after which the pilot lands the aircraft visually with the aid of the lights, since radio beacons are not sufficiently accurate at short range.

The bad-visibility landing procedure may now be described more fully. Radio or radar is used to guide the aircraft as accurately as possible along the ideal glide path, that is, a line in the same vertical plane as the centre-line of the runway, sloping down at an angle of about $2^1/_2{}^\circ$ and cutting the above-

1) G. L. van Heel, The illumination and beaconing of aerodromes, Philips techn. Rev. 4, 13-99, 1939.
2) Th. J. J. A. Manders, Aerodrome illumination by means of water-cooled mercury lamps, Philips tech. Rev. 6, 33-38, 1941.

mentioned centre-line at a point about 1000 ft beyond the start (or "threshold") of the runway. By the time the aircraft has descended to a height of about 200 ft (at which the horizontal distance to the threshold of the runway is still about 3000 ft), visual "ground-contact" should be established. From this contact the pilot must obtain an immediate and unmistakable impression of a number of data: his direction of flight relative to the direction of the runway; the horizon; the altitude of the aircraft; the glide slope; the distance to the threshold; the lateral displacement from the centre-line of the runway (produced) and the ground speed of the aircraft. Experience has shown that it is possible to provide the pilot with all this information by the mere observation of a number of light-beacons provided these beacons are arranged in a suitable pattern, and show if necessary, distinctive colours. On the basis of this information obtained visually the pilot performs all further manoevres for completing the landing.

It will be evident from the above that the lights employed as visual guides do not have to be visible from all directions; fortunately, then, light sources of reasonable power can be used by beaming the luminous flux in the required direction. Again, the restriction of the light-yield to certain directions prevents excessive diffusion of the light in heavy mist, which would make the pattern of the lights difficult to distinguish from an approaching aircraft.

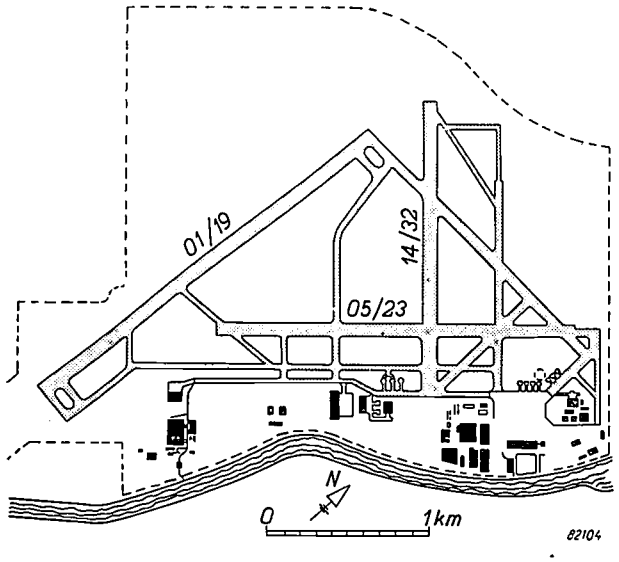


Fig. 1. Plan showing the present layout of Schiphol airport, Amsterdam. Each runway offers two possible directions for landing or take-off, depending on the wind, and must therefore be provided with approach and threshold lighting at both ends, and runway lights shining in both directions. In view of the directions of the prevailing winds, Schiphol is now equipped with two main runways, i.e. runway 05/23 (take-off and landing directions 50° or 230° to the S-N direction) and runway 01/19 (10° or 190° to the S-N direction). The beacon systems on these runways are formed by lights designed by Philips. (This plan, the diagram of fig. 3 and photographs of fig. 4, fig. 14 and fig. 21, is reproduced here by permission of the Schiphol Airport Authorities.

In this article it will be shown how the required luminous intensity and light-distribution of the beacons can be determined. The construction of some of the lights will then be described, with special reference to the lighting actually in use at a number of airports, e.g. on the runways at Schiphol Airport, Amsterdam (*fig. 1*).

For the sake of clarity, however, it is first necessary to deal more fully with how the pilot obtains visual guidance during a landing.

Configuration of the lights

Since the second world war, many experiments have been carried out on airfields, and in the laboratory, with the object of determining the most suitable pattern for runway beacons [3]). That the runway itself must be indicated by placing rows of lights along the sides and at the threshold is self-evident, but a certain amount of controversy has arisen in connection with the so-called approach lighting, that is, the lighting to indicate the position of the runway before the latter is actually seen through mist or fog. When once the IATA (International Air Transport Association) had decided its standpoint with regard to this question, the ICAO (International Civil Aviation Organisation) meeting in Montreal in November 1952 made a fairly clear choice between the various systems of approach lighting which had been proposed. *Fig. 2* shows one of the systems consistent with the recommendations of this organization, namely the Calvert system. Here, the extension of the centre-line of the runway is indicated by a luminous line, at right-angles to which, at specified intervals, are luminous transverse lines ("cross-bars"), arranged in order of diminishing length towards the runway. Each of these luminous lines really consists of a number of lights; hence the lines are usually seen as straight rows of luminous dots rather than as continuous lines.

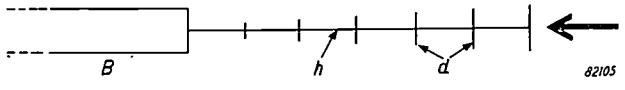


Fig. 2. Diagram showing the Calvert system of runway beaconing; the thick arrow indicates the landing direction. *B* runway, *h* centre-line, *d* cross-bars.

[3]) E. S. Calvert, J. Roy. Aeronaut. Soc. **52**, 439, 1948; E. S. Calvert, Trans Ill. Eng. Soc. (London) **15**, 183, 1950; H. J. Cory Pearson, C. A. A. Techn. Dev. Rep. Nos. 104 and 167, March 1950 and April 1952; A. N. Baldino, Bull. Soc. Trans. Electr. (6) **9**, 442, 1949; G. J. Malouin, Report to Flight-Technical Group IATA, New York, October 1951; J. W. Sparke and H. F. Ringe, Light and Lighting **43**, 259, 1950; F. C. Breckenridge, Illum. Eng. **47**, 1952.

A variation of this system, as shown in *fig. 3*, is employed on main runway 23 at Schiphol. *Fig. 4* is a photograph of this beacon system as seen from the cockpit of an approaching aircraft.

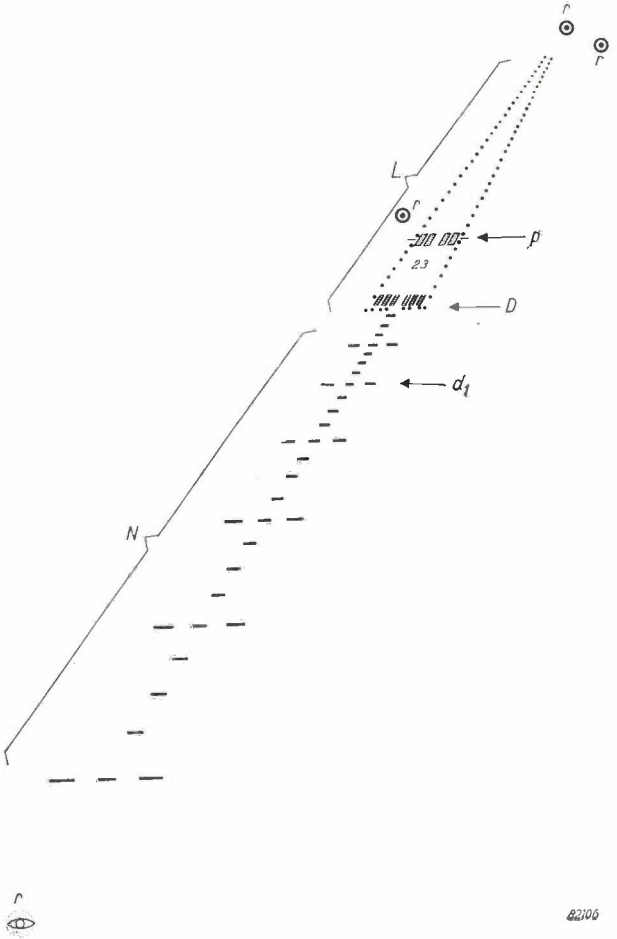


Fig. 3. Diagram showing approach lighting (*N*), threshold lights (*D*) and runway-marking (*L*) of runway 23 at Schiphol. d_1 is a special cross-bar in the approach lighting system coloured differently from the others (red), the so-called warning sign, and *p* is the so-called touch-down point, specially marked with paint and a number of lights, located 1000 ft beyond the threshold; *r* indicates the various radio beacons associated with the particular landing system.

Fig. 4. Beacon system of Schiphol runway 23 as seen from an aircraft coming in to land.

It is perhaps worth while to draw attention to the essential purpose of the cross-bars in the pattern. When an aircraft is approaching the runway along the "ideal" glide path and in the absence of cross-wind, the pilot will see the centre-line of the pattern as a vertical line straight ahead of him (a pilot, like a motorist can judge from experience what point corresponds to the direction "straight ahead" from its relation to the edge of his windscreen or the frame of his machine. In these circumstances, the pilot will be able to judge the distance to the threshold correctly if a suitable code is incorporated in the centre-line (e.g. differences in the colour or grouping of the individual lights). Assuming once more that there is no cross-wind, then, if the pilot sees the centre-line in the perspective pattern in another position, this may indicate: *a*) that the longitudinal axis of the aircraft is not parallel to the centre-line; *b*) that the aircraft is flying wide of the centre-line; or *c*) that the line joining the wing-tips is not horizontal. To illustrate these ambiguities, *fig. 5* shows two different situations in which the centre-line will be seen by the pilot as running in an identical direction. In the case shown in fig. 5*a*, the aircraft

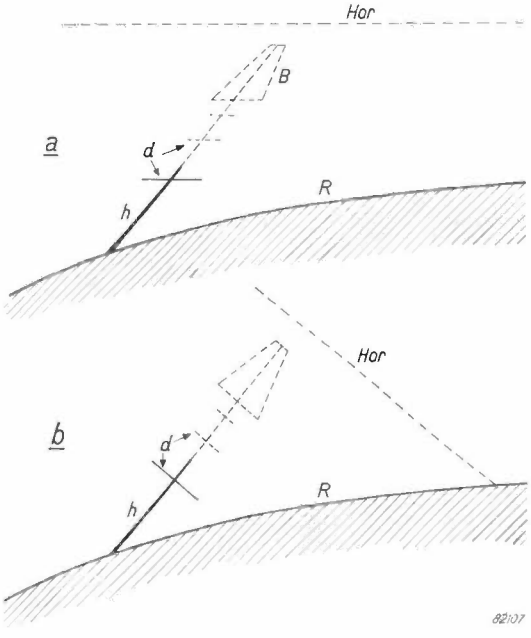


Fig. 5. Diagram to illustrate two different situations in which the pilot of an approaching aircraft will see the centre-line (*h*) of a runway (*B*) in exactly the same position relative to his cockpit cut-off. The horizon (*Hor*) is assumed to be completely obscured by fog in both cases. It is only by observing the attitude of the cross-bars (*d*) that the pilot is able to judge his position and direction of movement correctly.

is flying with the line joining its wing-tips horizontal, on a course parallel to, but slightly to the right of the runway: fig. 5*b* shows the situation when the aircraft is above the approach centre-line of the runway with its fuselage momentarily parallel to this line, but is banking to the left, so that the line joining the wing-tips slopes downwards to the left. It is only by observing the relative attitudes of the cross-bars that the pilot can judge which of these two interpretations corresponds to the real situation and so choose the proper manoeuvre to correct his course.

In view of the speed at which the aircraft travels, it is an interesting psychological problem as to how the pilot is able to deduce his position and movement in space from the movements of the different visible points and lines within his field of view. Some light is shed on this question by Calvert's so-called parafoveal streamer theory [3]). In view of the importance of this theory as a means of establishing present-day requirements for the light-distribution of runway beacons, Calvert's arguments will now be outlined.

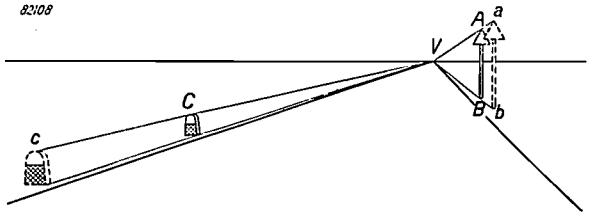


Fig. 6. Diagram showing the field of view of a motorist driving along a straight road. V is the vanishing point. Note the traffic sign (AB) and the mile post (C) as seen by the driver at a particular moment, and the same two objects (ab and c) as seen a moment later. From the apparent motion of all points outwards from the vanishing point, the driver is able to judge his own direction and speed.

Fig. 6 is a simplified diagrammatic representation of a perspective image as seen by a rapidly moving observer, e.g. the driver of a car travelling along a straight road. All the points within the field of view of such an observer appear to move outwards from a single fixed point, that is, the vanishing point (V), the point on the horizon corresponding to the infinitely distant "end" of the road). The speed of movement of a given point is proportional to the square of its apparent distance from V, and inversely proportional to the distance of this point from the line along which the observer's eye is moving. The more or less instinctive observations of the driver (or pilot) as to the direction and speed of movement of such points (in what may be described as the "dynamic field of view") enable him to estimate the direction and speed of his own movement; the less he alters his direction of view, the more accurate will his estimate be. In fact, films taken by Capt. Majendie of BOAC have shown that during the last phase of a landing, a pilot habitually "fixes" his head and eyeballs, that is, stares ahead at a point corresponding roughly to the point on the runway where the wheels of the aircraft will touch the ground. His line of sight then makes a small angle with the horizon: the zone visible to the pilot between the horizontal and about 10° below it is therefore the most important. Hence the light radiated by the beacons within the angle between 0° and 10° above the horizon contributes most to the visual guidance during landing.

Minimum requirements for visual guidance

The requirements to be imposed on approach systems and runway lights will now be defined more fully.

When once the pilot has established "ground-contact" (at a distance of 3-4 thousand feet from the runway-threshold), he must be able to see at any given moment a certain length of the system of lights in order to appreciate the overall pattern. This is necessary not only for the ideal glide path but also for other approach-tracks, within certain limits. These limits, or more precisely, the space within which visual guidance is to be assured, cannot be determined objectively. The information required to establish it can be supplied only by experienced pilots from their practical knowledge of flying. Recommendations concerning this zone were issued in September 1952 by the Flight-Technical Group of the I.A.T.A., an association well qualified to advise on such matters, since its members include many experienced pilots employed by the world's major air-lines. According to these recommendations, the space within which visual guidance must be provided may be imagined as a shaft of oval cross-section around the ideal glide path; the dimensions of this approach channel are indicated in *fig. 7*. An ordinary commercial aircraft approaching the runway anywhere outside this channel cannot readily, if at all, be so manouvred as to complete a safe landing; hence visual guidance outside the channel is of little value. On the other hand, standard radio or radar aids are quite accurate enough to "home" an aircraft to the "entrance" of the visual guidance channel defined by fig. 7.

Originally, it was considered necessary to require that a segment of constant length of the beacon system be visible from all points within the guidance channel. The light just visible above the bottom edge of the windscreen, the so-called cockpit cut-off (about 15° or 20° below the horizon, *fig. 8*), is a natural choice for the near limit of the visible segment. The consequence of the above consideration would be that the maximum visual range

(or luminous intensity) of the *approach* lights must be in quite a steep upward direction, and the maximum luminous intensity required of the *runway* lights may decrease at increasing distance from the runway-threshold.

It will be evident that such a system is altogether inconsistent with the parafoveal streamer theory, which holds that it is especially necessary that the pilot be able to see the lights within the angular zone between 0° and 10° below this horizon: this also includes the runway lights far ahead, whose luminous intensity would clearly have to be considerably higher than those of the approach lights and the lights at shorter distances.

Since the fulfilment of the above requirement would lead to enormous luminous intensities (see below), a compromise has been adopted: it is now required that all the lights become visible to the pilot at the same distance, i.e. 1500 feet. No light need be radiated at angles greater than 12-15° above the horizon except by the lights in the near part of the approach system, which must become visible to the pilot before he has "fixed" his eyes.

It will be seen that the required luminous intensity data of each individual light in all directions of radiation can be deduced from the above data and with the aid of the diagram shown in fig. 7, provided that the minimum conditions of visibility in which aircraft must be able to land are known. These conditions cover *a*) the background-brightness with which the lights must contrast and *b*) the transmission of the atmosphere. The luminous intensity to give the lights the required range will be considerably lower for night landings in good visibility than for daytime landings in fog. The I.A.T.A. recommends that a daytime landing in a meteorological visibility of only 1000 ft be considered the most unfavourable case in which full visual guidance should be provided. "Visibility" as a quantitative measure of the transmission of the atmosphere may be defined as the greatest distance (V) at which a dark object can be distinguished by day. The relationship between V and the transmission (r) of the atmosphere is defined in the formula of Koschmieder: $t^V = c$, where c is the contrast-sensitivity of the human eye. For the present type of observations, c may be taken equal to 0.04.

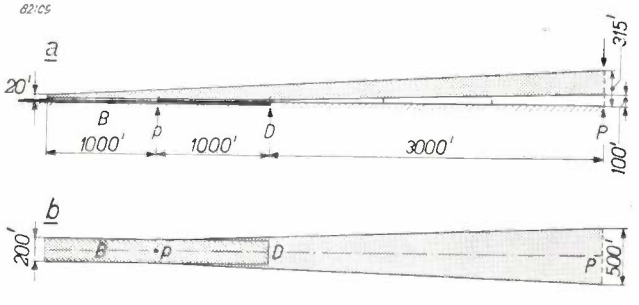


Fig. 7. *a*) Side view, and *b*) top view, of the imaginary shaft within which visual guidance should be given to an aircraft approaching for a landing. B runway, D threshold, p touch-down point, and P entrance of the visual guidance channel.

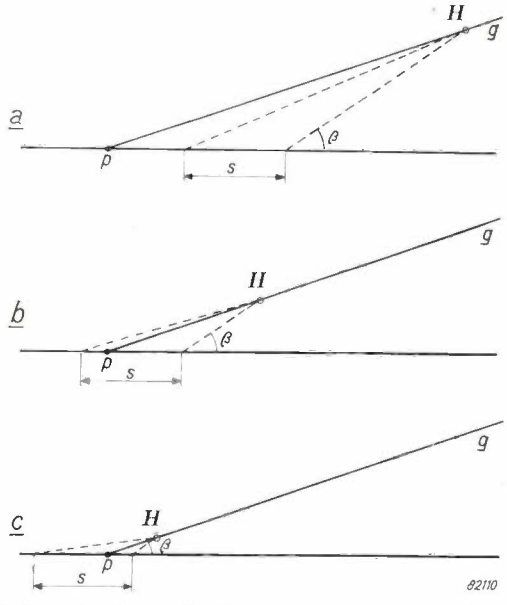


Fig. 8. In order to satisfy the requirement that the length (s) of the beacon system visible from an aircraft (H) approaching along the ideal glide path (g) be the same at all points along this path, the visual range of the initial approach lights would have to be longer than that of the others, in directions slanting quite steeply upwards. β is the limiting angle at which the pilot of an approaching aircraft will just be able to see a particular light above his cockpit cut-off. To illustrate the point more clearly, the glide path is here drawn at an exaggerated angle.

For reasons of economy, only those airfield runways corresponding to wind-directions frequently associated with relatively dense mist will be provided with approach lighting for $V = 300$ m. At Amsterdam, for example, a north-east wind very rarely brings fog; hence the two principal runways (01/19 and 05/23, see fig. 1) are provided with strong approach lights only at the north and north-east ends, respectively, the lighting at the south and south-west ends of these runways, where aircraft land in a prevailing north-east wind, being considerably weaker. The north-west to south-east runway (14/23 in fig. 1) is provided with relatively weak approach lighting at both ends.

Determination of the required luminous intensity and light-distribution of the beacons

The relationship between the visual range of r (in metres) and the luminous intensity I (in candelas) of a light source is defined by Allard's

formula:

$$I = \frac{E_c \cdot r^2}{t^r},$$

where t is the transmission of the atmosphere per metre and E_c the minimum eye-illumination (in lux) to make the light visible to the observer. Apart from the background-luminance, the required eye-illumination E_c also depends on several other factors, e.g. the apparent size, the form and the colour of the particular light source [4]). For night-flying, 10^{-6} lux is a suitable value of E_c, and for day-time flying 10^{-3} lux.

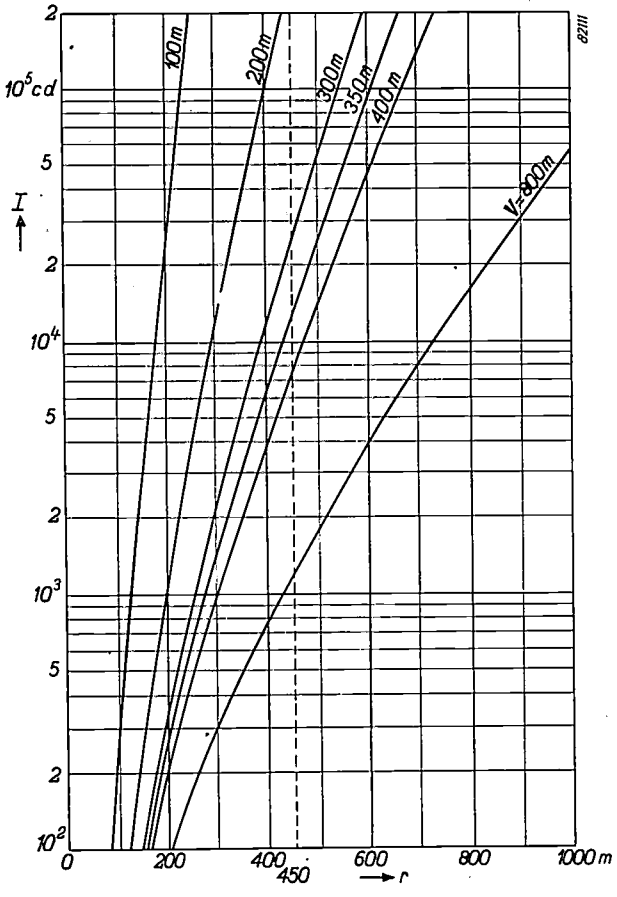


Fig. 9. Diagram showing the relationship between luminous intensity (I, in candelas) and visual range (r, in metres) for different meteorological visibilities (V, in metres). The curves are valid for an eye-illumination level $E_c = 10^{-3}$ lux.

Fig. 9, plotted for $E_c = 10^{-3}$ lux, shows that the effect of meteorological visibility on the luminous intensities required to ensure the prescribed visual range is considerable. It may easily be deduced from this diagram that any tightening of the requirements as to the range at which the lights must be visible or the maximum density of fog in which a landing may be attempted (V) will involve considerable expense. For a visibility $V = 400$ m (1300 ft), a light of 4000 cd luminous intensity can be seen at a distance of 400 m; to increase the range to 600 m (2000 ft) it, would be necessary to raise the luminous intensity of the light source to 45 000 cd. Again, in a visibility V of only 200 m, (670 ft) a range of 400 m will require a luminous intensity of 100 000 cd, and a range of 600 m some million cd. To appreciate the significance of these figures it should, be borne in mind that whereas a luminous intensity of the last-mentioned order is not unusual for light-houses, of which there are only fifteen along the entire coast of the Netherlands, it is altogether out of the question for runway lights, of which several hundreds are required on a single runway.

At $V = 300$ m, already referred to as the visibility upon which our calculations should be based, a luminous intensity of 13 000 cd is required for a range of 400 m in daylight; this would have to be raised to more than 200 000 cd to increase the range to 600 m. Hence the above-mentioned decision to specify a range which is not unduly long, that is, 450 m, the corresponding luminous intensity being 25 000 cd.

Let us now consider the required light-distribution.

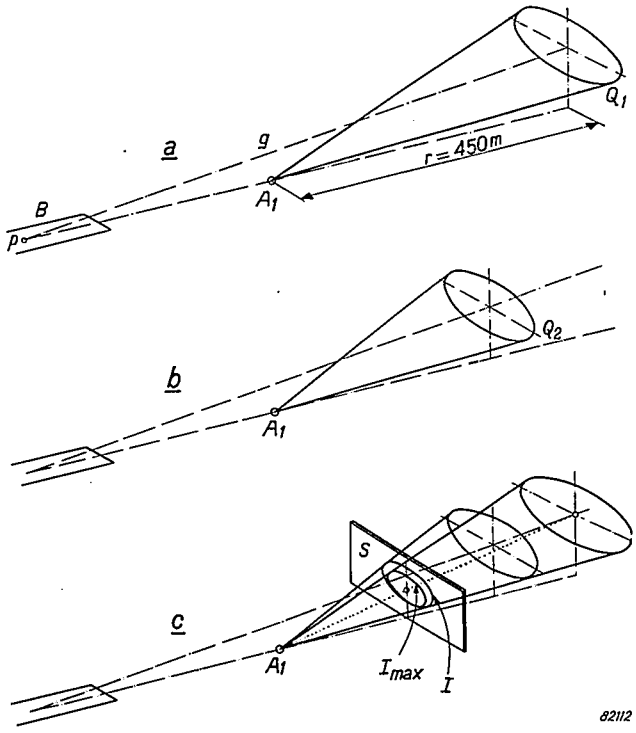


Fig. 10. *a*) A light (A_1) at a distance $r = 450$ m (1500 ft) from a particular cross-section (Q_1) of the visual guidance channel should possess a luminous intensity $I_{max} = 25\,000$ candelas in the directions of all the points within that cross-section (the small differences in distance of these points from the light may be ignored).
b) A lower luminous intensity (I) is sufficient in the directions of points within another cross-section of the channel (Q_2), closer to A_1.
c) Combination of diagrams (*a*) and (*b*), by projection on an imaginary screen (S), forming an isocandela diagram which defines the required light-distribution of the light A_1.

[4]) J. B. de Boer, Visibility of approach and runway lights, Philips Research Reports 6, 224-239, 1951.

An idea of this may be obtained from *fig. 10*. Fig. 10*a* shows a vertical cross-section of the region of visual guidance (fig. 7), taken at a point 450 m from one of the light sources A_1 of the approach lighting system.Since A_1 must be just visible from every point within this cross-section, its luminous intensity must be at least 25 000 cd everywhere within the solid angle subtended by the cross-section at Q_1. Fig. 10*b* shows another cross-section closer to the light source. A lower luminous intensity (in accordance with fig. 9) will be sufficient for directions within the solid angle or cone associated with this cross-section. Of course, the luminous intensity specified for the directions within the cone of fig. 10*a* must be maintained. In fig. 10*c* diagrams *a* and *b* are combined to show the two cones of fig. 10*a* and fig. 10*b* and their intersection with an imaginary surface at right-angles to the extended centre-line of the runway. The resultant curves are "isocandela lines", since they define the limits within which a given luminous intensity, in candela, is required. Accordingly, a complete isocandela diagram fully describing the required light-distribution is obtained by considering a series of cross-sections of the channel of guidance, one after another. Three such diagrams, referring to lights located at three different points are shown in *fig. 11*.

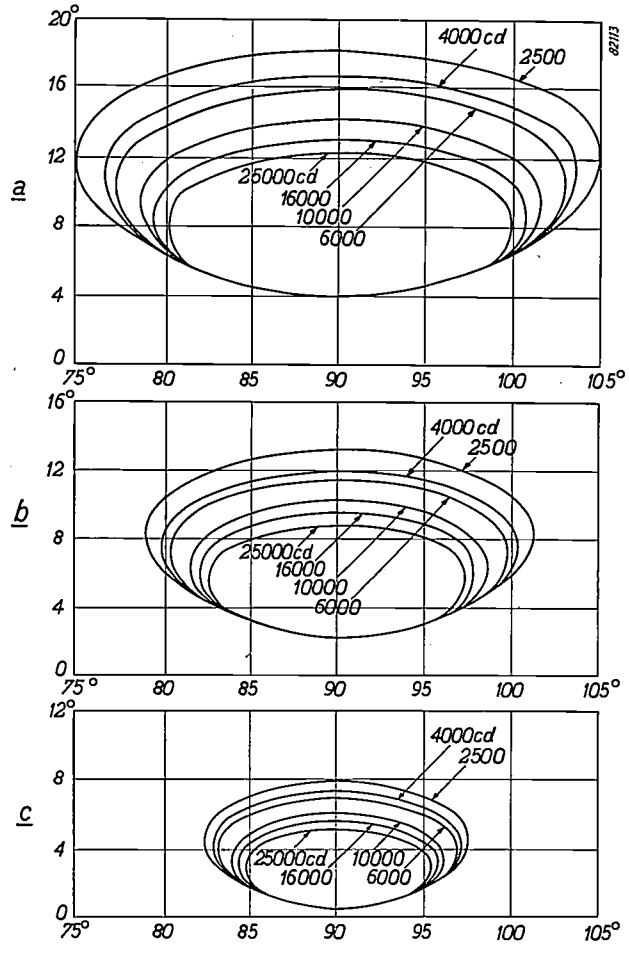


Fig. 11. Diagrams showing the required light-distribution (isocandela diagrams) of *a*) initial approach light, *b*) a threshold light, and *c*) a runway light 450 metres (1500 ft) beyond the threshold.

Fig. 10, shows the way in which an isocandela diagram for a light on the centre-line of the system is constructed. The light-distribution of a light displaced laterally with respect to the centre-line, e.g. on a cross-bar, will be virtually the same, but in this case the axis of the beam should be slightly toed-in instead of running parallel to the centre-line. Space does not permit us to amplify this brief explanation. A more precise description of the calculation of the required light-distributions may be found elsewhere [5]).

Given a light having the exact distribution required, it will be just visible to the pilot of an aircraft entering near the "ceiling" of the visual guidance channel; flying lower, however, he will see the light at a luminous intensity greater than the minimum required to make it visible to him. In fact, the pilot of an aircraft flying nearly at the "floor" of the guidance channel will be exposed to the full luminous intensity corresponding to a visual range of 450 m, until he has nearly reached the particular light source. It will be seen that this involves a risk of dazzle, which is all the greater owing to the fact that it is, of course, impossible to design a light whose luminous intensity is exactly the same *on* the inner isocandela line as *within* it; the luminous intensity will inevitably be greater inside the curve, reaching a maximum at the "axis" of the light-beam.

To avoid dazzle, it is necessary to ensure that the luminous intensity does not exceed a certain maximum value. The *a priori* possibility of satisfying both equally important requirements viz. visibility of the light at the specified range and avoidance of glare at the shortest distance at which the light disappears under the cockpit (angle β in fig. 8) if the pilot follows the least favourable track approach (near the "floor" of the channel), is very much open to question. Fortunately, investigations have shown [6]) that the latitude between the minimum eye-illumination necessary for discerning a light and the maximum eye-illumination tolerable in view of glase, both measured, of course, under the same conditions) is quite large. This latitude may be taken as a factor of 200 at night, and about 1000 by day. The geometrical dimensions of the visual guidance channel and those of the approach

[5]) J. B. de Boer, Calculations on the light distribution of approach and runway lights, Philips Research Reports **6**, 241-250, 1951.

[6]) G. A. W. Rutgers, Visibility of signal lights, Electrotechniek **26**, 36-40, 1948 (In Dutch).

lighting system are such that, given an ideal light-distribution as shown in fig. 11 and a visibility not less than 300 m, the above-mentioned latitude will not be exceeded by the luminous intensity anywhere within the channel.

It will be evident that for a day landing in relatively better visibility, as for a night landing, a lower luminous intensity of all the lamps is not only permissible, but essential for the avoidance of glare. Hence the overall luminous intensity is controlled from a central point by varying the supply current. In fact, one of the methods employed enables the pilot of an approaching aircraft to vary the luminous intensity of the beacon-system to suit his own immediate requirements by means of a radio-operated servo-mechanism.

The supply system of an airfield beacon system will be described in a separate article in this Review.

The practical realization of the calculated light-distributions involves its own special problems. For reasons of economy it will be necessary to obtain the required light-concentration with the aid of fairly simple optical systems; hence a reasonable approximation to the ideal light distributions is the most that can be expected in practice. The principal aims of the designer will be:

1) To reduce the above-mentioned peak luminous-intensity in the beams as far as possible, and so limit glare.
2) To minimize the amount of light radiated outside the prescribed beam, since the light so "spilled" not only constitutes a loss in itself, but, owing to diffusion in fog or mist, increases the overall background luminance and so necessitates a higher level of eye-illumination: i.e. the range associated with a particular luminous intensity is reduced. For the same reason, it is necessary to ensure that the amount of light radiated upwards is no larger than strictly necessary.

Design of the different lights

The three types of light constituting the beacon system of a runway, i.e. approach lights, threshold lights and the actual runway lights (see fig. 3), necessarily differ considerably in design, firstly by reason of the difference in the required light-distribution (see fig. 11) and secondly — a very important factor in the design — because they vary as regards the extent to which they may be raised above ground level. The initial approach lights, stationed at a point which approaching aircraft will clear by a hundred feet or so, may safely be raised some feet above the ground (the "rough"). On the other hand, the lights at the threshold of the runway, where aircraft frequently touch down, should preferably be virtually or entirely "flush". The lights at the edges of the runway may protrude above the rough, but, since there is a possibility of the aircraft colliding with them, must then be so designed as to ensure that such a collision cannot damage the aircraft.

Another important feature of design, common to all the lights, is adjustability as regards azimuth and elevation (aiming adjustments). This is necessary as a means of ensuring correct coverage of the visual guidance channel with the particular light-distribution provided (fig. 7). Such aiming necessitates special features of the lamps as well as certain auxiliary equipment, some examples of which will be described briefly at the end of this article.

Incandescent lamps are employed for all the lights considered, since the concentrated beam of light indicated by the diagrams in fig. 11 would be virtually unattainable from relatively large light sources of low luminance, such as fluorescent, or sodium lamps. High-pressure mercury vapour lamps would be suitable by virtue of their very high luminance-level, but are less simple to operate (e.g. it is more difficult to vary their light output) than incandescent lamps; hence they have not been employed hitherto for this purpose.

The various components of the system will now be described more fully, viz. the approach lights, the runway lights and finally intermediate the threshold lights.

Approach lights

The approach lights require a greater vertical and lateral beam-spread than either of the other two classes of light referred to in the above (fig. 11). However, they may be mounted entirely elevated, and need not be unduly restricted as to size and weight.

A special incandescent lamp (see *fig. 12*) has been designed to furnish the required light-distribution. The rear bulb-wall of this is silvered on the inside and has the shape of a paraboloid of revolution, provided with vertical ridges, whose form governs the horizontal spread of the beam emitted by the lamp. *Fig. 13* shows the isocandela diagram of this lamp, which is fitted with a 400 watt low-voltage filament (24 V). Comparing this diagram with fig. 11*a*, we see that the lamp amply satisfies the requirements imposed. Although it is not within the scope of this article to enter into the reasons for choosing such a low operating voltage (or, rather, a relatively strong current) which necessitates a separate transformer for each lamp, it may be noted that a low operating voltage is consistent with a thick, compact filament, which facilitates the concentration of the light.

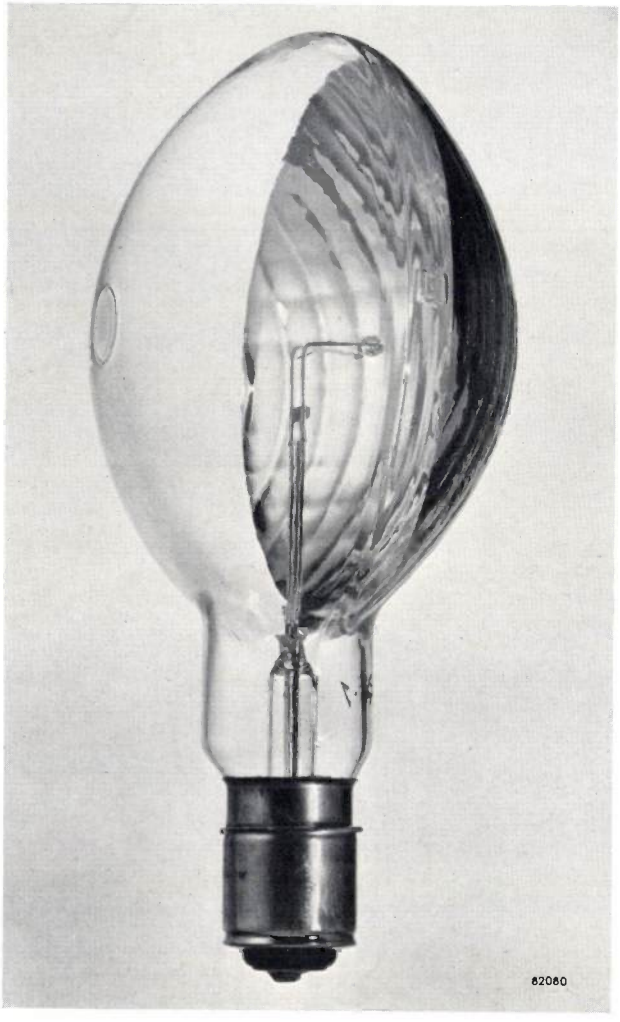


Fig. 12. Special 400 W, 24 V incandescent lamp for use in approach lights. Part of the bulb has the form of a paraboloid of revolution, silvered on the inside and vertically ridged. Height of lamp: 11 inches.

The lamp is fitted with a so-called "prefocus" cap (type P40) having two dissimilar sector-shaped lugs by virtue of which the lamp can be placed in only one precisely defined position in the lamp holder. The orientation of the beam relative to the cap is accurately adjustedand fixed during the manufacture of the lamp, so that correct beam-alignment is assured whenever a new lamp is inserted in one of the lamp-holders which are aimed once and for all when they are installed.

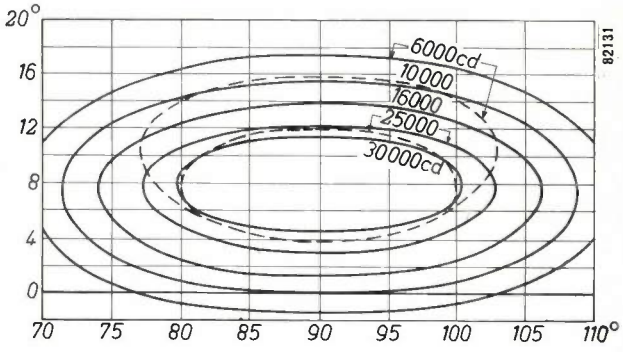


Fig. 13. Isocandela diagram of the 400 W lamp shown in fig. 12. Two curves from fig. 11*a*, drawn as dotted lines, are also included in this diagram.

The bulbs of these lamps are made of hard glass and therefore require no protection from the weather. The principal advantage of such special lamps is that they require no fitting other than an adjustable lamp-holder suitable for outdoor use: hence all difficulties arising from the infiltration of dust and dirt into optical systems and the weathering of mirrors are avoided.

The cross-bars of the approach-lighting system are built-up by arranging several lamps in a row. On one of the Schiphol runways this is accomplished by the method illustrated in *fig. 14*. Here, such a bar comprises a number of units of 10 lamps mounted side-by-side on a bracket, the advantage being that all 10 lamps can be adjusted simultaneously to

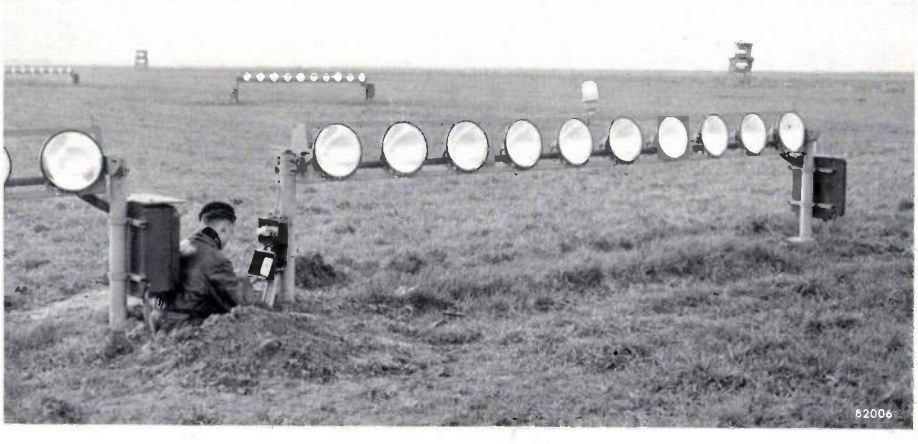


Fig. 14. Bracket carrying 10 approach lights (of a type other than that shown in fig. 12). Several of these brackets make up one "cross-bar" in the approach lighting system of runway 23 at Schiphol.

the correct beam-elevation merely by turning the bracket. A small omni-directional light (that is, an ordinary incandescent lamp in a barrel-shaped Fresnel lens of pressed glass) is also provided at the centre of each bracket. Provided that the visibility is not unduly bad, these omni-directional lights enable the pilots of aircraft circling the airfield to maintain their sense of direction relative to the runway.

Runway lights

As we have already seen, the runway lights may protrude above the level of the rough, if so designed as to cause no damage to any aircraft happening to collide with them. Such raised lights have a number

Fig. 15. Elevated runway light, as employed on a number of airfields in the Netherlands. Two special lamps, similar to that shown in fig. 12 but rated at only 100 W, are suspended back to back to shine in the two landing directions of the runway. The cover contains an omni-directional light.

of advanatages over "flush" lights. They not only enable the requirements as to luminous intensity and beam-spread to be satisfied far more easily, but (if suitably designed) are less likely to be covered by snow in winter.

Raised runway lights are employed on several airfields in the Netherlands. *Fig. 15* shows the construction of such a light. It includes two adjustable lamp-holders from which two special lamps are suspended. These lamps have bulbs similar to that shown in fig. 12 (though slightly smaller), and the same type of prefocussing lamp cap, but the filament-power is only 100 W. The light-distribution in the beam of each individual lamp (*fig. 16*) agrees reasonably with the requirement as defined in fig. 11*c*.

The two lamps throw their beams in opposite directions, that is, one in each direction from which aircraft may land on the runway. Like the approach unit, the runway light also includes a small omni-directional lamp (35 W).

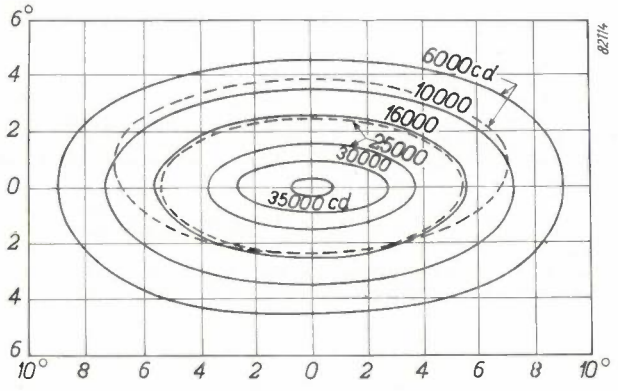


Fig. 16. Isocandela diagram of the 100 W lamp shown in fig. 15, including two curves (dotted lines) from fig. 11*c*.

These runway lights are provided with so-called breakable coupling, that is, specially weakened joints in the supports at a point just above ground level, which break readily if the fitting is struck by an aircraft. Within these mechanical joints are breakable circuit connections, i.e. plug-socket connections each with a rubber socket in the lead between the lamps and the buried transformers.

Fig. 17 shows a still simpler solution to the problem, involving the use of separate lights for the

Fig. 17. The same lamp as that shown in fig. 15, mounted in a single adjustable fitting to give a very simple runway light.

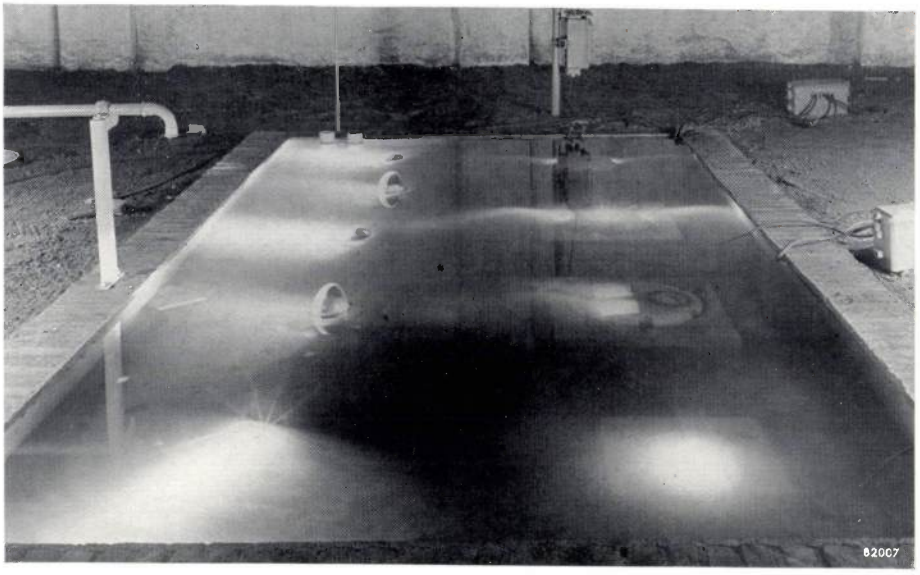


Fig. 18. Testing tank for outdoor lighting fittings. Flush (or semi-flush) lights are mounted as in practice in the tank. The tank is periodically flooded with water and emptied. Five lamps of the type shown in fig. 17 can be seen immersed or semi-immersed on the left-hand side of the picture.

two landing-directions. Here, the special lamp already described is fitted upright in a lamp-holder fixed either in the runway itself or in a separate block of concrete, so that it does not protrude above the level of the runway. The bottom of the holder is a ring, resting on the concrete and secured to it by two plates. These plates can be loosened so as to allow the holder to be turned on its axis. Two screws passing through lugs on the holder also enable the latter to be tilted to some extent about a horizontal axis. If struck by a passing aircraft, the hard glass bulb of such a lamp will splinter without damaging the aircraft; on the other hand, experience has shown that the bulb, with a wall $1^1/_2$ mm thick, is strong enough to ensure that it will not be damaged by stones scattered low over the ground by the aircraft using the runway.

This light, being of a type intermediate between raised and flush lights, involves the problem common to all flush lights, that is, the protection of the fitting from water seeping through the ground; hence a rubber socket is fitted round the lamp cap to provide a water-tight seal when once the lamp has been inserted. Below the clamping nut screwed to the top of the lamp holder is a rubber ring, which when the nut is tightened, presses against the rubber socket and against the inner wall of the lamp holder The effectiveness of the seal is tested in an installation specially designed to test outdoor lighting fittings. Part of this installation is shown in *fig. 18*, from which it will be seen that in order to simulate practical conditions the lights on test are mounted in a tank which is periodically flooded with water.

In case the replacement of lamps broken by the wheels of aircraft is felt to be an objection, but cleaning the lights in the event of a snow-fall presents no difficulty, flush runway lights will be preferred. This system is particularly suitable when the lights, in order to obtain better height indication during the last phase of the landing, are to be arranged in two parallel rows relatively close together, say, about 150 ft apart, instead of along the edges of the paved runway, which may be as much as 200 ft wide. The probability of collision is

Fig. 19. Special 35-W lamp employed in a flush runway light. Part of the bulb, in the form of an ellipsoid of revolution, is silvered on the inside. Height of lamp: $2^1/_2$″.

then too great to permit of the use of elevated lights.

To ensure that flush lights will neither damage, nor be damaged by aircraft running over them, the mounting must be so designed that the lights do not protrude more than a few inches above the level of the runway. Philips have developed a special lamp enabling the required light beam to be obtained despite the above-mentioned limitation and without employing an expensive optical system (an important condition having regard to the fact that a large number of lights is required for each runway). This lamp is shown in *fig. 19*. One side (the back) of the bulb is an ellipsoid of revolution; the filament of the lamp is at one focus of this ellipsoid, and the image of the filament, whichs acts as the light source proper for the optical system, is at the other focus, outside the bulb. The optical system is a rectangular section cut from an aspherical lens of pressed glass (*fig. 20*). The "image" of the filament in the second focus of the ellipsoid really comprises a series of images projected one on top of the other by the different zones of the bulb-mirror and therefore differently magnified. This prevents any sharp definition of the filament structure either in the "image" or in the cross-section of the light beam, and furnishes a large image consistent with the required beam-spread.

Two systems of this type, fixed to adjustable brackets, are mounted back-to-back in the cover of a closed cast-iron housing, which is so embedded in the runway as to protrude less than 3″ above the latter; see *fig. 21*. This housing allows in a simple way the insertion of colour filters in the beam in order to code the beacon system. For further details see the captions of fig. 20 and fig. 21.

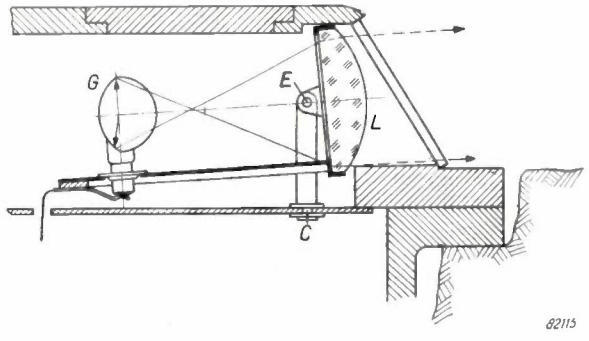


Fig. 20. Optical system of a flush runway light fitted with the lamp (*G*) shown in fig. 19. By virtue of the ellipsoidal mirror in the bulb, the light emitted by the filament is concentrated within a solid angle of nearly 2π, thus giving the required beam despite the fact that the height of the (square) lens (*L*) is only 55 mm. This runway light protrudes less than 3″ above the level of the "rough". To procure a beam of the same width and uniformity, with the same optical efficiency, by a more conventional optical system (e.g. that of the approach light, fig. 12), it would be necessary to employ a much larger system. The elevation of the beam is varied by tilting the optical system and the lamp on a horizontal spindle (*E*). Azimuth adjustment is by rotation about the vertical pin (*C*).

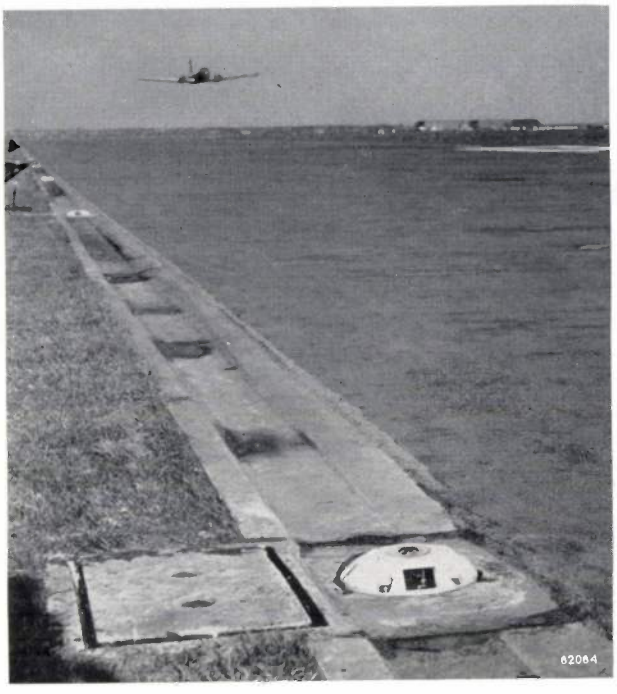


Fig. 21. Flush runway light on runway 05/23 at Schiphol. The light is housed in a cast-iron case, and contains two systems of the type shown in fig. 20, mounted back to back, and an omni-directional top light. To make the housing watertight, a rubber ring is inserted in a circular groove in the cover, close to the securing bolts. The lower part of the housing accomodates three supply transformers.

Threshold lights

When passing the threshold of a runway, aircraft in process of landing are still travelling at high speed; therefore, even the relatively slight elevation of 3 inches above the "rough" is not always considered permissible for lights situated at this point. Shallower lights of the type as that shown in fig. 21 have been designed for this purpose, but require a very special optical system to produce the required luminous intensity. The difficulty is enhanced by the fact that, according to the international recommendations, threshold lights should be green. To obtain green light with incandescent lamps it is necessary to employ a colour filter whose transmission can be only about 15%, so that in fact the luminous intensity of the beam before it passes through the filter must be in the region of 170,000 candelas.

Another solution to the problem, more attractive from the point of view of illumination engineering, is provided by what is known as a grid threshold light; see *fig. 22*. In this, the light supplying the required beam is below the surface of the concrete runway and protected by a steel coverplate. The light beam, projected upwards at a small angle, emerges through a slot cut in the runway in front of the light. Vertical plates of steel, strong enough to give adequate support to the wheel of

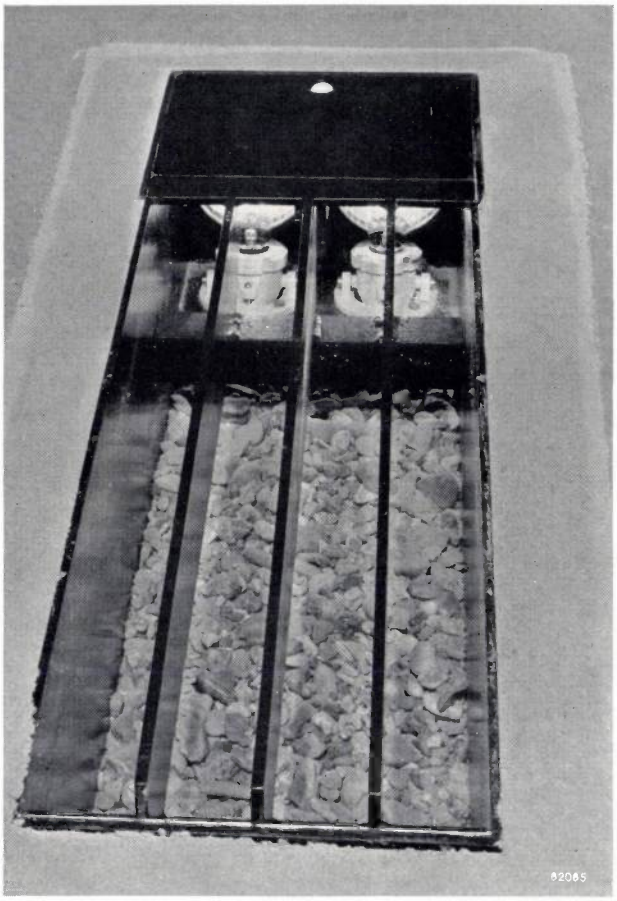


Fig. 22. Grid-type threshold light, fitted with two lamps of the type shown in fig. 15 and fig. 17.

any aircraft passing over them, are fixed in this slot. In the type of grid light illustrated, an internally silvered 100 watt lamp of the type shown in fig. 15 and fig. 17 is employed. Since it is possible to mount this lamp just below the cover-plate, the slot in the runway heed not be unduly long and the installation costs of the whole light can be kept within reasonable limits. Of course, the grid plates cause a certain loss of light. A spacing of these plates which is acceptable from the point of view of the width of aircraft wheels may well involve a light-loss of 20% along the axis of the beam, and of 50% in a direction deviating 5° from this axis. Hence each light should comprise several lamps arranged side-by-side to emit parallel beams, like the two lamps shown in fig. 22.

Aiming the beams

The required aiming of each individual light beam in a particular beacon system can be deduced direct from the diagram in fig. 10*a* ;the axis of the beam should pass through the centre of a cross-section of the visual guidance channel taken at a distance $r = 1500$ feet from the particular light (i.e. the required visual range) [7]). *Fig. 23* shows the required angle of elevation β, deduced from fig. 10*a*, plotted against the position of the light. The required azimuthal angle, α, depends entirely on the lateral distance b between the light and the centre-line of the runway ($\alpha = \tan^{-1} b/r$).

The means of azimuth and elevation adjustment provided in the different lights have already been referred to above. Two of the aiming devices designed to enable the lights to be adjusted quickly and conveniently to the required angles α and β will now be described.

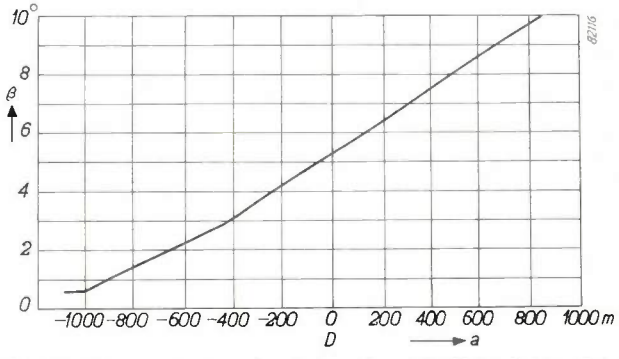


Fig. 23. The required angle of elevation (β) of the beam axis of individual lights in the approach and runway lighting system, plotted against the distance (a) between light and threshold.

Fig. 24 shows the device employed to aim the runway light illustrated in fig. 17. The device includes a prefocusing lamp-cap similar to that of the lamp to be fitted in this runway light. A moveable sight attached to the device can be adjusted to the required azimuthal angle α relative to the "beam-direction" of the prefocusing cap. The cap of the device is then inserted in the lampholder of the particular light, and this holder is rotated on its vertical axis until the sight is in line with the next light

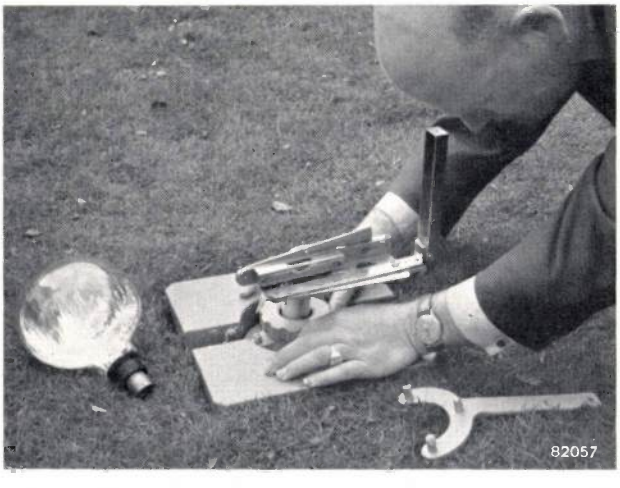


Fig. 24. Aiming device for runway lights of the type shown in fig. 17. The lamps used in such lights, being fitted with a prefocusing cap, will stand at an accurately pre-determined position when inserted in a lampholder. In view of this fact, the *lampholder* is aimed by means of this apparatus. The azimuthal deviation of the sight relative to the beam-direction the lamp will have can be varied. A prism is mounted in the sight to facilitate the adjustments by enabling the user to look down into the instrument.

[7]) This principle is not always adhered to. In some systems of approach lighting, all the beam-axes are in vertical planes parallel to the centre-line of the runway, or even slightly toed-out. Such aiming necessitates greater beam-spread.

in the row (that is, parallel to the landing-direction); a holder so aligned will give the correct azimuthal angle to the beam of the particular lamp inserted in it. The instrument also includes a spirit-level, so mounted that it can be tilted along a graduated arc parallel to the "beam direction" of the lamp cap. If the spirit-level be tilted to the desired angle β and the lampholder then tilted until the spirit-level is again horizontal, the light will assume the correct angle of elevation.

The apparatus shown in *fig. 25* is employed to aim flush runway lights of the type shown in fig. 12. It consists of a large, square-cut lens, mounted on a frame, a peep-sight parallel to the centre-line of lens and frame, and a screen, calibrated in degrees of azimuth and elevation, at right angles to this centre-line. The lens is placed immediately in front of the beam-emitting aperture of the runway light; the frame is levelled with the aid of two spirit-levels, and its centre-line is lined-up parallel to the runway by sighting on the next light in the row or on a marker post placed at the edge of the runway. The bright spot of light produced by the lens as an image of the light-beam on the screen then indicates the azimuth and elevation of the beam. With the aid of the adjusting screws provided in the light, the optical system forming the beam is then rotated about horizontal and vertical axes until the light-spot strikes the correct point on the screen.

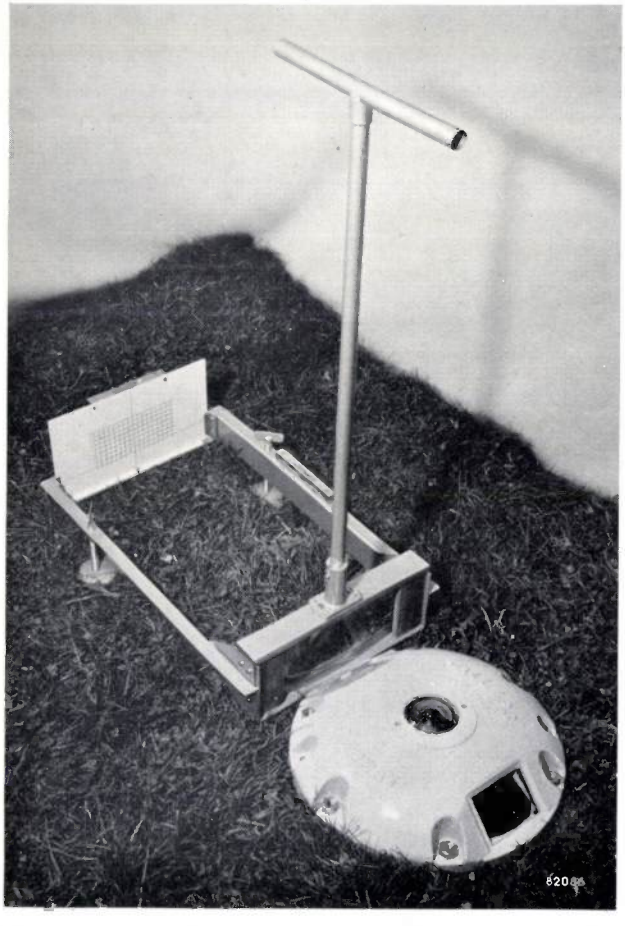


Fig. 25. Aiming apparatus for the flush runway light shown in fig. 21.

It may be worth mentioning, in conclusion, that the possibilities of further developments in landing techniques for aircraft are by no means exhausted. The present situation, that is, that an airfield must close down temporarily when the visibility drops below $V = 1000$ feet, will not be tolerated indefinitely. However, the solution to this problem lies less in increasing the luminous intensity of the beacons — as pointed out, this would be very expensive — than in the development of automatic control systems for the landing of aircraft.

Even when such a system is provided, the pilot will be loath to dispense with visual guidance entirely; however it will then be possible to reduce the altitude at present specified as the minimum for effective visual guidance. With the smaller visual guidance range then permissible, the present beacon system would be adequate in conditions of worse visibility. Indeed, it may probably be predicted that as far as the lights are concerned a point of development has now been reached where further major improvements are unlikely.

Summary. Apart from guidance by radar or radio beacons, which can home an aircraft to within about 3000 ft of the runway-threshold, the pilot coming in to land at night or in fog requires visual guidance to enable him to land safely. Such visual guidance is supplied by a system of beacons comprising approach lights, threshold lights (at the start of the paved runway) and runway lights. According to the recommendations of the I.A.T.A. and I.C.A.O., the approach lights should be arranged along the centre-line of the runway and along one or more transverse lines on this centre-line, and, given a visibility of only 1000 ft, all the lights should be visible, by night as well as in daytime fog, from a distance of at least 1500 ft in all directions within a certain "visual guidance channel". In accordance with these recommendations, it is calculated that the luminous intensity of all the lights should be at least 25,000 candelas, and the required light-distribution (isocandela diagram) of each light is determined. To satisfy these requirements, Philips have designed a series of special incandescent lamps and fittings. Lights belonging to each of the three classes referred to above are described in this article. The actual runway lights may be either "elevated" or "flush" the former type being so designed that if struck by an aircraft they will break without damage to the plane, whilst in the latter type special attention is given to water-tightness of the fitting. Threshold lights should preferably be entirely flush and should be green — a very stringent combination of requirements; one of the most suitable designs is the so-called grid-type threshold light. All the lights are provided with a means of varying the elevation and (where necessary) the azimuth of the beam: the beams of the lights installed on an airfield are aimed with the aid of special equipment.

FEEDBACK AMPLIFIERS FOR CARRIER TELEPHONE SYSTEMS

by J. te WINKEL.

621.375.232:621.395.44

In carrier telephony systems used for long-distance communications, techniques have been developed whish permit a large number of speech channels — up to some hundreds — to be carried on each line. The very wide frequency bands which then have to be transmitted by the cable introduce some interesting problems regarding the design of the line amplifiers that have to be included at regular intervals in the cable circuit.

Introduction

To cope with the vastly increased long-distance telephone traffic it is expedient for economic reasons to transmit a number of channels simultaneously over one pair of conductors. For this purpose each audio signal, usually restricted to frequencies between 300 and 3400 c/s, is modulated onto a different H.F carrier frequency, the result being effectively a displacement of the frequency band. The number of channels that can thus be transmitted over one conductor pair when these bands are placed side by side in the frequency spectrum is determined by the total bandwith that can be transmitted by the circuit. In modern telephone cables, containing a number of conductor pairs and termed symmetrical cables (as distinguished from the coaxial cables to be mentioned later) the number of channels operated over each conductor pair is nowadays as much as 48 or 60. The limit to this number of channels, and hence to the width of the total frequency band to be transmitted, is entirely determined by the properties of the cable. Crosstalk is fundamentally the limiting factor, since this increases with increasing frequency and thus also when the number of channels is increased [1]). Special means, which will be discussed later, are always used to compensate this crosstalk as much as possible. The remaining amount, however, still constitutes an upper limit to the frequency. The designer of amplifiers for use in conjunction with this type of cable should therefore consider the total bandwidth as the primary design parameter. If it is desired to carry even more channels over one pair of conductors by means of conventional single-sideband modulation, other types of cable have to be employed. The coaxial cable is particularly suitable for this purpose. In this type of cable, the inner conductor is completely surrounded by the other, the tubular outer conductor. Owing to the skin effect in the outer conductor hardly any external field is created (provided the frequency is not too low), so that no crosstalk occurs. In this respect there is no limit to the number of channels and thus to the total bandwidth. Since, however, the attenuation increases with the frequency also in this type of cable, the degree of amplification must become greater the wider the frequency band, or the amplifiers have to be spaced at shorter intervals. This constitutes a fundamental limitation on the possible number of channels. In practice bands up to 4 to 8 Mc/s wide, with room for 1000 to 1800 channels, can be used. This large number of available channels is in itself an advantage of the coaxial cable. On the other hand it gives rise to several particular difficulties, so that in practice both coaxial and symmetrical cables have found their own fields of application [2]). The amplifiers which have to be incorporated in the transmission line at regular intervals, known as line amplifiers, are the subject of this article. They have to meet a number of specific requirements. The nature of these requirements, and under what conditions they can be satisfied will be explained below. Finally certain types will be examined more closely.

Design criteria for line amplifiers

The various properties to be considered when designing a line amplifier for carrier telephony are as follows:

1) Large absolute bandwith
2) Large relative bandwidth
3) Very constant amplification (high stability).

[1]) See for example, G. H. Bast, D. Goedhart and J. F. Schouten, A 48-channel carrier telephone system, Philips tech. Rev. 9, 161, 1947/48.

[2]) For a more detailed discussion of the two systems, see H. N. Hansen and H. Feiner, Coaxial cable as a transmission medium for carrier telephony, Philips tech. Rev. 14, 141-150, 1952/53.

4) Amplification to be independent of the amplitude of the signal (linearity).
5) Sufficient degree of power amplification.
6) Well-defined input and output impedances (i.e. the impedances presented by the amplifier to the line).

These requirements will now be discussed in turn.

Large absolute bandwidth

The absolute bandwidth, i.e. the difference between the highest and the lowest frequency of the signal to be amplified, is larger according to the number of channels used. With symmetrical cables, for the transmission of 48 channels, the band 12-204 kc/s is used, and for 60 channels the band 12-252 kc/s With coaxial cables the permissible bandwidth is far greater and is, as mentioned in the introduction, limited only by the properties of the amplifier. At the present state of amplifier technique, it is possible to design amplifiers for the bands up to 60-4100 kc/s or even 0.3-8 Mc/s.

These bandwidths are of the same order of magnitude as those required for the transmission of a TV signal (5 Mc/s for the 625-line standard). The above-mentioned amplifier for 0.3 to 8 Mc/s has been developed also with a view to this application. It has, therefore, to meet certain additional requirements applicable to TV transmission, but these will not be considered here.

Large relative bandwidth

The relative bandwidth, i.e. the ratio of the highest to the lowest frequency, will always have to be large, because at a given absolute bandwidth it will always be attempted to keep the lowest frequency (and hence also the highest) as low as possible in view of the fact that the attenuation increases with the frequency for both types of cable. For the above-mentioned bands of 0.3-8 Mc/s and 60-4100 kc/s the relative bandwidth amounts to approx. 27:1 and 70:1 respectively.

A large relative bandwidth causes the following difficulties: The amplifier includes a number of components, such as inductances and capacitors, whose values and hence whose dimensions have to be greater as the lower frequency limit becomes lower. The larger their dimensions, the more are all these components liable to give rise to all sorts of secondary effects (e.g. the self-capacitance of a coil, the leakage inductance of a transformer, the earth capacitance of a capacitor, etc.). As a rule these secondary effects either directly or indirectly reduce the amplification at higher frequencies. Clearly this constitutes a limitation on the relative bandwidth. The designer of components can make a useful contribution here by reducing the disturbing effects, for example by using new materials and thus decreasing the dimensions.

High stability

Several factors may cause the amplification to vary. Fluctuation of temperature and supply voltages may influence the properties of several components. This may also occur without external influences, simply by a process of ageing, particularly where valves are concerned.

As regards the amplification necessary in coaxial circuits, strict limitations are imposed upon amplification variations. Now, it has been found in practice that the distance between two repeater stations must never exceed 10 km. If it were possible to make each amplifier constant within 1% (which means an amplification variation of $\pm$ 0.09 dB), then a succession of 100 amplifiers (a transmission distance of 1000 km is by no means exceptional) would result in a variation that could amount to as much as $\pm$ 9 dB. It is true that since the attenuation of the cable varies with the temperature, it is anyhow necessary to provide for a certain compensation — either automatic or not — for the fluctuations in amplification of the whole system of cable and amplifiers. In view of the fact that this compensation introduces complications, however, it is advisable not to compensate more than is strictly necessary, so that fluctuations in amplification should be kept as small as possible.

For amplifiers in transmission systems using symmetrical cables this requirement is less strict, because here the repeaters may be spaced at intervals of 18-25 km, so that a smaller number is required to bridge a given distance.

Amplification independent of the amplitude (linearity)

If the instantaneous output voltage of an amplifier is not a purely linear function of the instantaneous input voltage, this gives rise, in the case of a complex input signal, to harmonics and sum-and-difference frequencies in the output signal. If many channels are amplified at the same time, interference signals will appear roughly uniformly distributed throughout the band, which may be regarded as noise. This is called *intermodulation noise* because it is mainly caused by the interaction of different frequencies. Although the sum and difference frequencies are as a rule of greater importance, it is possible to calculate the intermodulation noise from measurements of the second or higher harmonics using a single variable frequency [3]). From this the quantitative re-

[3]) See R. A. Brockbank and C. A. Wass, J. Inst. El. Engrs. III, 92, 45-56, 1945.

quirements regarding the linearity of the amplifier can be derived.

The intermodulation noise can be reduced by keeping the amplitude of the input signals very small, but this remedy is restricted by the fact that noise is also caused by other factors (thermal noise, valve noise) already present at the amplifier input. This noise is independent of the signal level. It thus sets a certain lower limit to the input voltage if one is to maintain a certain miminum signal-to-noise ratio.

Power amplification

The production of sufficient power does not constitute a problem in itself. There are several types of amplifying valves, up to the large transmitting valves, that are capable of dissipating large powers. Raising the output power, however, requires larger valves as well as larger output transformers and thus gives rise to the disturbing secondary effects mentioned under (2). In practice, moreover, it has been found that there is little point in raising the output power above a certain level, as we shall now demonstrate.

The purpose of the amplifier is obviously to compensate the attenuation occurring in a certain length of line. As an example let us consider a coaxial cable with line amplifiers spaced at intervals of 10 km. Each amplifier has a gain of 50 dB and a maximum output power of 500 mW. If the voltage amplification is to be increased by a factor of $\sqrt{10}$, i.e. 10 dB, then, in view of the fact that the input signal is practically fixed, the output power must become 10 times greater, i.e. 5 W.

This would make the amplifier considerably more complicated and also require a larger power-supply. If all these objections were acceptable, then we would have obtained an amplification of 60 dB. The interval between the repeaters could then be enlarged to

$$\frac{60}{50} \times 10 = 12 \text{ km},$$

which does not make very much difference in the number of amplifiers required. To obtain 20 dB more amplification and a spacing interval of 14 km, an output power of 50 W would be necessary.

Input and output impedance

With amplifiers for symmetrical cables the impedance, presented to the cable by the input and output terminals, must be equal to the characteristic impedance of the cable. This is necessary in order to keep crosstalk at a low level, as may be seen with the aid of *fig. 1*. Lines *1* and *2* represent two adjacent conductor pairs. At the ends marked *B* are the output terminals of two line amplifiers; at *C* are the input terminals of the two next amplifiers. Let us assume that at *A* a crosstalk signal from conductor pair *2* is transferred to conductor pair *1*. This signal will be propagated in both directions. The part going to the right will arrive at *C* together with the original signal on conductor pair *2* irrespective of the position *A* where the signal originated.

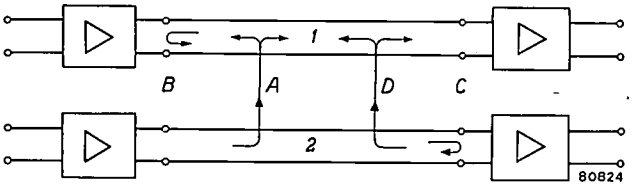


Fig. 1. Two conductor pairs (*1* and *2*) of a symmetrical cable link. At *B* are the output terminals of two amplifiers, at *C* the input terminals of the next two amplifiers. A signal producing crosstalk from *2* to *1* at *A*, will be propagated in *1* in both directions. That going to the right (together with the initial signal) can be compensated at *C*, because it is in phase with the initial signal. The part going to the left must not be reflected at *B*, for the phase difference (which depends on the position of *A*) cannot be compensated. Nor must reflection occur at *C*, since the crosstalk from this reflected signal (e.g. at *D*) cannot be compensated either.

By using additional adjustable coupling elements between the two conductor pairs this crosstalk signal can be fairly well compensated for all positions of *A* (this is known as *balancing* of the conductor pairs). The part of the crosstalk signal going from *A* to the left and being reflected at *B*, however, cannot be compensated at *C*, because it has become out of phase with respect to the direct signal, due to its longer path. In order to keep the crosstalk level as low as possible these reflections should be avoided whenever possible by terminating the conductor pairs at *B* with their characteristic impedances (matching). The termination at *C* should be similarly matched to avoid reflection of the signal itself. This reflected signal can also give rise to crosstalk, as indicated by *D* in the figure. This crosstalk signal also cannot be compensated.

In the case of amplifiers for coaxial cables, where crosstalk becomes negligible at higher frequencies, such matching is, as a rule, not required for telephony transmission (although it is necessary for TV signals, but for different reasons).

Design of the amplifier; negative feedback

One is thus faced with the following problem: a reliable, constant and linear amplification has to be effected over a wide band. A solution for this problem, well-known in amplifying technique, is the application of considerable negative feedback. The subject of negative feedback has been discussed

more than once in this Review [4]), so it will suffice here to recall briefly its principle. *Fig. 2* shows an amplifier A and a feedback circuit B. The gain factor of A is μ. Part of the output voltage V_u is

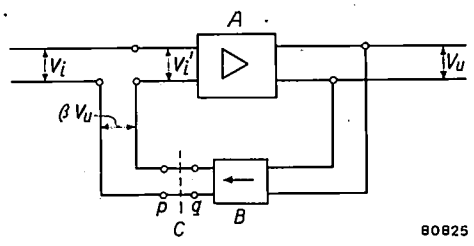


Fig. 2. Amplifier A with feedback circuit B. A fraction βV_u of the output voltage V_u is tapped off by the circuit B and applied in series with the input voltage of A (in opposite phase). The total amplification thus becomes more linear and less dependent on variable quantities.

tapped off to the circuit B, in such a way that a part βV_u is deducted from the input signal V_i. The following relationships then exist between the effective input voltage V_i' and the output voltage V_u of the amplifier A:

$$V_i' = V_i - \beta V_u,$$
$$V_u = \mu V_i'.$$

and hence:

$$V_u = V_i \cdot \frac{1}{\beta} \cdot \frac{\mu\beta}{1+\mu\beta}.$$

If $\mu\beta$ is large compared to 1, the voltage gain becomes approximately equal to $1/\beta$ and, therefore, no longer dependent on μ, i.e. on the valves and other fluctuating or distortion-causing elements in the separate stages of the so called μ-circuit. β on the other hand, is determined by the ratio between impedances or between the numbers of turns on a transformer, and can thus readily be made constant and linear. A more rigorous treatment shows that all variations in amplification caused by fluctuations in valves and other components of the μ-circuit, as well as the non-linearity, are reduced by a factor $1 + \mu\beta$.

The relative change of amplification caused by a change $\Delta\mu$ is, without feedback, $\Delta\mu/\mu$, and with feedback:

$$\frac{\Delta\mu \cdot \frac{\mathrm{d}}{\mathrm{d}\mu}\left(\frac{\mu}{1+\mu\beta}\right)}{\frac{\mu}{1+\mu\beta}},$$

which is

$$\frac{1}{1+\mu\beta} \cdot \frac{\Delta\mu}{\mu}.$$

The change $\Delta\mu$ may have been caused by variations of any component in the amplifier μ-circuit.

[4]) See, for example, C. J. van Loon, Philips tech. Rev. 1, 264-279, 1936; B. D. H. Tellegen, Philips tech. Rev. 2, 289-294, 1937.

It will thus be our aim to make the product $\mu\beta$, termed the feedback factor, as large as possible. The resulting amplification, i.e. the gain between input and output terminals, which, as stated, is nearly equal to $1/\beta$, is determined by the design of the whole transmission system. Hence μ should be given the highest possible value. We find, however, that the large bandwidth imposes certain limitations upon μ, which will be discussed below.

Limitation of the amplification at large bandwidth

For a single amplifying stage, which in simplest form may be considered to be a valve with slope S and anode impedance Z, the amplification amounts to SZ (so long as Z is small compared to the internal resistance of the valve).

One component of Z is due to the total stray capaci-

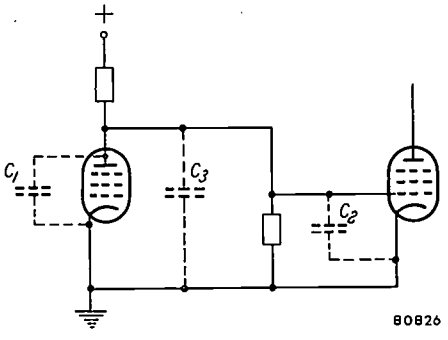


Fig. 3. Part of an amplifier. For determining the anode impedance, the following stray capacitances have to be taken into account:
C_1: the anode-cathode capacitance of the valve,
C_2: the grid-cathode capacitance of the next valve,
C_3: the wiring capacitance.

tance $C_p = C_1 + C_2 + C_3$ (*fig. 3*), i.e. the sum of the anode-cathode and grid-cathode capacitances and the wiring capacitance. It is not possible to connect anything in series with this capacitance, but a network of impedances may be arranged in parallel with it. This impedance should be chosen so as to obtain the greatest possible amplification over the

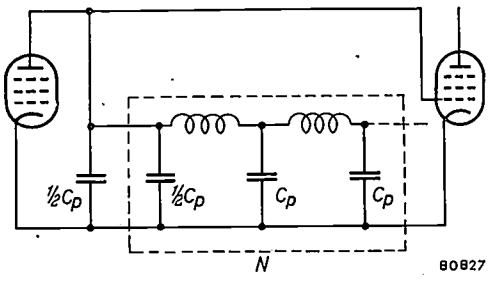


Fig. 4. Circuit for obtaining the optimum amplification for a given bandwidth. N is a low-pass filter with a cut-off at the highest frequency to be amplified. The first element of N is formed by half the stray capacitance $C_p/2$.

given bandwith. According to Bode [5]) this should be done as follows: The stray capacitance is imagined to be divided into two equal parts, one part forming the capacitance of the first section of an ideal low-pass filter N (see *fig. 4*), the cut-off frequency of which is to correspond to the highest frequency f_m to be transmitted. It is known from the theory of filters that the admittance Y_p of such a filter at a frequency f is given by the general formula:

$$Y_p = Y_0 \sqrt{1 - \frac{f^2}{f^2_m}}, \quad \ldots . \quad (1)$$

where Y_0 represents the D.C. admittance.

For the case of this filter $Y_0 = \pi f_m C_p$.

The total admittance Y in the anode circuit is therefore:

$$Y = Y_p + j\pi f C_p = \pi f_m C_p \sqrt{1 - \frac{f^2}{f^2_m}} + j\pi f C_p. \quad (2)$$

For $f \leqq f_m$, the modulus of this is:

$$|Y| = \pi f_m C_p.$$

The amplification μ then becomes:

$$\mu = S|Z| = \frac{S}{|Y|} = \frac{S}{\pi f_m C_p},$$

and is thus constant for all frequencies between 0 and f_m. We see that the absolute values of the amplification and of the impedance become uniform throughout the whole range. This impedance is equal to double the impedance of the stray capacitance at the cut-off frequency. This can be readily understood if we consider that the impedance of the filter is infinitely high at the cut-off frequency (cf. equation 1). Half the stray capacitance has, as it were, been removed from the circuit.

An ideal filter contains an infinite number of elements: in practice this has to be approximated to by a distributed network of a finite number of elements. With such a distributed network it is possible to attain 80-90% of the theoretical maximum amplification. When this is taken into account, it is found that the amplification, according to the above considerations, is determined by the highest frequency f_m and by the ratio S/C_p, and the latter cannot possibly exceed $S/(C_1+C_2)$ (see fig. 3). The ratio $S/(C_1+C_2)$, which is also an important quantity in other amplifying circuits, is termed the figure of merit of the valve in question. For the sake of simplicity this will be denoted henceforth S/C.

Once we have selected a tube with the best possible figure of merit, the maximum attainable amplification per stage is more or less settled. We should then employ as many stages in series as are necessary for the total amplification μ to reach the required high value. If more than two stages are necessary, which is very probable where coaxial amplifiers are concerned, certain complications may arise, due to the fact that the amplifier begins to oscillate spontaneously. This effect can be explained as follows. An amplifier with negative feedback possesses a second amplification circuit, the $\mu\beta$-circuit, which we can imagine if the feedback circuit is interrupted, say, at C in fig.2, thus creating the terminal pairs p and q. It is assumed that at p a slight potential difference v occurs, caused e.g. by the stray pick up of an interfering signal, whether the input voltage V_i at that particular moment is zero or not. Owing to the feedback, a voltage $\mu\beta v$ then prevails at q. This circuit has the inherent property (on which the whole idea of negative feedback is based) that the voltage at q is in anti-phase with the voltage at p. It is clear, therefore, that if at some frequency the output voltage is given an additional phase shift of 180° and thus be comes in-phase with the input voltage the whole system will start oscillating.

The precise conditions under which oscillation will occur are given by Nyquist's criterion [6]), which we shall recall briefly here. The ratio of the output to the input voltage in the feedback circuit of an amplifier is, as a rule, a complex quantity, which can be plotted as a point in the complex plane. This quantity is a function of the frequency, and as the frequency is varied the point in the complex plane describes a curve, generally a closed curve, since the amplification both at very high and at very low frequencies approaches zero.

Nyquist showed that a negative feedback amplifier will not oscillate unless this curve encircles the point —1 on the real axis [7]). In practice a more stringent criterion must be maintained, which is in effect the condition stated above, viz. the phase shift of the output must be less than 180° when the amplification is greater than unity.

Considering the phase shift between input and output voltage per stage, the risk of oscillation is considerable. At low frequencies the anode impedance will behave very nearly like a resistance. At frequencies equal to or higher than f_m the maximum frequency to be transmitted, equation (2) shows that the anode impedance is purely imaginary, so that at frequencies equal to or higher than f_m the phase shift is 90° more than at low frequencies. Using two valves, we would thus just reach the condition for oscillation viz. 180° extra phase shift in the $\mu\beta$-circuit. With more than two valves the phase difference might be even greater and, with

[5]) H. W. Bode, Network analysis and feedback amplifier design, Van Nostrand, New York, 1945.

[6]) H. Nyquist, Bell Syst. tech. J. **11**, 125-147, 1932.

[7]) See also the article by B. D. H. Tellegen referred to in [4]).

an amplification in the circuit which could be greater than unity, the amplifier would promptly start oscillating.

Here we encounter a new problem: how to obtain in this second amplifier circuit — i.e. from p to q in fig. 2 — not only the greatest possible amplification $\mu\beta$ over a given frequency range, but at the same time a phase shift which does not exceed 180° for all those frequencies for which $\mu\beta$ is greater than unity. This problem too has been studied theoretically by Bode. His study shows that for each network, irrespective of its circuitry or number of valves there must exist a certain relationship between the amplification and the phase shift. To a first approximation the phase shift is proportional to the ratio of the change of the logarithm of the amplification to that of the logarithm of the frequency, i.e.

$$\varphi = c\,\frac{\mathrm{d}\log\mu\beta}{\mathrm{d}\log f}. \quad \ldots\ldots (3)$$

The exact formula may be found in the book referred to in [5]): the above formula (3) applies if $\mathrm{d}\log\mu\beta/\mathrm{d}\log f$, i.e. φ, varies relatively slowly with the frequency.

In order to avoid oscillation of the amplifier, φ should be less than 180° for all values of $\mu\beta$ greater than unity. To give a concrete example, we assume φ to be equal to a constant φ_0 for all frequencies between f_m, the highest frequency transmitted, at which $\mu\beta$ has nearly the maximum value, and the frequency f_0, at which $\mu\beta$ is unity. We can then integrate equation (3) between f_0 and f_m and find:

$$\mu\beta_{max} = \left(\frac{f_0}{f_m}\right)^{-\frac{\varphi_0}{c}}.$$

An exact calculation shows that for $\varphi_0 = -180°$ and for three stages of amplification,

$$\mu\beta_{max} = \left(\frac{4}{3}\,\frac{f_0}{f_m}\right)^2.$$

The parameter f_m may be considered as given. The maximum value of $\mu\beta$, i.e. the maximum feedback factor between $f = 0$ and $f = f_m$, is thus known if we know the frequency f_0. This frequency, for which $\mu\beta$ has dropped as low as unity, is so high that the anode impedances of the various stages are nearly equal to the stray capacitances.

As a concrete example, let us assume that for this frequency $\beta = 1$, then also $\mu = 1$ and the amplification per stage is 1, so that:

$$\frac{S}{2\pi f_0 C_p} = 1, \text{ or } f_0 = \frac{1}{2\pi}\,\frac{S}{C_p}.$$

Since the value of β is generally somewhat smaller than 1, dependent on the circuit, f_0 will assume a slightly lower value, but the essential fact remains that f_0 bears a direct relationship to S/C ,the figure of merit of the type of tube in question.

We now arrive at the following result. For a given type of valve with a given S/C value, the maximum attainable feedback $\mu\beta$ is determined by the highest frequency to be transmitted f_m, $\mu\beta$ decreasing as f_m increases.

We see that enlarging the bandwidth is achieved at the expense of the feedback and hence at the expense of the linearity and of the stability of amplification. The better the quality of the valves, expressed as the figure of merit, the larger bandwidth and/or feedback is permissible.

Having dealt with the fundamentals of these amplifiers we shall now consider some practical designs.

Practical designs

Amplifier for symmetrical cables (12-204 kc/s)

In the foregoing we have shown that the order of magnitude of both input and output signal is more or less determined. The input level has to lie a certain amount above the total noise level and the output power should be such that it can be delivered by a tube of normal dimensions. The magnitude of the resulting amplification $1/\beta$ is therefore approximately determined. If, furthermore, the feedback factor $\mu\beta$ is specified, μ can be found. μ is the amplification without feedback and hence the product of all amplifications per stage. Its numerical value is such that two amplifying stages, having equal amplification for all frequencies of the band, are not sufficient. This requirement of a uniform amplification for all frequencies, however, is not essential; the fact that higher frequencies are attenuated to a greater extent by the cable than the lower ones makes it permissible for $1/\beta$ to decrease at lower frequencies. If the feedback is to be kept constant, μ also has to decrease in proportion to the frequency, in fact, in a ratio that is inversely proportional to β This can be realized by means of complementary networks in the μ and β circuits [8]).

Owing to the fact that μ decreases with decreasing frequency, it is possible to reach higher values of μ at higher frequencies than could be achieved if μ were uniform throughout the frequency band, and in this way two amplifying stages will be sufficient. This means a considerable saving as compared to a 3-stage amplifier. In addition, the requirement of

[8]) See H. van de Weg, The equalization of telephone cables, Philips tech. Rev. 7, 184-191, 1942.

stability can be better satisfied, as demonstrated above.

Fig. 5 shows the rising frequency characteristic within the band 12-204 kc/s. At 204 kc/s the amplification is 68 dB. The feedback factor $\mu\beta$ is 32. With present symmetrical cables and a band of 12-204 kc/s these repeaters can be placed at intervals of 25 km.

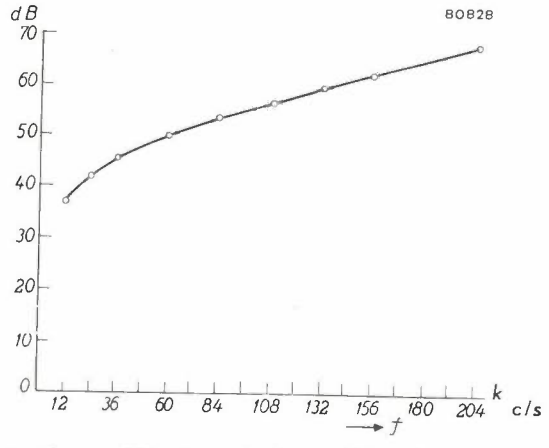


Fig. 5 The amplification of the amplifier for symmetrical cables as a function of frequency. In order to use only two amplifying stages, the amplification is made to rise with the frequency.

The amplifier under consideration uses a combination of current and voltage feedback in order to match its output impedance to that of the cable and thus to keep the cross-talk level low. Due to the voltage feedback (as shown in fig. 2) the terminal voltage will be kept nearly constant by the amplifier during variations of the load resistance. The internal resistance measured at the output terminals is low in this case. In the case of current feedback a voltage is returned to the input terminals that is proportional to the output current. As a result of this, variations in the output current are suppressed, so that the internal resistance is high. Without any calculations, it will be seen that by combining current and voltage feedback any desired resistance may be obtained. (Matching could also be effected with only voltage feedback or only current feedback by introducing an additional parallel or series resistor, but this would always result in a loss of output power.)

Fig. 6 shows a simplified diagram of the amplifier with the combined feedback. Current feedback is established across resistor R and voltage feedback through the secondary winding S of the output

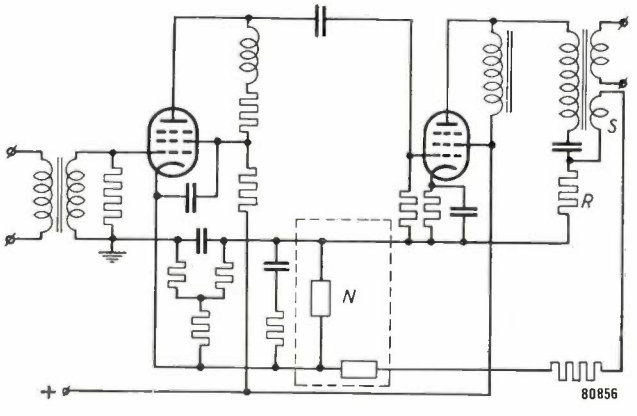


Fig. 6. Simplified diagram of a two-stage amplifier for symmetrical cables. Current feedback is provided by resistor R, whilst voltage feedback is effected by the secondary winding S. The equalizing network N ensures a straight frequency-response curve of cable plus amplifier.

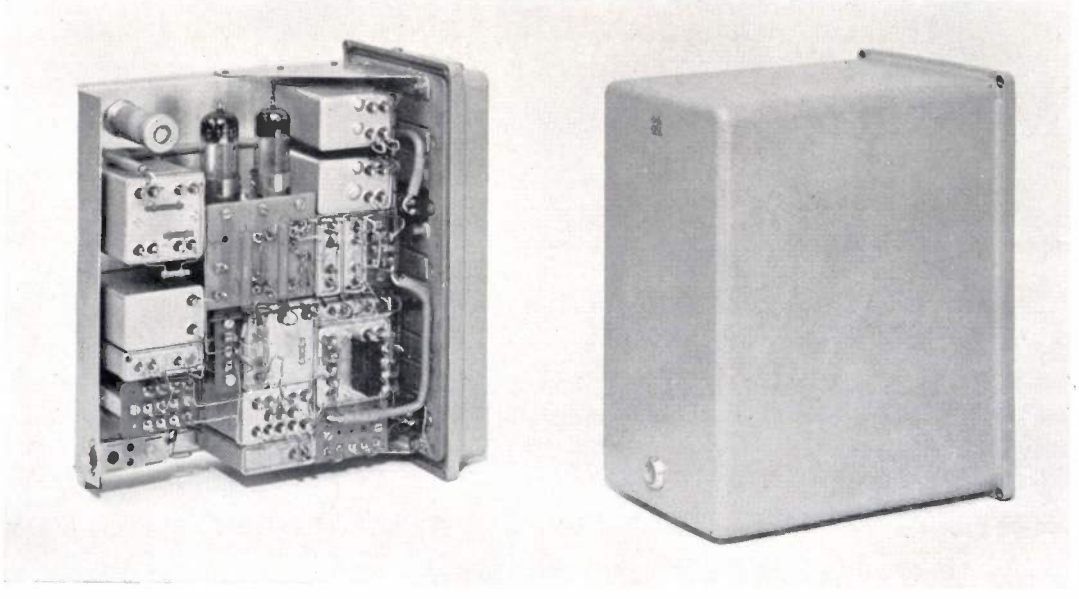


Fig. 7. The two-stage amplifier for symmetrical cables.

transformer. The equalizing network N has the function of straightening out the overall response curve of cable plus amplifier; it varies $1/\beta$ in such a way that at all frequencies the resulting amplification just compensates the cable attenuation. *Fig. 7* is a photograph of this amplifier.

Amplifier for coaxial cables (300-4200 kc/s)

As a second example we will consider a line amplifier for coaxial cables. This amplifier has been designed for the frequency band 300-4200 kc/s and is built up of three stages. The highest frequency is so much higher than that in the previous example that even by means of an amplification increasing with frequency the value of μ that could be reached with a two-stage amplfier would be too low. The diagram of *fig. 8* shows that only current feedback across the impedance Z has been applied. It should further be noticed that the coupling elements between the tubes are rather complex. This is necessary in order to obtain the closest possible approximation to the aforementioned theoretical value of the feedback factor $\mu\beta$.

All valves have a second identical valve in parallel, a necessary measure in view of the fact that in this case a single amplifier has to deal with 900 channels. *Fig. 9* shows what happens if one, two or three of the additional parallel valves are put out of action for some reason. Even in the extreme case of three non-effective valves, the change in amplification will not exceed approximately 1 dB, thanks to the feedback. We see that the overall amplification between 300 and 4200 kc/s is approximately 41 dB and if all valves are functioning, remains constant

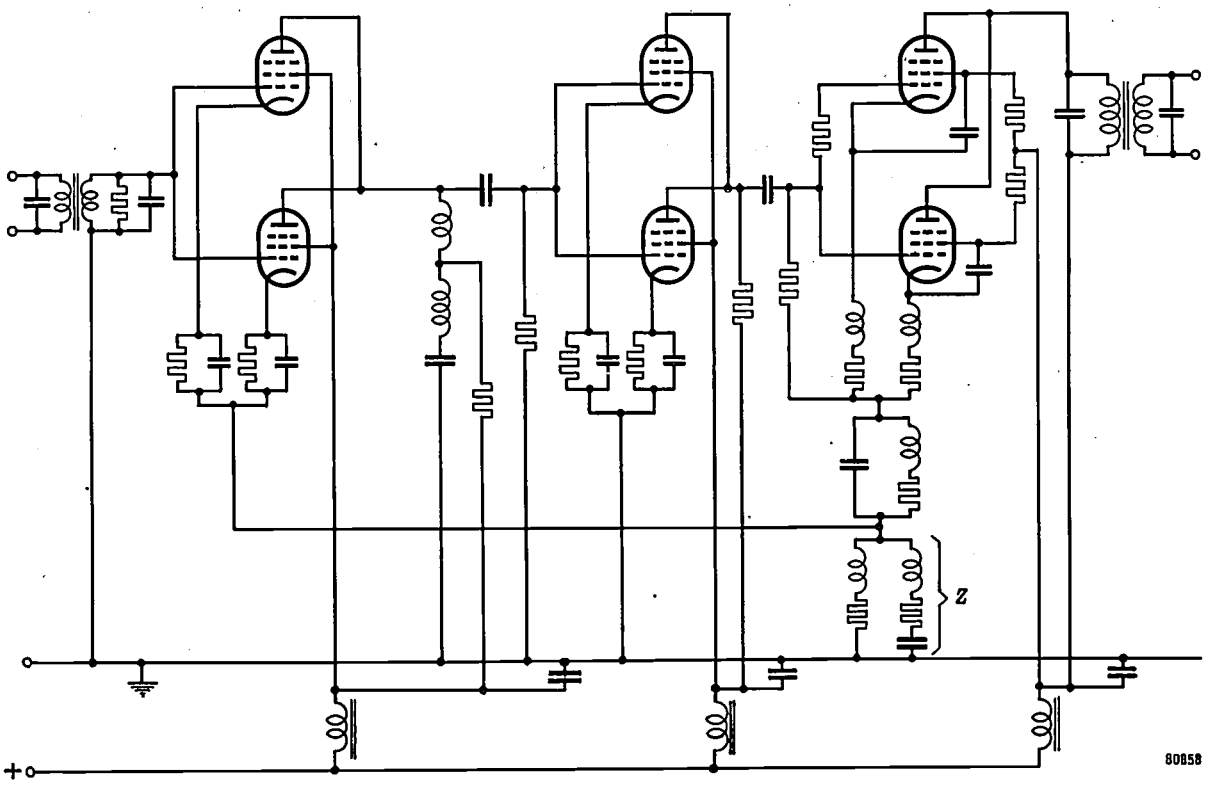


Fig. 8. Simplified diagram of the amplifier for coaxial cables. Current feedback is obtained across the impedance Z. All tubes work as parallel pairs. The coupling elements are such that approximately 80% of the theoretical maximum value of $\mu\beta$ can be attained.

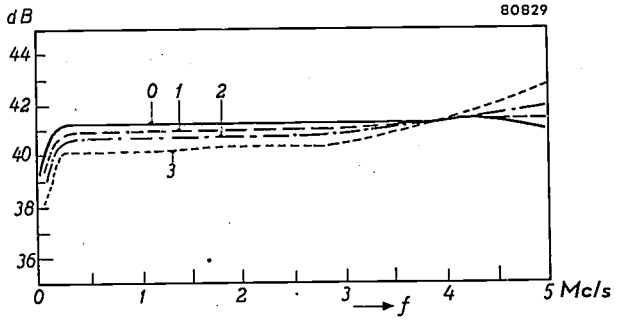


Fig. 9. Amplification as a funetion of the frequency for the amplifier for coaxial cables: *0*, all valves functioning (amplification is then constant within 0.1 dB throughout the range 0.3-4.2 Mc/s); *1*, one valve in one stage failed; *2*, one valve in each of two stages failed; *3*, one valve in each of three stages failed. Even in the last case the deviation is no more than 1 dB.

within 0.1 dB. We have demonstrated above that both the amplitude and the phase of the feedback factor $\mu\beta$ are of importance. This factor has therefore been plotted in a polar diagram in *fig. 10*. We see that in the band to be amplified $\mu\beta$ is almost constant and amounts to 30 dB (31×), but that at higher frequencies, at a phase angle of about −150° the amplitude decreases to below 1, so that the point (−1, 0) (amplification 1 at phase angle 180°) falls entirely outside the curve. Thus the condition for non-oscillation is satisfied. We see that in this respect also the behaviour of the amplifier at frequencies up to approximately 50 Mc/s is important [9]). At these frequencies parasitic effects may be very troublesome. A wire of 10 cm length then represents an inductive impedance of about 30 Ω. Very careful assembly is thus necessary and, as a rule, all connections should be kept as short as possible. The photographs of this amplifier in *fig. 11* give an impression of the assembly arrangement. The valve-holders are as near to each other as possible and the various coupling elements are mounted between them. Capacitative coupling between the valves has been avoided by building into the chassis two mounting plates, in the form of a cross, and by locating the valve holders in a suitable way with respect to these two crossed plates.

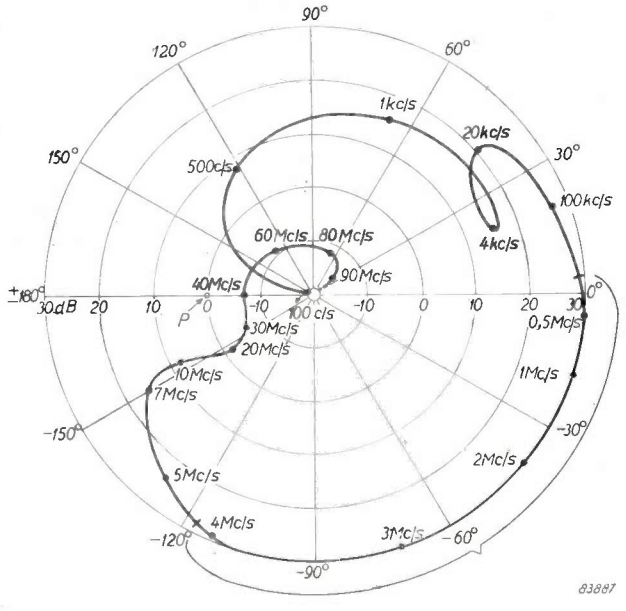
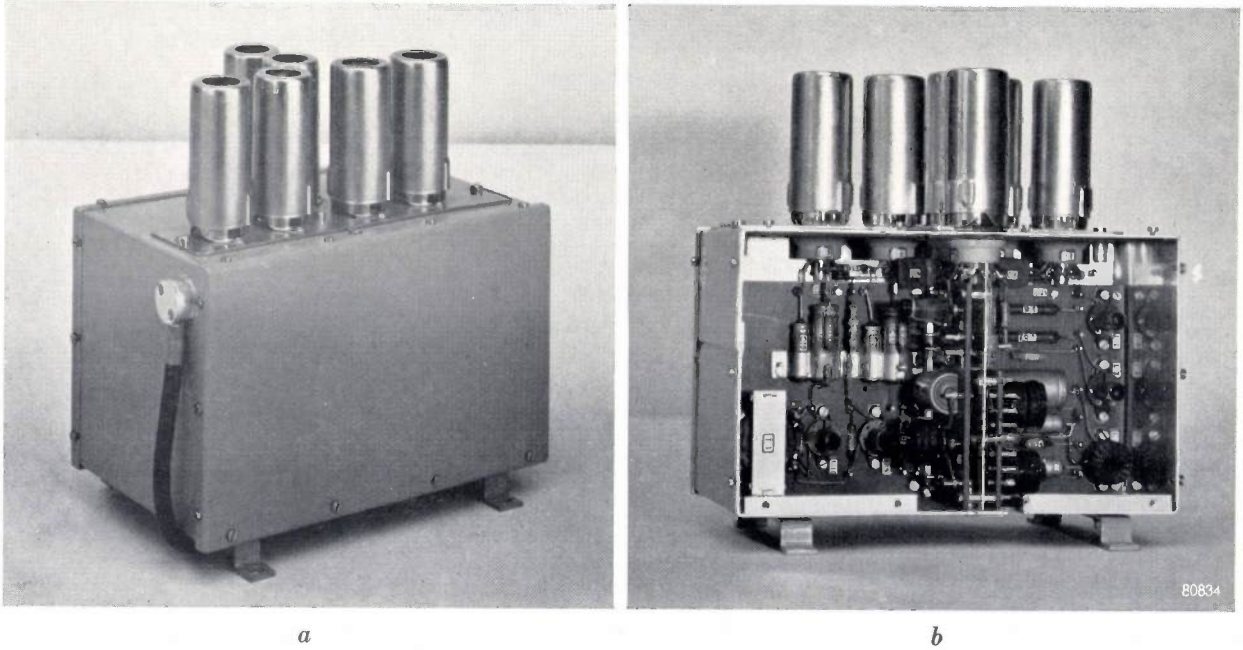


Fig. 10. Polar plot of the value $\mu\beta$ of the amplifier for coaxial cables. In the band to be amplified $\mu\beta$ is approximately 31; this range is indicated by the bracket. In view of the rather large variations to be shown, the amplitude has been plotted logarithmically (in dB) (20 dB corresponds to a factor of 10). Point P (0 dB, 180°) is clearly outside the curve, so that the amplifier, according to Nyquist's criterion, cannot oscillate. It can be seen that in order to be sure of this, it is necessary to know the value of $\mu\beta$ for frequencies up to about 50 Mc/s.

[9]) See also H. Thirup, An instrument for measuring complex voltage ratios in the frequency range 1-100 Mc/s, Philips tech. Rev. **14**, 102-114, 1952/53.

Summary. In carrier line-telephony attenuation in the cable makes it necessary to place line amplifiers at regular intervals along the cable, capable of highly stable and linear amplification over a large bandwidth. It is found that, at a given bandwidth, both the coupling impedance between the valves of such an amplifier and the amount of the amplification per stage are mainly determined by the figure of merit S/C of the valves. In order to improve the linearity and to reduce amplifi-



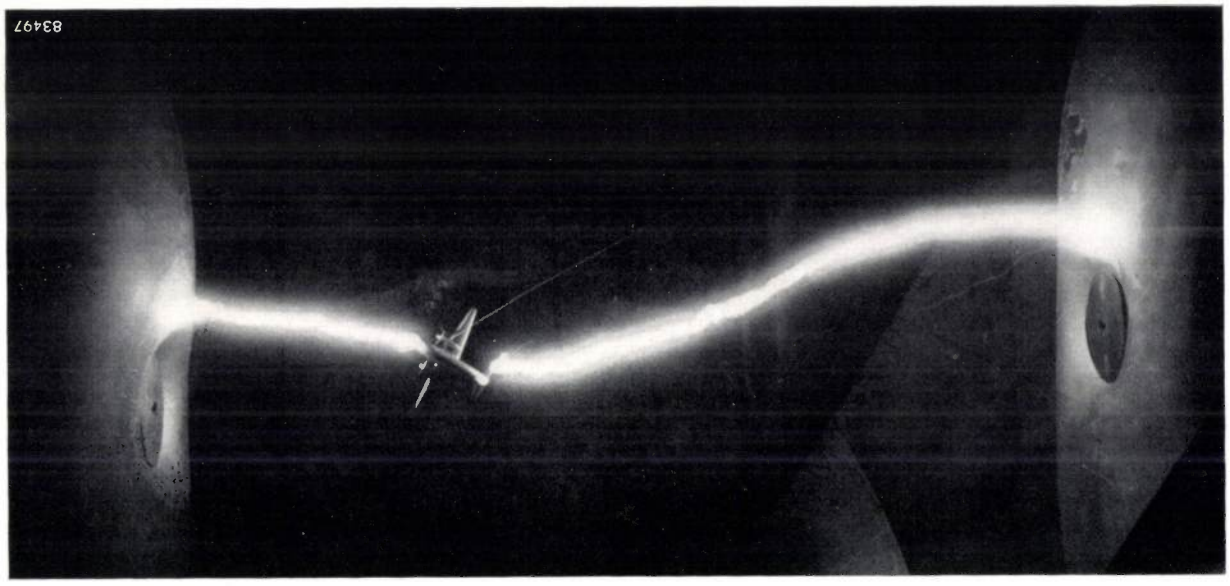
a *b*

Fig. 11. Amplifier for coaxial cables. *a*) Front. *b*) Rear.

cation variations, feedback is applied. The article demonstrates that S/C similarly limits the amount of the feedback, which, if exceeded, may put the feedback in anti-phase at very high frequences, so that the amplifier oscillates. These considerations are applied to two types of cable currently used in modern carrier telephony — the symmetrical and the coaxial cable. An amplifier for each of these types is described. The amplifier for symmetrical cables operates in the range 12-204 kc/s, which is the bandwidth required for 48 channels. By using an amplifier with a rising frequency characteristic, it is possible to use only two amplifying stages, A correct output impedance is achieved by a combined current and voltage feedback. The maximum amplification is 68 dB at a feedback factor of 32. The amplifier for coaxial cables operates in the band 0.3-4.2 Mc/s (which can accomodate a good 900 channels). It has a constant amplification of 41 dB at a feedback factor of 31. Three stages are necessary in this amplifier. To ensure a high degree of reliability all valves are paired. In the construction of this amplifier special steps are taken to reduce parasitic effects.

LIGHTNING AND AIRCRAFT

551.594.221:629.135



Thunderstorms are violent atmospheric phenomena, and on such a large scale that it hardly seems feasible to similate them in small-scale laboratory experiments.

However, it is possible to study certain aspects of these phenomena by means of scale experiments, in those cases in which the effects are essentially unchanged by the scale. An example is the impulse testing of insulators and transformers at high tension, to simulate the effects of lightning striking a power line.

The photographs reproduced in this article are illustrative of quite a different investigation undertaken recently in Eindhoven. The problem is concerned with the question as to what parts of an aircraft flying through a thunderstorm region are most likely to be struck by lightning. With the increasing speed and size of modern aircraft, it has been found that incidents of this kind are becoming more and more frequent. Although accidents only rarely result from aircraft being struck by lighting, some damage, often only slight, usually results. In order to maintain full airworthiness the damage will probably have to be repaired before the next flight; this means loss of time which can have serious effects on the economics of operations.

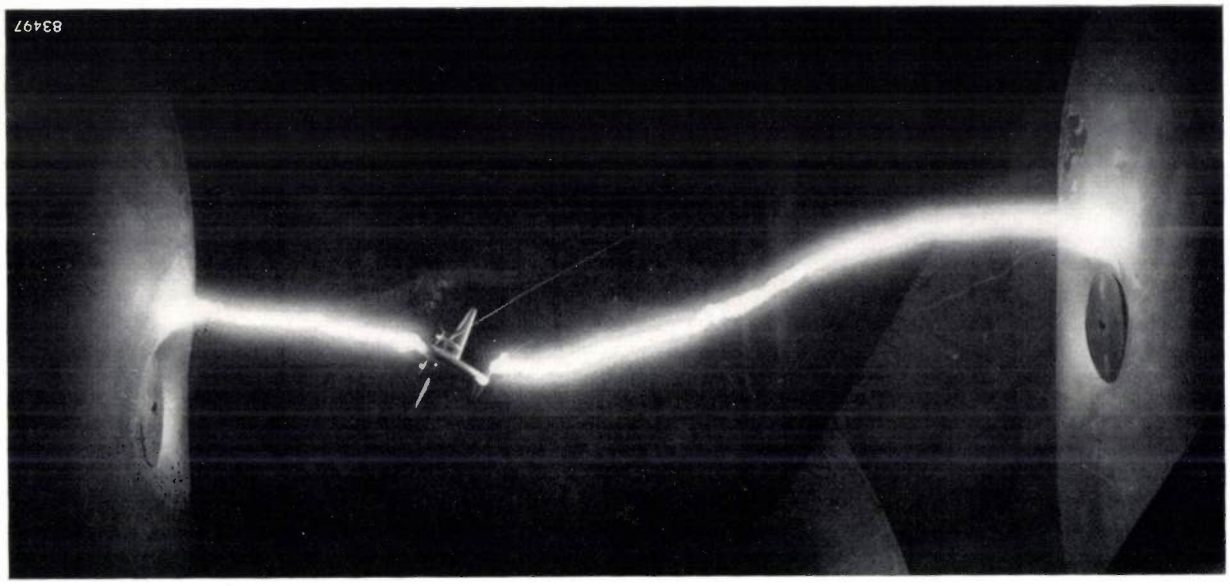
To investigate which parts of an aircraft are most susceptible to a lightening stroke, a model aircraft, scale 1 : 100, was suspended from insulating threads between two high-tension electrodes (of dimensions of some yards, see photograph at the head of this article). One electrode is connected to earth, the other to the positive high-tension terminal of a million-volt cascade generator [1]). The separation of the electrodes is such that spark-over would occur spontaneously at any smaller separation. Only a faint brushing (corona discharge) occurs. This is caused by the inevitable roughness of the electrodes, dust particles on the surface, etc; brushing is also sometimes observed on the model aircraft. The total value of such corona currents is very small (a fraction of 1 mA). In the space between the

[1]) See e.g. A. Kuntke, Philips tech. Rev. 2, 161, 1937; and A. Brouwers and F. A. Heyn, Philips tech. Rev. 6, 46, 1941.

cation variations, feedback is applied. The article demonstrates that S/C similarly limits the amount of the feedback, which, if exceeded, may put the feedback in anti-phase at very high frequences, so that the amplifier oscillates. These considerations are applied to two types of cable currently used in modern carrier telephony — the symmetrical and the coaxial cable. An amplifier for each of these types is described. The amplifier for symmetrical cables operates in the range 12-204 kc/s, which is the bandwidth required for 48 channels. By using an amplifier with a rising frequency characteristic, it is possible to use only two amplifying stages, A correct output impedance is achieved by a combined current and voltage feedback. The maximum amplification is 68 dB at a feedback factor of 32. The amplifier for coaxial cables operates in the band 0.3-4.2 Mc/s (which can accomodate a good 900 channels). It has a constant amplification of 41 dB at a feedback factor of 31. Three stages are necessary in this amplifier. To ensure a high degree of reliability all valves are paired. In the construction of this amplifier special steps are taken to reduce parasitic effects.

LIGHTNING AND AIRCRAFT

551.594.221:629.135



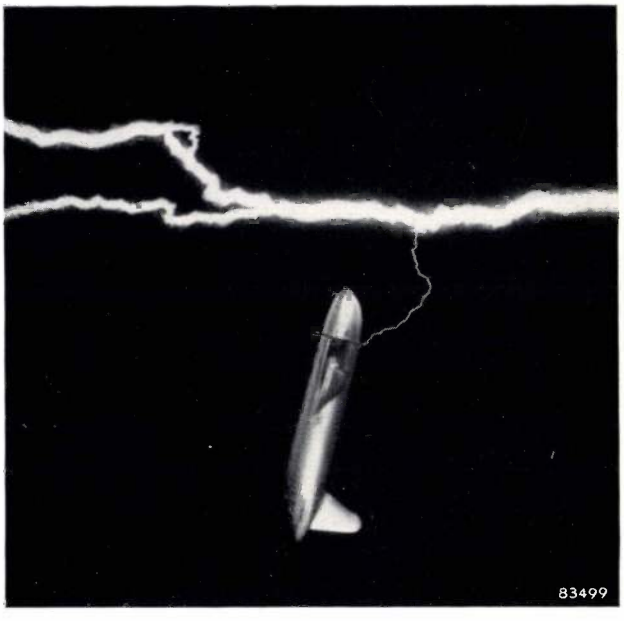
Thunderstorms are violent atmospheric phenomena, and on such a large scale that it hardly seems feasible to similate them in small-scale laboratory experiments.

However, it is possible to study certain aspects of these phenomena by means of scale experiments, in those cases in which the effects are essentially unchanged by the scale. An example is the impulse testing of insulators and transformers at high tension, to simulate the effects of lightning striking a power line.

The photographs reproduced in this article are illustrative of quite a different investigation undertaken recently in Eindhoven. The problem is concerned with the question as to what parts of an aircraft flying through a thunderstorm region are most likely to be struck by lightning. With the increasing speed and size of modern aircraft, it has been found that incidents of this kind are becoming more and more frequent. Although accidents only rarely result from aircraft being struck by lighting, some damage, often only slight, usually results. In order to maintain full airworthiness the damage will probably have to be repaired before the next flight; this means loss of time which can have serious effects on the economics of operations.

To investigate which parts of an aircraft are most susceptible to a lightening stroke, a model aircraft, scale 1 : 100, was suspended from insulating threads between two high-tension electrodes (of dimensions of some yards, see photograph at the head of this article). One electrode is connected to earth, the other to the positive high-tension terminal of a million-volt cascade generator [1]). The separation of the electrodes is such that spark-over would occur spontaneously at any smaller separation. Only a faint brushing (corona discharge) occurs. This is caused by the inevitable roughness of the electrodes, dust particles on the surface, etc; brushing is also sometimes observed on the model aircraft. The total value of such corona currents is very small (a fraction of 1 mA). In the space between the

[1]) See e.g. A. Kuntke, Philips tech. Rev. 2, 161, 1937; and A. Brouwers and F. A. Heyn, Philips tech. Rev. 6, 46, 1941.



electrodes, around the model airplane, however, a critical condition of electric stress prevails, which may break down into a disruptive discharge by any slight change. A similar condition, on the verge of a discharge, may also prevail around an aircraft in a thunderstorm. A discharge can then be triggered by any adventitious factor causing a momentary increase in field strength at some point. As the experiments involved a large number of bolts of lightning — sufficient, that is, for a statistical evaluation of the results — it was obviously impracticable to wait each time for a chance disturbance. Instead, a disturbance was induced by shooting a pellet from an airgun against the high tension electrode. Once the voltage had been adjusted near enough to the critical value, virtually every shot would trigger a full discharge [2]).

Not all, but a large number of the discharges thus triggered, chose a path via the model aircraft. On one or two occasions the model was struck by a secondary discharge branching off from the main discharge (see photograph), a phenomenon which has also been observed with natural lightning.

The place at which the model aircraft was struck by the discharge was photographically recorded. The aircraft model was placed in every conceivable spatial attitude with respect to the electric field. From the many hundreds of photographs is was possible to form some idea of the parts of the model most likely to be struck by the discharge.

On the basis of experiments of this kind it may well be possible in the future to provide certain points on aircraft with lightning arresters to prevent or minimize damage.

A. C. van DORSTEN.

[2]) The phenomenon that a rapid mechanical disturbance may cause a discharge across a gap when the applied voltage is just below the critical value was, as far as is known, first observed and studied by A. Kuntke in the Laboratory of C. H. F. Müller in Hamburg; see VDE-Fachberiche 12, 157-164, 1948.

ENTROPY IN SCIENCE AND TECHNOLOGY

II. EXAMPLES AND APPLICATIONS

by J. D. FAST.

536.75

In a previous article [1]) definitions and general aspects of entropy were treated; this article illustrates with examples the application of the entropy concept in chemistry, physics and engineering. As an example in the chemical field the reduction reactions of metal oxides, which play such an important part in metallurgy, are treated. In the field of physics, in which the statistical interpretation of entropy is explicitly expressed, the subject of paramagnetism is discussed, and the technique of paramagnetic demagnetization for the attainment of very low temperatures is explained. As examples of "classical" thermodynamics in engineering, hot-air engines, refrigerators and heat pumps are considered. The article is concluded with a discussion of the thermodynamical aspects of electromagnetic radiation.

In the first article of this series (I) [1]) the concept of entropy has been examined from various viewpoints. In this article and in the two following ones we shall demonstrate with examples the importance of the entropy concept in widely divergent fields of chemistry, physics and technology.

These examples are taken more or less at random and are merely representative of the innumerable possible applications. Unfortunately restrictions of space do not permit us to enter into one of the finest examples, viz. accurately determining the equilibrium constant of reversible gaseous reactions (e.g. $H_2 + Br_2 \rightleftarrows 2\,HBr$) by the methods of statistical thermodynamics. We can only point out that the values of the entropy of polyatomic gases, on which such determinations are based, can be derived from their spectra. The various possible states of rotation and vibration of the gas molecules can be found from the wavelengths of the spectral lines. The values of the entropy can then be calculated from the available translational, rotational and vibrational states with the aid of the formula $S = k \ln m$ (cf. I)[2]).

Chemistry and metallurgy

In article I we have seen that a system whose temperature and volume are kept constant, strives towards a minimum value of the Helmholtz free energy

$$F = U - TS.$$

On the other hand, at constant temperature and *pressure*, the value of the Gibbs free energy (also known as free enthalpy and thermodynamic potential)

$$G = U + pV - TS = H - TS$$

strives towards a minimum.

Instead of the "competition" between the energy U and the entropy S, discussed under the heading *The free energy* in part I, here a similar thing occurs between the enthalpy H and the entropy S, viz. the striving of $H = U + pV$ towards a minimum and that of S towards a maximum value.

If a system of constant temperature and pressure contains different substances, then, according to the above, only those chemical reactions will occur spontaneously between the substances for which the change in Gibbs' free energy

$$\Delta G = \Delta H - T\Delta S \quad . \quad . \quad . \quad . \quad . \quad (II, 1)$$

is negative.

In this equation ΔH is the heat of reaction at constant pressure; i.e. the amount of heat absorbed during an irreversible reaction at constant values of p and T, whilst ΔS stands for the increase in entropy as a result of the reaction.

The fact that ΔH is the ordinary (irreversible) heat of reaction at constant pressure, follows directly from the first law of thermodynamics, which can be stated here in the form:

$$Q_{irr} = \Delta U + p\Delta V = \Delta H. \quad . \quad (II, 2)$$

ΔU and ΔH, both representing the heat of reaction at constant pressure and constant volume, have nearly the same value as long as no gaseous components play a part in the reaction (e.g. $PbS + Fe \rightarrow Pb + FeS$), or if the number of gas molecules does not change during the reaction,

[1]) J. D. Fast, Entropy in science and technology I, Philips tech. Rev. **16**, 258-269, 1954/55, further to be referred to as I.

[2]) For an explanation of the methods used, cf. J. D. Fast, Entropie (in Dutch), Centen's, publishers, Amsterdam 1948, or R. W. Gurney, Introduction to statistical mechanics, McGraw Hill, New York, 1949.

e.g. ($H_2 + Cl_2 \rightarrow 2\,HCl$). If, on the other hand, the number of gas molecules formed during the reaction is not equal to the number of those that have disappeared (as in the reaction $2NH_3 \rightarrow N_2 + 3\,H_2$), then ΔU and ΔH will differ, according to (II, 2), by an amount

$$p\Delta V = \Delta(pV) \approx n\,RT,$$

in which n stands for the difference between the number of gram- molecules of newly-formed gas and that of the disappeared gram-molecules of gas. In chemistry, ΔH is more important than ΔU, since most reactions take place at a constant pressure. According to the convention adopted in I, heat added to a system is designated positive (cf. I), so that an exothermic reaction has a negative heat of reaction.

When considered thermodynamically, it can be said that a chemical reaction will occur the more readily the more negative is the value of ΔG for this reaction. In thermodynamics, therefore, $-\Delta G = -\Delta H + T\Delta S$ is called the affinity of the reaction. This affinity increases if ΔH is more negative, i.e. if the reaction is more exothermic, and also if ΔS has a greater positive value, which means a greater increase in entropy in the course of the reaction.

Chemical reactions in which only solids participate usually involve only a very slight entropy change, and are consequently always exothermic if the reaction occurs spontaneously. If gases participate in the reaction, ΔS will be usually very small if the number of molecules remains unaltered during the reaction.

Conversely, if the number of gas molecules increases by the reaction, then generally the entropy will likewise increase substantially, because the number of available micro-states m in the gaseous state is far greater than in the condensed state (cf. I, in which it was shown that $S = k \ln m$).

Let us consider as an example the oxidation of carbon according to the two equations

$$C + O_2 \rightarrow CO_2, \quad . \; . \; . \; . \quad (II, 3)$$

$$2\,C + O_2 \rightarrow 2\,CO. \quad . \; . \; . \; . \quad (II, 4)$$

In the course of (II, 3) far more heat is evolved (the enthalpy decreases substantially more) than in the course of (II, 4). The entropy, on the other hand, hardly changes in the course of (II, 3), whereas during (II, 4), due to the doubling of the number of gas molecules, it increases considerably. The result of this is that at relatively low temperatures, where the energy (or enthalpy) term prevails, carbon oxidises to CO_2, whereas at high temperatures, at which the entropy term prevails due to the factor T (cf. equation II, 1), the combustion of carbon produces CO. In the latter case, according to (II, 1) (ΔS positive) ΔG will become increasingly negative at rising temperature, i.e. the affinity of carbon to oxygen, as regards the formation of CO, shows a continuous increase as the temperature rises.

Conversely, in the case of oxidation of solid or liquid metals into solid or liquid oxides, such as

$$2\,Fe + O_2 \rightarrow 2\,FeO, \; . \; . \; . \; . \; . \quad (II, 5)$$

$$4\,Al + 3O_2 \rightarrow 2\,Al_2O_3, \; . \; . \; . \; . \quad (II, 6)$$

the affinity decreases at rising temperature, since ΔS is negative due to the reduction of the number of gas molecules.

The fact that at rising temperature the affinity of some oxidation reactions (e.g. II, 5, or II, 6) diminishes, whilst that of others (reaction II, 4) increases, has the consequence that for every metal there is a temperature above which its affinity with oxygen at a certain pressure is less than that of carbon with oxygen. We have thus found an entropy effect that is especially important in metallurgy: at a sufficiently high temperature all liquid and solid metallic oxides can be reduced by carbon. For a more quantitative study we have to know the dependence upon the gas pressure. We shall not enter into the derivation, but only mention that for oxidation reactions of the type (II, 5) and (II, 6) at not excessively high O_2-pressures, we have the equation

$$\Delta G = \Delta G^0 - RT \ln p_{O_2}. \; . \; . \; . \; . \; (II, 7)$$

where ΔG and ΔG^0 have a specific value for each reaction. ΔG^0 denotes the change of Gibbs free energy *at standard pressure*, and depends only on the temperature.

For reaction (II, 4) the equation (II, 7) becomes

$$\Delta G_{(4)} = \Delta G^0_{(4)} + RT \ln \frac{p^2_{CO}}{p_{O_2}}. \; . \; . \quad (II, 8)$$

In the equilibrium state $\Delta G = 0$, so for a simple oxidation reaction of the type (II, 5) and (II, 6), we can write from (II, 7),

$$\Delta G^0 = RT \ln p_{O_2}. \; . \; . \; . \; . \quad (II, 9)$$

Hence, from (II, 8), for reaction (II, 4) in equilibrium,

$$\Delta G^0_{(4)} = -RT \ln \frac{p^2_{CO}}{p_{O_2}}. \; . \; . \; . \quad (II, 10)$$

Knowledge of ΔG^0 at a certain temperature for e.g. reaction (II, 6) gives directly, according to (II, 9), the dissociation pressure of Al_2O_3 at that temperature, i.e. the oxygen pressure for which the

affinity of reaction (II, 6) is zero. For oxygen at 1 atm, Al has an affinity of $-\Delta G = -\Delta G^0$, according to (II, 7). This is termed the "standard" affinity $-\Delta G^0$, which means the affinity when all substances that are formed or that disappear are at standard pressure. This implies in the present examples that all solids and liquids are in a pure state and that all gases have partial pressures of 1 atm. and are assumed to be perfect gases.

Fig. 1 shows the standard affinity of various oxidation reactions as a function of the temperature. All lines in the diagram relate to the reaction with 1 gram-molecule of oxygen. Starting in the lower left-hand corner we see that ΔG^0 for the reaction

$$4\,Ag + O_2 \rightarrow 2\,Ag_2O \quad . \; . \; . \; . \quad (II, 1)$$

has a small negative value (and thus $-\Delta G^0$ has a small positive value). As explained above, the affinity decreases at rising temperature. This relationship is nearly linear, since ΔH^0 and ΔS^0 change only little with the temperature.

The slope of this "straight line" given by

$$\frac{\partial(-\Delta G^0)}{\partial T} = \Delta S^0 \quad . \; . \; . \; . \quad (II, 12)$$

is a measure of the entropy change occurring in the reaction. The standard affinity of reaction (II, 11) reaches zero at approximately 200 °C, at which temperature Ag and Ag_2O are in equilibrium with oxygen at 1 atm: at that temperature the dissociation pressure of Ag_2O is 1 atm. Above 200 °C, at an O_2-pressure of 1 atm., Ag_2O dissociates into silver and oxygen.

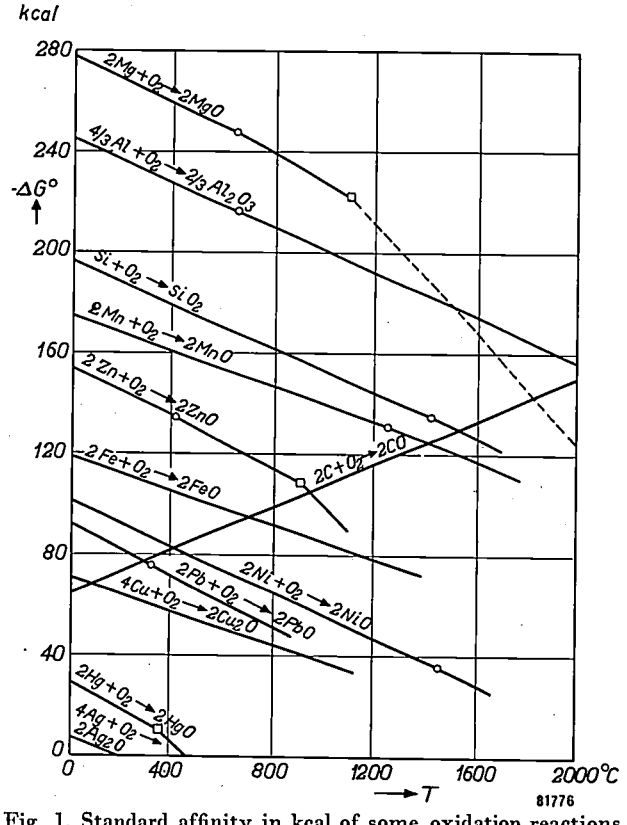


Fig. 1. Standard affinity in kcal of some oxidation reactions plotted as functions of the temperature. The circles in the diagram corresponds to the melting points, and the squares to the boiling points of the various metals. The affinity lines have not been produced beyond the melting points and boiling points of the oxides.

At the point where the metal or the oxide undergoes a change of state, there is a discontinuity in the affinity curve. A number of the lines in the diagram show this feature, e.g. the second from the bottom, relating to the reaction

$$2\,Hg + O_2 \rightarrow 2\,HgO. \quad . \; . \; . \; . \quad (II, 13)$$

The discontinuity in this case occurs at 357 °C, the boiling point of mercury. Above this temperature, mercury at 1 atm. is in the gaseous state; its entropy is then far greater than in the liquid state. $\Delta S^0_{(13)}$ is therefore more negative above 357 °C than below this temperature and the curve shows a steeper slope. Melting points and crystallographic transition points, on the other hand, cause only slight changes in the direction of the affinity curves, due to their relatively small entropy changes. For the solid and the liquid state of metal and oxide all the lines have roughly the same slope. This is due to the fact that the entropy change of all these reactions is mainly determined by the disappearance of 1 molecule of O_2.

From a metallurgical point of view, the most important exception to these roughly equal slopes is the curve for the reaction $2C + O_2 \rightarrow 2\,CO$, which rises with temperature as already mentioned. Wherever it intersects another line, e.g. that of

$$Si + O_2 \rightarrow SiO_2 \quad . \; . \; . \; . \quad (II, 14)$$

the standard affinities of the reactions (II, 4) and (II, 14) are equal, which means that the standard affinity of the reaction

$$Si + 2\,CO \rightarrow SiO_2 + 2\,C, \quad . \; . \quad (II, 15)$$

obtained by subtracting (II, 4) from (II, 14), is zero at the temperature of the point of intersection. This means that the CO equilibrium pressure of this reaction, and hence also of the inverse reaction

$$SiO_2 + 2\,C \rightarrow Si + 2\,CO \quad . \; . \; . \quad (II, 16)$$

is 1 atm. at this temperature. In fact, this pressure is reached at a lower temperature, as there exists rather a large affinity between Si and C, because of which the reaction with carbon in reality progresses further, according to:

$$SiO_2 + 3\,C \rightarrow SiC + 2\,CO. \quad . \; . \quad (II, 17)$$

This reaction corresponds to higher CO-pressures than reaction (II, 16). Also the formation of gaseous SiO should be reckoned with, which likewise causes the interaction between C and SiO_2 to occur at lower temperatures than would be expected from (II, 16).

It should be mentioned that the afinity values do not give any information on the *speed* of a chemical reaction. There are cases in which ΔG is strongly negative, yet the speed of reaction is extremely slow. Fig. 1, for example, indicates that the standard affinity of reaction (II, 6) is very great at normal room temperature, whereas it is known from experience that aluminium objects are not appreciably corroded by oxygen. Further examination shows that a very thin coating of oxide has rapidly formed on the metal, thus protecting it from further attack, since neither Al, nor 0-atoms (or ions) are capable of penetrating this skin.

The investigation of the speeds of chemical reactions falls outside the scope of thermodynamics. The latter is concerned only with states of equilibrium. The great value of thermodynamics in chemistry however, is that it can be established whether a given reaction will proceed in the desired direction simply by checking that its ΔG is negative. Of equal importance are the quantitative conclusions which may be drawn, viz. the chemical equilibrium of a reaction can be derived from ΔG^0; i.e. the maximum possible yield of reaction products can be calculated in advance. As an example, let us consider reaction (II, 4), for which we can put, using equations (II, 1) and (II, 10):

$$\ln \frac{p_{CO}^2}{p_{O_2}} = -\frac{\Delta H^0}{4.575\,T} + \frac{\Delta S^0}{4.575}. \qquad (II, 18)$$

Instead of calculating the equilibrium directly from ΔG^0, which is difficult to be determined directly, it can also be found from ΔH^0 and ΔS^0. The former of the two, the constant-pressure heat of reaction under standard pressure, is known from calorimetric measurements. The latter (ΔS^0), the entropy change at standard pressure can be calculated from the specific-heat values which are known from calorimetric and spectroscopic data.

The particular importance of the knowledge of thermodynamics to the metallurgist was strikingly demonstrated in the past century by the great French chemist Le Chartelier. His line of reasoning was as follows. It is known that in blast-furnaces the reduction of iron oxides is mainly effected by their reaction with CO, with the formation of CO_2. The gas leaving the blast-furnace at the top, however, contains a considerable percentage of CO, which was considered wasteful and undesirable. Since the incompleteness of the reaction was attributed to the insufficiently prolonged contact between CO and iron ore, blast-furnaces of abnormal heights were errected. The percentage of CO in the escaping gas, however, did not diminish. These costly experiments showed that there is a certain upper limit to the efficiency of CO as a reducing agent for iron oxides, which cannot be exceeded. Had the laws governing the chemical equilibrium been known, the same conclusion could have been reached far more quickly and at far less cost. It is only necessary to know the reaction enthalpy ΔH^0 and the reaction-entropy ΔS^0 of the reaction

$$FeO + CO \rightarrow Fe + CO_2 \quad . \quad . \quad . \quad (II, 19)$$

for the standard conditions, to be able to calculate directly the maximum fraction of the CO that can be converted into CO_2 from the equation

$$\ln \frac{p_{CO_2}}{p_{CO}} = -\frac{\Delta H^0}{4.575\,T} + \frac{\Delta S^0}{4.575} \quad . \quad (II, 20)$$

Reaction (II, 19) invloves only relatively small changes in enthalpy and entropy, viz. at 25 °C $\Delta H^0 = -3800$ cal and $\Delta S^0 = -3.0$ cal/degree. With the aid of these values, and disregarding their temperature dependence, it is found by (II, 20) that escaping ($CO + CO_2$)-mixture will have at least 40%-60% of CO in the temperature range 1000-2000 °K [3]).

Paramagnetism

The molecules of paramagnetic substances possess a magnetic moment, i.e. they behave to a certain extent as small magnets. The property of paramagnetism is a consequence of the movements of the electrons in the molecule as well as of the angular momentum (spin) of the electrons. In most molecules the motions of the electrons are coupled in such a way that the resulting magnetic moment is zero.

An example of a gaseous paramagnetic substance is oxygen. The magnetic properties of O_2 are based on the fact that two of the electrons of each oxygen molecule are "unpaired", i.e. have similarly directed (i.e. parallel) angular momenta (spins) and consequently also parallel magnetic moments of the same sign. All the other electrons in the O_2 molecule, on the other hand, are paired. A gaseous alkali metal provides an even simpler example, because here the paramagnetic properties are based on the presence of only one non-compensated electron spin per atom

[3]) A more accurate calculation, taking into account the temperature-dependence of ΔH° and ΔS°, gives even higher percentages of CO.

of the gas. Several hydrated salts show a marked analogy to these paramagnetic gases, because the spin carriers in these salts are separated by such a distance that their interaction, just as in the gases, may be neglected to a first approximation.

If oxygen, gaseous sodium or a paramagnetic salt, is brought into a magnetic field, then the striving towards a minimum value of the energy corresponds to the striving of the permanent atomic magnets to assume an orientation with their axes parallel to the field, just as compass needle in the magnetic field of the earth. A parallel orientation of all magnetic axes is by no means attained, at least not at high temperature, since the thermal agitation associated with the striving of the entropy towards a maximum value is opposed to this. The orientation of each particle with respect to the direction of the field is continuously changing due to the thermal agitation but, averaged over all molecules, there always remains a small resultant moment in the direction of the field.

In view the two opposed tendencies — the energy and the entropy effect — it is clear that the resultant moment (the magnetization) will be related directly to the field strength H but inversely to the absolute temperature T. In fact, it is found to be proportional to H and inversely proportional to T. The dependence on the temperature is known as Curie's law.

If the field strength is further increased, the magnetization is, in the long run, unable to keep up the same rate of increase. Finally a state of saturation is bound to occur in which all magnetic axes are in the parallel orientation. Fields strong enough to attain this saturation at normal room temperature, however, cannot be generated in the laboratory but by applying a combination of very strong fields and very low temperatures this state of saturation has been very slosely approached for several paramagnetic salts. The same principle has also been employed to reach temperatures considerably lower than hitherto possible. Before we describe this in the following section, we shall derive first the dependence of the magnetization on H and T for a simple case.

We shall consider the case that the paramagnetic properties are based on the presence of one non-compensated electron spin per molecule [4]) and that the directions of the magnetic moments are practically independent of each other. The quantum theory states that the component of the spin in the direction of the magnetic field can assume only the values $+\frac{1}{2}$ and $-\frac{1}{2}$ (expressed in units of $h/2\pi$). This corresponds to a magnetic moment of $+\mu_B$ or $-\mu_B$ in the direction of the field, if the symbol μ_B represents a Bohr magneton. The energy levels of these two states coincide when the field strength approaches zero. At a field strength H, however, the energy levels relating to the field-free state have the values $-\mu_B H$ and $+\mu_B H$. In this case, then, we are concerned with only two energy levels. The distribution of the magnets between these levels as a function of H and T can be derived as follows.

In a system of N molecules, at a given value of H a larger fraction will be formed at the lower than at the higher energy level. If the difference between the two populations is denoted by n, then $\frac{1}{2}(N+n)$ molecules have their magnetic moments in the direction of the field and $\frac{1}{2}(N-n)$ in the opposite direction. The system then possesses a moment $n\mu_B$ in the direction of the field and an energy

$$U = U_0 - \tfrac{1}{2}(N+n)\mu_B H + \tfrac{1}{2}(N-n)\mu_B H = U_0 - n\mu_B H, \qquad \text{(II, 21)}$$

where U_0 represents the energy in the absence of the magnetic field. According to I, the entropy of the system is determined by the number of possible distributions of N molecules, subject to the condition that a fraction $\frac{1}{2}(N+n)$ is in one state and a fraction of $\frac{1}{2}(N-n)$ in the other state:

$$S = S_0 + k \ln \frac{N!}{\{\frac{1}{2}(N+n)\}!\,\{\frac{1}{2}(N-n)\}!}. \qquad \text{(II, 22)}$$

In this equation S_0 represents that part of the entropy which is independent of the orientation of the moments.

The free energy $F = U - TS$ contains, therefore apart from a term $U_0 - TS_0$ independent of n, also a term that can be approximated by means of Stirling's formula (I, 3) by

$$F(n) = -n\mu_B H + kT\left\{\frac{N+n}{2}\ln\frac{N+n}{2N} + \frac{N-n}{2}\ln\frac{N-n}{2N}\right\}. \qquad \text{(II, 23)}$$

If the system could submit completely to the tendency towards a minimum value of its energy, all molecules would be at the lowest level ($n = N$). If, on the other hand, it could submit to the tendency towards a maximum entropy value, then the molecules would be equally divided between the two levels ($n = 0$). The compromise (the equilibrium state) lies according to I at the point where the free energy is minimum, i.e. where $\mathrm{d}F(n)/\mathrm{d}n = 0$.

[4]) For the sake of convenience the term "molecule" is used here. Clearly the discussion also applies to atoms and ions.

Hence $$\frac{kT}{2}\ln\frac{N+n}{N-n}=\mu_B H$$

or, after re-arranging,

$$\frac{n}{N}=\frac{e^{\mu_B H/kT}-e^{-\mu_B H/kT}}{e^{\mu_B H/kT}+e^{-\mu_B H/kT}}=\tanh(\mu_B H/kT).$$

The magnetic moment of the system is, therefore,

$$n\mu_B = N\mu_B \tanh(\mu_B H/kT).$$

The moment per unit volume is

$$I = N'\mu_B \tanh(\mu_B H/kT). \quad . \quad . \quad \text{(II, 24)}$$

where N' represents the number of molecules per unit volume. In case of $\mu_B H/kT \ll 1$, i.e. where relatively weak fields or relatively high temperatures are concerned, $\tanh(\mu_B H/kT) \approx \mu_B H/kT$, and consequently

$$I \approx \frac{N'\mu_B^2}{kT}H. \quad . \quad . \quad . \quad . \quad \text{(II, 25)}$$

For this limiting case we have thus arrived at the relationship between I and H and between I and $1/T$, mentioned earlier.

The susceptibility χ in this region is given by

$$\chi = \frac{I}{H} = \frac{N'\mu_B^2}{kT}. \quad . \quad . \quad . \quad . \quad \text{(II,26)}$$

If a molecule possesses more than one non-compensated electron-spin, these spins often combine into one resultant molecular spin. Such a resultant spin has more than two possible orientations in a magnetic field.

For a spin quantum number s ($s = \frac{1}{2}$ for one non-compensated electron spin) the general rule is that the component in the direction of the magnetic field can assume only the values s, s$-$1, s$-$2, ..., $-$s + 1, $-$s. For the case under discussion ($s = \frac{1}{2}$) the components $+\frac{1}{2}$ and $-\frac{1}{2}$ were possible. For $s = 1$ (two similar spin orientations per molecule) we find the three possibilities +1, 0 and -1. As a rule there are 2s + 1 different possibilities of orientation of the resultant spin.

A calculation analogous to the one carried out for $s = \frac{1}{2}$, shows that for $\mu_B H/kT \ll 1$,

$$\chi = \frac{N'4s(s+1)\mu_B^2}{3\,kT}, \quad . \quad . \quad . \quad \text{(II, 27)}$$

which reduces to (I, 26) for $s = \frac{1}{2}$.

Formula (II, 27) is frequently written as

$$\chi = \frac{N'\mu_{eff}^2}{3\,kT}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(II, 28)}$$

where μ_{eff} represents the effective magnetic moment defined by:

$$\mu_{eff}^2 = 4s(s+1)\mu_B^2.$$

Equation (II, 28) is identical in form to the equation for the contribution of a dipole to the electric susceptibility, χ_{dip}, of a rarefied gas of molecules such as H_2O, NH_3, HCl, which possess a permanent electric moment μ_{el}:

$$\chi_{dip} = \frac{N'\mu_{el}^2}{3\,kT}. \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(II, 29)}$$

The theoretical background of the formulae (II, 28) and (II, 29), however, is somewhat different. In very strong fields or at extremely low temperatures the electric polarization is given by $N'\mu_{el}$, whereas the magnetization is *not* defined by the analogous expression $N'\mu_{eff} = N'2\sqrt{s(s+1)}\,\mu_B$, but by $N'2s\mu_B$.

Fig. 2 gives the curves of the magnetic moment as a function of H/T for three paramagnetic salts. Consider the middle curve, which refers to $FeNH_4(SO_4)_2 \cdot 12H_2O$, i.e. ferric ammonium alum. In this compound the ferric ion embodies the magnetic

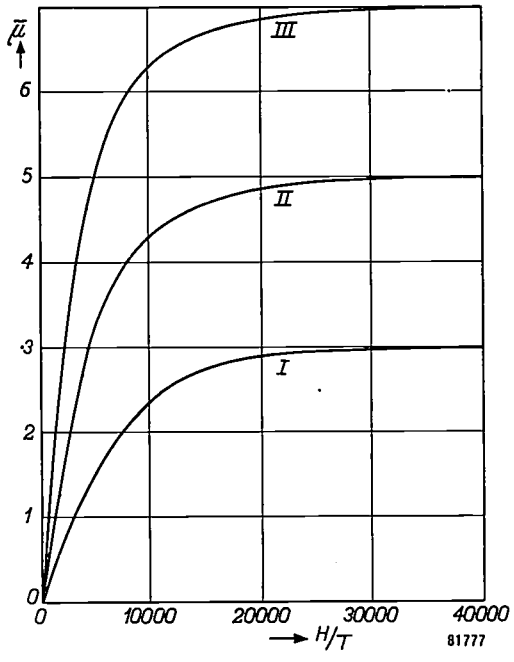


Fig. 2. Average magnetic moment $\overline{\mu}$ in Bohr magnetons per molecule as functions of H/T in oersteds per °K for (*I*) sodium-chrome alum ($s = 3/2$), (*II*) ferric-ammonium alum ($s = 5/2$) and (*III*) gadolinium sulphate octahydrate ($s = 7/2$). From measurements of C. J. Gorter, W. J. de Haas and J. van den Handel, Proc. Kon. Ned. Akad. Wetensch. **36**, 158-167, 1933, and W. G. Henry, Phys. Rev. **88**, 559-562, 1952.

properties. In its outer electron shell it contains 2 + 6 + 5 electrons ($3s^2\ 3p^6\ 3d^5$). Of these, the five $3d$-electrons are unpaired, i.e. they have parallel spins. All other electrons are paired and cannot be influenced by a magnetic field. The spin quantum number of this ion is hence $s = 5/2$. In accordance with this, the curve shows that the mean magnetic moment in the direction of the field, at large values of H/T, reaches a saturation value of 5 Bohr

magnetons per ion. The curve shows also that this mean moment in the direction of the field is proportional to H/T at small values of the latter. The two other curves refer to chrome alum ($s = 3/2$) and gadolinium sulphate-octahydrate ($s = 7/2$).

Attaining very low temperatures

By allowing liquid helium to boil under reduced pressure, a temperature not much lower than about 1 °K can be attained. Relatively much lower temperatures may be reached by employing the paramagnetic properties of a salt such as the above-mentioned ferric-ammonium alum. Pioneer work in this field was carried out in the Netherlands in the Kamerling-Onnes Laboratory of Leiden University [5]).

We have shown in the above that the resultant of five parallel spins and hence also the resultant spin of a ferric ion can assume only six different orientations relative to a magnetic field. If the field is very weak, there is little preference energetically for any of the six orientations, i.e. the six energy levels corresponding to the possible orientations are almost coincident in a weak field, and at not too low temperatures they will be equally occupied.

At small values of H/T, therefore, each molecule of the ferric-ammonium alum may occur in any one of six different but equally probable paramagnetic states. For each of the six possible states of the first molecule, the second molecule may be present in any one of the six states. There are thus 6^2 different states for two molecules, and 6^{N_0} different states for a gram-molecule (N_0 molecules). That part of the entropy dependent on the paramagnetic properties of the salt, (also called paramagnetic entropy, or, more specifically, spin entropy), has, according to the well-known formula $S = k \ln m$, the value

$$S = k \ln 6^{N_0} = R \ln 6$$

per gram-molecule, R being the gas constant. To put it more generally: if s is the quantum number of the resultant spin, then the spin entropy per gram-molecule is given by

$$S = R \ln (2s + 1) \ldots\ldots\ldots \text{(II, 30)}$$

The energy levels of the $(2s + 1)$ different states become more separated as the magnetic field becomes stronger. At a very strong field and a continuously dropping temperature the higher energy levels will be gradually emptied until ultimately the state of saturation, discussed earlier, is approached, in which practically all molecules are at the lowest level. The paramagnetic entropy per gram-molecule for this extreme case is given by

$$S \approx R \ln 1 = 0.$$

A considerable lowering of the temperature can now be attained by bringing about an adiabatic and reversible demagnetization of the magnetically saturated substance. The effect of this is completely analogous to the temperature drop occurring with the adiabatic -reversible expansion of a gas. In either case the temperature drop can be achieved by a two-step procedure. First stage: the volume of the gas is reduced in an isothermal (reversible) way, or the magnetization of the salt is isothermally and reversibly increased. In either case the result is a decrease of the entropy (in the case of the gas, as in I, p. 209), which means (since $T\mathrm{d}S = \mathrm{d}Q_{\mathrm{rev}}$) a flow of heat out of the gas or the salt. This is carried off to a heat sink at constant temperature (a quantity of liquid helium, boiling under reduced pressure in the case of the paramagnetism). Second stage: the thermal contact with the surroundings is broken, after which the pressure on the gas is reversibly reduced, or the magnetic field is reversibly decreased to zero. During this adiabatic-reversible process the total entropy of the gas or salt, since $\mathrm{d}S = \mathrm{d}Q_{\mathrm{rev}}/T$ and $\mathrm{d}Q_{\mathrm{rev}} = 0$, remains unaltered.

In the case of the gas, it can be seen directly that the increase of the entropy due to the expansion has to be compensated by a reduction of the thermal entropy. This implies a decrease of the temperature [6]) The case of the paramagnetic salt is not quite so simple. An analogous reasoning would be as follows. When the magnetic field is reduced to zero, the paramagnetic entropy $S = R \ln (2s + 1)$ increases and this happens at the cost of the vibrational entropy of the atoms, since the total entropy remains constant. In fact, this latter entropy is extremely small at an initial temperature of approximately 1 °K, far smaller than the paramagnetic entropy. At first sight it could thus be expected that during the adiabatic demagnetization the temperature would drop to 0 °K, whilst, moreover, a considerable part of the magnetic order would be retained. The latter is indeed the case, but the temperature does not drop to 0 °K. The lowest temperatures that can be

[5]) See, e.g., W. J. de Haas, E. C. Wiersma and H. A. Kramers, Physica **1**, 1-13, 1934.

[6]) That this adiabatic-reversible process is extremely suitable for producing low temperatures is demonstrated by the development of the gas refrigerating machine, which operates on this principle. See the articles by G. W. L. Köhler and C. O. Jonkers, Philips tech. Rev. **16**, 69-78, 1954/55 (No. 3) and **16**, 105-115, 1954/55 (No. 4).

attained in this way are between 0.01 and 0.001 °K. That the temperature cannot be reduced further is due to a slight interaction among the magnetic moments and between the moments and the crystal lattice, which have not been taken into account up to now. These interactions have the effect that the energy levels of the $(2s + 1)$ possible states of the paramagnetic ions do not coincide completely, even in the absence of an external magnetic field. In accordance with the third law of thermodynamics (cf. I) this causes all spins to have a fixed orientation at absolute zero. This means that at very low temperatures and in the absence of a magnetic field, the disorder of the spins automatically disappears under the influence of the interaction, causing the entropy to drop to zero. The curve of *fig. 3* represents the entropy of a gram-molecule of ferric-ammonium alum (in the absence of an external magnetic field) as a function of the temperature. Between 0.5 °K and 5 °K the entropy remains nearly constant at the earlier mentioned value of $R \ln 6$. The effect of the lattice vibrations on the entropy only becomes discernible above 5 °K.

In the preceding section we have seen that the difference between two energy levels of an electron in a field H is $2\,\mu_B H$. To a first approximation, we may assume that the $(2s + 1)$ levels of the paramagnetic ions are separated by uniform distances $2\mu_B H$ in a strong field H, and that they are situated at far smaller intervals $k\Theta_0$ in a zero field (see *fig. 4*). Furthermore, (see above) we may disregard the entropy of the lattice vibrations at 1 °K or lower with respect to the paramagnetic entropy. The latter therefore, remains nearly constant during the adiabatic demagnetization, which means in the statistical concept of entropy that the electron spins at the final temperature T_2 (after removal of the field) are

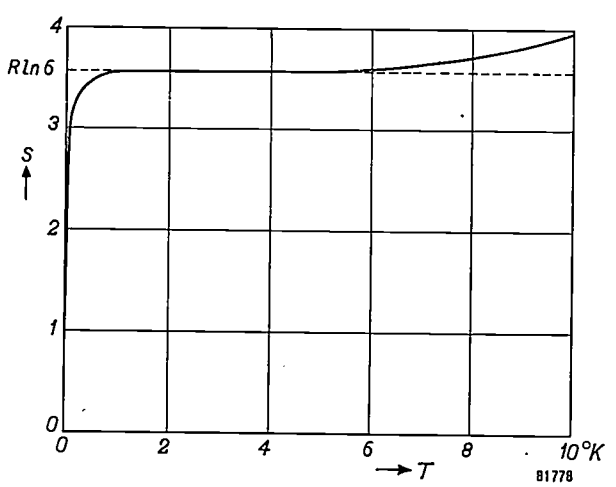


Fig. 3. Entropy in cal per °K per gram-molecule of ferric-ammonium alum (in the absence of a magnetic field) as a function of the temperature.

distributed among the right-hand group of levels in fig. 4 in the same arrangement as they were at the initial temperature T_1 in the field H among the middle group. According to I this distribution is

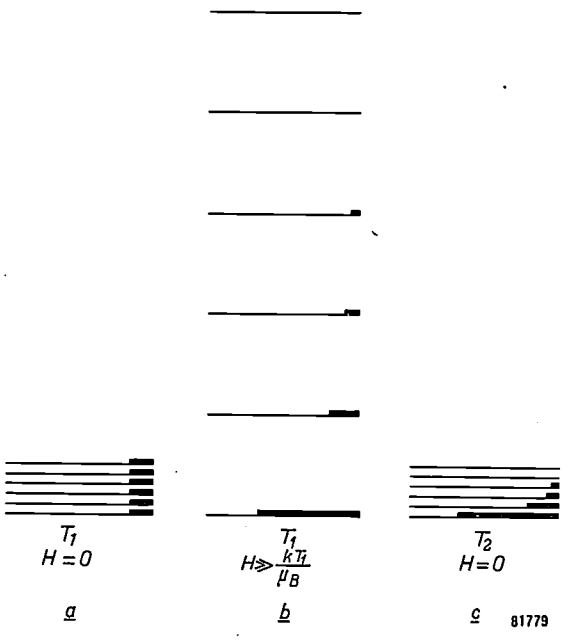


Fig. 4. Adiabatic demagnetization.
a) At the initial temperature T, kT_1 is considerably greater than the separation of the energy levels; the spins are uniformly distributed among the six levels. This is indicated by the thick parts of the lines.
b) When a strong field such that $\mu_B H \gg kT$, is applied isothermally, the higher levels will be emptied.
c) When the field is adiabatically removed, the entropy remains constant, which means that although the levels come as near together as in the beginning, the distributions remain the same as in (*b*). (From an article by N. Kurti in the booklet "Low temperature physics", Pergamon Press, London 1952.

determined by the ratio of the separation between the levels to kT, i.e. by $2\mu_B H/kT$ and $k\Theta_0/kT_2$ respectively. To a rough approximation, therefore,

$$\frac{2\mu_B H}{kT_1} \approx \frac{k\Theta_0}{kT_2}, \quad \text{or} \quad T_2 \approx \frac{k\Theta_0}{2\,\mu_B H}\,T_1. \quad \text{(II, 31)}$$

The final temperature T_2 is thus lower as the interaction between the magnetic moments (and hence $k\Theta_0$) is less, the field H stronger, and the starting temperature T_1 lower. The first condition is satisfied by using salts such as $FeNH_4(SO_4)_2 . 12H_2O$ in which the paramagnetic ions (Fe^{+++}) lie far apart. It is to be expected that in future even lower temperatures can be attained by employing the very weak coupling of the spins of certain atomic nuclei.

The foregoing shows that below 1 °K the concept of temperature is not so much coupled to the vibrational disorder but to the disorder of the spins of the system. The emptying of the higher energy levels is already perceptible at a temperature for

which kT is equal to the separation of the levels, i.e. at the temperature $T = \Theta_0$. This "characteristic" temperature may be considered as a Curie-temperature. Its approximate value can be derived from specific-heat measurements.

Heat engines

Considerations of entropy play an important part in the study of machines which convert heat into work, which may be collectively termed heat engines (steam engines, hot-gas engines, etc.). The motions of the atoms of a body, e.g. the piston of a steam engine, only provide useful work when they are uniformly directed, i.e. when the piston moves as a whole. The conversion of heat into work means in principle the conversion of the completely random motions of atoms into a uniformly directed (ordered) motion. i.e. a decrease of the entropy. According to the second law of thermodynamics this is not feasible unless compensated in some way. Indeed, one of the earliest formulations of the second law of thermodynamics was: It is impossible to abstract heat from a heat-reservoir and to convert it *completely* into work.

The reverse process, on the other hand, the complete conversion of work into heat (e.g. by friction which occurs spontaneously and therefore irreversibly, is a commonplace. This is sometimes referred to as the "degeneration of energy", since here a form of energy valuable to mankind is changed into a less useful form.

How do we succeed, in spite of all this, in converting heat into work? By using the striving of the free energy towards a minimum value, discussed in the first article, it is possible to decrease the entropy, provided that this decrease is compensated by a simultaneous decrease of the energy of the (non-isolated) system. This method cannot be used in practice, however, because a heat engine has to complete the same cycle again and again, whereby at the end of each cycle the piston returns to its initial position and the initernal energy reaches the initial value again.

The requirement of the second law has therefore to be met in another way, viz. by the fact that only part of the heat of combustion of the fuel is converted into work, while another part is carried off to a cooler, i.e. to a heat sink at a lower temperature level. The reduction of the entropy, corresponding to the conversion of heat into mechanical energy of the piston, is thus offset by an entropy increase caused by the flow of another quantity of heat from the engine to the cooler.

Taking the steam engine as an example, its working cycle is as follows. The steam absorbs a quantity of heat Q_1 at a high temperature T_1 from the boiler, and then discharges a smaller absolute heat value Q_2 at a lower temperature T_2 to the condenser. The difference $Q_1 - Q_2$ is gained in the form of work.

According to the second law in its form $dQ/T \leqq dS$, which becomes $\oint dQ/T \leqq 0$ for a cyclic process, the following equations applies to our simplified model of a steam engine:

$$\frac{Q_1}{T_1} - \frac{Q_2}{T_2} \leqq 0,$$

or:

$$\frac{Q_1 - Q_2}{Q_1} \leqq \frac{T_1 - T_2}{T_1}. \quad \ldots \quad \text{(II, 32)}$$

The first term of this equation simply represents the proportion of the applied heat that has been converted into work, i.e. the efficiency of the heat engine. The formula indicates that the highest possible efficiency would be achieved by a machine with completely reversible cycle (Carnot cycle); this optimum efficiency would then equal $(T_1 - T_2)/T_2$.

For many types of steam engines we may take $T_1 \approx 430$ °K and $T_2 \approx 300$ °K. The optimum efficiency is therefore about 0.3 (30 %). In practice the efficiency of this type of machine remains below 20 %.

Refrigerators and heat-pumps

The heat engine discussed above can also be made to function in the reverse direction. Instead of displacing heat from a reservoir of high temperature to one of low temperature and converting the greatest possible portion of it into work "on the way", we now perform work in order to transfer heat from the reservoir at low temperature to that at high temperature. Whereas with the former cycle the working fluid (the gas or the vapour) took up more heat at high temperature than it discharged at low temperature, it will now discharge more heat at high temperature than it absorbs at low temperature The difference constitutes the minimum external energy W that is required for driving the machine. Such a machine is termed a refrigerator if its function consists primarily of abstracting heat from the reservoir at low temperature; if its task consists essentially of delivering heat to the reservoir at high temperature it is termed a heat-pump. In the former case we are interested in Q_2, and the efficiency is given by:

$$\frac{Q_2}{W} \leqq \frac{T_2}{T_1 - T_2}. \quad \ldots \quad \text{(II, 33)}$$

For a heat-pump the essential factor is Q_1 and the efficiency is given by

$$\frac{Q_1}{W} \leqq \frac{T_1}{T_1 - T_2} \quad \ldots\ldots \quad (II, 34)$$

Let us illustrate our point by considering the case that $T_1 = 300$ °K and $T_2 = 270$ °K. When our machine is used as a domestic refrigerator, it has the purpose of keeping the interior of the box at 270 °K = —3 °C, whilst the ambient temperature is 300 °K = 27 °C. When functioning as a heat-pump it has the task of abstracting heat from outside at —3 °C and using it for keeping the interior of a building at 27 °C. The upper limit of the "efficiency" (or better, the *coefficient of performance*) is 270/30 = 9 for the refrigerator and 300/30 = 10 for the heat-pump. In practice, of course, the efficiency remains considerably below these values [7]).

Compared with a heat pump, heating by means of an electric resistance heater is very uneconomical. At first sight one is inclined to think that there is hardly a more favourable proposition than direct electric heating, since all the energy is converted into heat. The foregoing, however, shows that a heat pump may be several hundred percent more "efficient". The heat pump is as yet used only to a very limited extent. In practice it commonly takes the form of a liquid with high latent heat of evaporation (e.g. CF_2Cl_2) which is made to evaporate at low temperature, during which it abstracts the nccessary heat from a source at low temperature, e.g. the outside atmosphere or the ground. Condensation is then made to take place at the higher temperature, delivering the heat of condensation to where it is required e.g. the interior of a building. It can be safely predicted that as conventional sources of energy become more and more exhausted, interest in the heat pump will increase. This machine is particularly suited to the heating of buildings, because its efficiency is especially high for small values of $T_1 - T_2$ (see eq. II, 34).

Radiation of heat and light

All bodies radiate energy, or, in other words, emit photons. The quantity of energy emitted per unit time and per unit area increases rapidly with the temperature and depends on the nature of the radiating body. The "blacker" the body, i.e. the more radiation it absorbs, the better emitter of radiation it will be. Maximum emission occurs with the "perfect black body", the body that absorbs all incident light. Such a body does not exist, but can be very closely approximated by a small aperture in the wall of an otherwise light-tight chamber. Any radiation entering through the hole is scattered and absorbed by repeated reflection (*fig. 5*), so that only an infinitesimal fraction escapes. If the chamber is at a temperature T, the aperture radiates as a perfect black body of that temperature. What is emitted is the same radiation that is present inside the chamber.

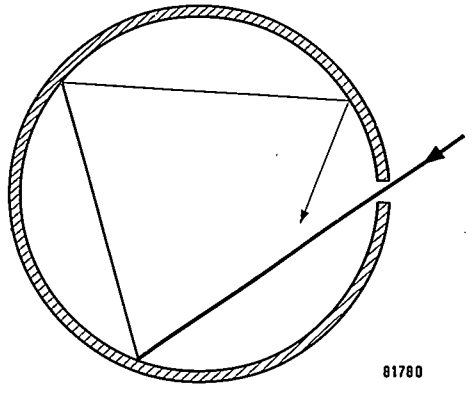


Fig. 5. A small aperture in a hollow body acts as a perfect black body, i.e. a perfect emitter and absorber of radiation.

We shall not concern ourselves here with the distribution of the radiation among the different wavelengths, but merely use thermodynamical arguments to arrive at the Stefan-Boltzmann law for the total radiation of a black body.

Consider a hollow chamber in the form of a cylinder provided with a weightless piston that moves without friction. The bottom or another part of the vessel is a black body, all other walls, and the face of the piston, being perfectly reflecting. The black end of the cylinder continuously emits and absorbs radiation. Since the radiation has a finite speed of propagation, there is always a quantity of radiation energy U present inside the hollow chamber. This chamber filled with radiation is in several respects analogous to a gas-filled vessel. An important difference, however, is that the density of the radiation energy $u = U/V$ depends exclusively on the temperature T, whereas the density of the gas can only be altered by altering the volume. The total number of photons in a hollow chamber changes with the temperature; the total number of gas molecules remains constant.

The pressure exerted by the radiation (photons) on the wall of our (gas-free) chamber is given by $p = \frac{1}{3}\mu$, according to electromagnetic theory.

---

[7]) It is usually preferable to indicate the performance of a refrigerator or heat pump by a quantity that cannot exceed unity. The concept then introduced is the *figure of merit*, this being the ratio of the actual to the theoretical (Carnot) efficiency. In this connection, see the passage in small print on p. 73 of the first article quoted under [6]), and fig. 2 of the second article.

Work can be done on the system by compressing the radiation "gas" by displacing the piston, and heat can be supplied to it by bringing the black end of the cylinder in contact with a source at higher temperature. Because $p = \frac{1}{3}u$, the first law takes the form

$$dU = dQ - \tfrac{1}{3}u\,dV.$$

The radiation in the chamber also possesses a certain entropy. This statement is statistically justified by considering the radiation as a gas composed of photons, each of which can occur in different energy states. It may be justified in an even simpler way by classical thermodynamics: if the heat source delivers a quantity of heat dQ reversibly to the cylinder, the entropy of the source drops. Since the total entropy must remain constant with a reversible transfer, the entropy S of the radiation (the photon-gas), must increase as a result of the transfer. The combination of the first and second law of thermodynamics thus gives:

$$dU = T\,dS - \tfrac{1}{3}u\,dV. \quad . \quad . \quad . \quad \text{(II, 35)}$$

where T is the temperature of the black-body radiation, i.e. the temperature of the black wall with which the radiation is in equilibrium. Employing the relation $U = Vu$ or $dU = V\,du + u\,dV$, we obtain

$$dS = \frac{V}{T}\frac{du}{dT}dT + \frac{4}{3}\frac{u}{T}dV. \quad \text{(II, 36)}$$

Now dS is a pure differential in the sense of I (a "total differential") and can therefore be written:

$$dS = \frac{\partial S}{\partial T}dT + \frac{\partial S}{\partial V}dV.$$

Consequently $$\frac{\partial^2 S}{\partial T\,\partial V} = \frac{\partial^2 S}{\partial V\,\partial T}$$

and hence, from (II, 36) we have

$$\frac{1}{T}\frac{du}{dT} = \frac{4}{3}\frac{1}{T}\frac{du}{dT} - \frac{4}{3}\frac{u}{T^2},$$

i.e. $$\frac{du}{u} = 4\frac{dT}{T}.$$

Hence

$$u = a\,T^4, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(II, 37)}$$

where a is an integration constant. This is the Stefan-Boltzmann law. As the radiation energy, i.e. the quantity of radiation impinging per second upon 1 cm^2, is proportional to the quantity of energy per cm^3, this law may also be stated: the total radiation of a black body is proportional to T^4. Thus, substituting (II, 37) for u in (II, 36), we obtain:

$$dS = 4\,aV\,T^2\,dT + \frac{4}{3}a\,T^3\,dV$$

so that $$\left(\frac{\partial S}{\partial V}\right)_T = \frac{4}{3}a\,T^3. \quad . \quad . \quad . \quad \text{(II, 38)}$$

The energy per unit volume of the radiation is thus proportional to the fourth power of the absolute temperature, and the entropy per unit volume to its cube.

ERRATA

THE THEORY AND CONSTRUCTION OF GERMANIUM DIODES

In the article by J. Z. van Wessem, The theory and construction of germanium diodes, which appeared in the February issue of this Review, the heading "The p and n concentrations in the absence of an applied voltage" on p. 218, should read: "The p and n concentrations in the *presence* of an applied voltage".

"MÜLLER" UGX APPARATUS FOR X-RAY DIAGNOSTICS

In the short notice concerning the above apparatus which appeared in our February issue (p. 237), it was stated that the patient can be rotated about an axis *parallel* to the X-ray beam. It should be pointed out that the axis of rotation is not merely parallel but *coincident* with the axis of the beam.

VOL. 16 No. 11, pp. 309-340 MAY 1955

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

REFLECTION AND IMPEDANCE MEASUREMENTS BY MEANS OF A LONG TRANSMISSION LINE

by J. C. van den HOOGENBAND and J. STOLK.

621.317.332.6

In systems employing metre (and shorter) waves, it is often necessary to match, say, an aerial to a transmission line, or a transmission line to a receiver. A method providing a direct visual indication of matching, by means of an oscilloscope, is described in this article. This method also enables the reflection factor to be deduced quite readily from the oscillograms, and hence the correct line-terminating impedance. Given this impedance, it is a relatively simple matter to determine the measures necessary to ensure correct matching.

Many methods of measuring impedances at very high frequencies are known. One of them, i.e. measurement of the unknown impedance in a bridge circuit, is used in several different forms. However, these measurements present some difficulty at frequencies higher than about 300 Mc/s (decimetre waves), owing to the fact that here the wavelength is of the same order of magnitude as the dimensions of the object on test and the length of the connecting cable between it and the measuring equipment. The effects which therefore occur at such frequencies may be attributed to the stray self-inductance and capacitance; corrections to measurements, although possible, are never very accurate.

According to another principle, the impedance can be measured from the detuning and damping effects when it is in parallel with a tuned circuit such as a lecher system (transmission line) [1]). Again, however, this method is accurate only when the impedance to be measured is either appreciably lower, or appreciably higher, than the characteristic impedance of the transmission line.

Another method, in fact an elaboration of the one just described, produces accurate results in the case of an unknown impedance of the same order of magnitude as the characteristic impedance ζ (i.e. from about 0.1 ζ to 10 ζ). Here, the voltage distribution along an untuned transmission line is determined by means of a movable detector (standing wave indicator.) From the ratio of the voltage maxima and minima (standing wave ratio), and the position of the minima, the reflection coefficient (to be defined later) can be deduced. Both components of the required impedance can then be calculated from this coefficient [2]) [3]). In many cases the object of the measurement is to determine the reflection factor itself; it indicates in how far the transmission line is terminated with an impedance equivalent to its characteristic impedance, that is, in how far these impedances are matched.

In very high frequency applications, many of the objects tested are specially designed for connection to a transmission line [4]). If the characteristic impedance of the test-object is the same as that of the transmission line, the result of the measurement will be a true picture of the behaviour of the line and the impedance combined, which is precisely the information required in practice.

[1]) J. M. van Hofweegen, The measurement of impedances, particularly on decimeter waves, Philips tech. Rev. **8**, 16-24, 1946.

[2]) J. M. van Hofweegen, Impedance measurements with a non-tuned Lecher system. Philips tech. Rev. **8**, 278-286, 1946.

[3]) H. J. Lindenhovius, The measurement of impedances at high frequencies and applications of the standing-wave indicator, T. Ned. Radiogenootschap **12**, 65-82, 1947 (in Dutch).

[4]) A. E. Pannenborg, A measuring arrangement for waveguides, Philips tech. Rev. **12**, 15-24, 1950/51.

Although measurements carried out with the standing-wave indicator produce quite accurate results, the mechanical quality of the indicator itself must conform to very stringent requirements. The method that will now be described, although perhaps less accurate, is much simpler both as regards the equipment required and the actual measurement. It resembles the method described in article [2]) in as far as the impedance to be determined is deduced from the reflection coefficient. However, the new method offers at least one considerable advantage, that is, that the reflection coefficient itself is readily determined from oscillograms; the oscillogram shows whether matching has been effected or not. Hence this method is most suitable in cases where the matching of a transmission line to, say, a signal generator or an aerial is to be adjusted or checked quickly over a wide range of frequencies.

The principle employed here for determining the absolute value of the reflection coefficient has already been described by Bayer [5]). However, he did not suggest the possibility of calculating the unknown impedance entirely from the reflection coefficient. The object of the present article is to develop the method outlined by Bauer, and to formulate the results mathematically and describe one or two practical applications.

Incident and reflected waves in a transmission line

Using the method of complex quantities, the voltage and current at an arbitrary point (y) on a transmission line (*fig. 1*) may be defined as $V(y)e^{j\omega t}$ and $I(y)e^{j\omega t}$, respectively, where ω is the angular frequency and $V(y)$ and $I(y)$ are complex quantities depending solely on the position (y), that is, independent of the time (t).

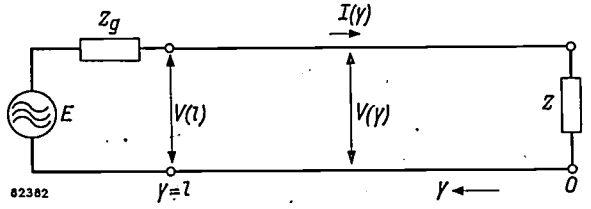


Fig. 1. Transmission line fed by a generator of e.m.f. E and internal impedance Z_g, and terminated in an impedance Z.

For $V(y)$ and $I(y)$ we have the relations

$$V(y) = A\,e^{\gamma y} + B\,e^{-\gamma y} \quad . \quad . \quad . \quad . \quad . \quad (1)$$

and

$$I(y) = \frac{A}{\zeta}\,e^{\gamma y} - \frac{B}{\zeta}\,e^{-\gamma y}, \quad . \quad . \quad . \quad (2)$$

where A and B are complex constants, independent of t and y, the precise values of which are governed by such boundary conditions as the manner in which the system is supplied and terminated. Also, γ is the propagation constant and ζ is the characteristic impedance of the transmission line.

In general, γ and ζ are complex; γ may then be written as:

$$\gamma = \alpha + j\beta, \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

where α is the attenuation constant (expressed in nepers per metre) and β the phase constant (in radians per metre). If v is the phase velocity of the waves in the transmission line and f is the frequency, we have:

$$\beta = \frac{\omega}{v} = \frac{2\pi f}{v}. \quad . \quad . \quad . \quad . \quad . \quad (4)$$

If no loss occurs in the system, $\gamma = j\beta$. The characteristic impedance ζ is then a real quantity i.e. a pure resistance; in the following it is assumed, unless otherwise stated, that there is no loss in the system (zero attenuation).

From (1) and (2), we have at the position of the unknown load impedance Z, i.e. at $y = 0$:

$$V(0) = A + B \quad \text{and} \quad I(0) = \frac{A-B}{\zeta}.$$

Also,

$$Z = \frac{V(0)}{I(0)}, \quad \text{hence} \quad Z = \frac{A+B}{A-B}\,\zeta,$$

or

$$B = A\,\frac{Z-\zeta}{Z+\zeta} = Ar,$$

where

$$r = \frac{Z-\zeta}{Z+\zeta} = |r|\,e^{j\varphi} \quad . \quad . \quad . \quad . \quad . \quad (5)$$

is complex.

Substituting $B = Ar$ in (1) and (2), we have:

$$V(y) = A(e^{\gamma y} + r\,e^{-\gamma y}) \quad . \quad . \quad . \quad . \quad (6a)$$

and

$$I(y) = \frac{A}{\zeta}(e^{\gamma y} - r\,e^{-\gamma y}). \quad . \quad . \quad . \quad (6b)$$

It is seen from these equations that the voltage distribution in the line may be considered as the superposition of an outgoing wave $Ae^{\gamma y}$ and a returning wave $Are^{-\gamma y}$; the latter may be imagined to arise from a partial reflection of the outgoing wave by the load impedance Z. Since r is a measure of this reflection, it is termed the *reflection coefficient*. From (5), r is determined uniquely by

[5]) J. A. Bauer, Special applications of ultra-high frequency wide-band sweep generators, R. C. A. Rev. 8, 564-575, 1947.

the ratio of the terminating impedance Z to the characteristic impedance ζ. Given $Z = \zeta$, then, $r = 0$. In this case, we have only an outgoing wave ("travelling wave"). If $Z = 0$ (end short-circuited), $r = -1$; if $Z = \infty$ (open-circuit end), $r = +1$. In both cases the reflected wave has the same amplitude as the outgoing wave, i.e. there is *total reflection*. Thus a "standing wave" is produced.

Reasons for the undesirability of reflection

Provided that the condition that the impedance Z_g of the generator (fig. 1) shall equal ζ is satisfied, the power carried by the outgoing wave will be $(\frac{1}{2}E)^2/\zeta$, regardless of the value of the load impedance Z. From this "available power" $P_1 = (\frac{1}{2}E)^2/\zeta$, a part $P_2 = (\frac{1}{2}|r|E)^2/\zeta$ is reflected by Z and carried back by the reflected wave. Accordingly, the ratio:

$$\eta = \frac{P_1 - P_2}{P_1} = 1 - |r|^2$$

is the efficiency of the energy transfer.

The square root of η is equal to the ratio of the voltage across the resistive part (R) of the load impedance Z to the voltage that would be obtained across the same resistance R with proper matching, that is, if all the energy were concentrated in R. Hence $\sqrt{\eta}$ may be termed the "efficiency" of the voltage transfer. It will be seen from *fig. 2*, which shows η and $\sqrt{\eta}$ plotted against $|r|$, that in general the reflections should be minimized by a proper match.

There are even stronger reasons to minimize reflections. In the case of a television receiver, for example, the cable between set and aerial must be matched at *both* ends; otherwise, reflected waves, as well as the desired "principal wave", will enter the receiver. These reflected waves would distort abrupt black to white transients in the picture.

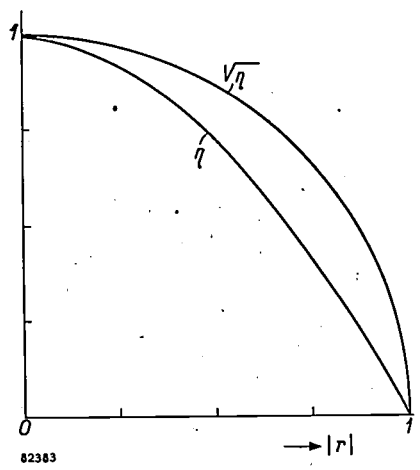


Fig. 2. Efficiency η of the energy transfer, and "voltage efficiency" $\sqrt{\eta}$, of a transmission line, plotted against the modulus $|r|$ of the reflection coefficient. High values of η and $\sqrt{\eta}$ are possible only when $|r|$ is small.

Smith's chart

Equation (5) can be appreciably clarified with the help of a Smith chart [6]) (*fig. 3*). Here, r is plotted in the complex plane. Given $Z = R + jX$, i.e. $Z/\zeta = R/\zeta + jX/\zeta$, the locus described by r when X/ζ varies and R/ζ remains constant can be determined with the aid of equation (5). For each value of R/ζ a circle is obtained whose centre lies on the real axis and which passes through the point $+1$. We thus get a system of circles all passing through the point $+1$. If X/ζ is taken as the parameter and R/ζ as the variable, another system of circles orthogonal to the first system will be obtained.

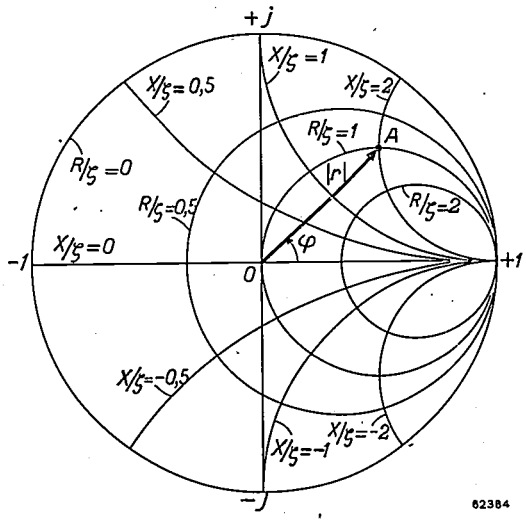


Fig. 3. Smith chart, showing the two orthogonal families of circles, the one referring to constant R/ζ and the other to constant X/ζ. A is the point of intersection of the circles corresponding to $R/\zeta = 1$ and $X/\zeta = 2$ (R being the resistance, and X the reactance of the terminating impedance Z from fig. 1). Vector OA is the reflection coefficient r with modulus $|r|$ and argument φ.

Now, for a given value of ζ, the point associated with any impedance $Z = R + jX$, at which the circle corresponding to R/ζ cuts that corresponding to X/ζ, can be established. This point determines the modulus $|r|$ and the argument φ of the reflection factor r.

Since $|r| \leqq 1$, the entire diagram is contained within a circle of radius unity. The dimensionless parameters R/ζ and X/ζ enable the diagram to be

[6]) P. H. Smith, An improved transmission-line calculator, Electronics **17**, 130-133 and 318-325, Jan. 1944. A similar diagram was worked out independently during the war by J. M. van Hofweegen, who describes it in his article [2]), pages 283-284.

adapted to any characteristic impedance ζ; the reading becomes inaccurate only at very low or very high values of these parameters.

Many results derived in the theory of transmission lines can be illustrated by means of the Smith chart. For example, the impedance of a line terminated in a given impedance $Z = R + jX$ can be determined as follows (*fig. 4a*). If this line

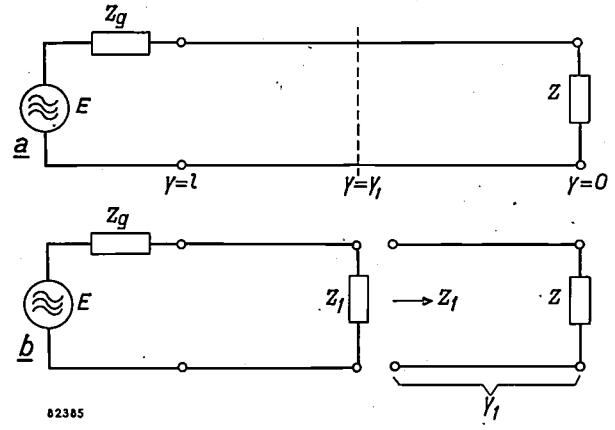


Fig. 4. *a*) Transmission line as shown in fig. 1. *b*) A certain length y_1 of the system is cut off and replaced by the equivalent impedance (Z_1).

be cut at a certain point $y = y_1$ and there terminated with an impedance matching the input impedance Z_1 of the piece cut off (fig. 4*b*), the outgoing and reflected waves in the remainder of the line will not change. Hence the amplitudes the outgoing and reflected waves at point $y = y_1$ will be the same after as before the line is cut, viz. (from (6a)) $Ae^{\gamma y_1}$ and $Are^{-\gamma y_1}$ respectively. In the case to which fig. 4*b* refers, then, a reflection coefficient of

$$r_1 = \frac{Are^{-\gamma y_1}}{Ae^{\gamma y_1}} = re^{-2\gamma y_1}. \quad . \quad . \quad . \quad . \quad (7)$$

is required at point $y = y_1$. From (3) and (4), assuming the line to be without loss, we may write for formula (7):

$$r_1 = re^{-j4\pi f y_1/v}. \quad . \quad . \quad . \quad . \quad . \quad (8)$$

Accordingly, r_1 is determined by rotating the vector r in the Smith chart through an angle $4\pi f y_1/v$ (*fig. 5*). The real and imaginary parts of impedance Z_1, seen at the input of the cut-off portion of the line, are then read from the families of circles associated with constant R/ζ and constant X/ζ.

In this way we find that, say, at $\zeta = 100$ ohms and $Z = 100 + j \times 100$ ohms (that is, R/ζ and X/ζ both unity, point A), the modulus of the reflection factor r is $|r| = 0.447$ and its argument is $\varphi = 63°$. If the line has a length l equal to $^1/_4$ wavelength ("$\frac{1}{4}\lambda$-line"), the angle of rotation $4\pi f l/v$ is π. The reading from fig. 5 corresponding to this value (point C) is: $R/\zeta = \frac{1}{2}$ and $X/\zeta = -\frac{1}{2}$; therefore $Z_1 = 50 - j \times 50$ ohms.

A similar method is employed to determine the impedance Z_l of a cable terminated with impedance Z, considered as a load on the voltage source at the input. Let us assume that this source is a generator of e.m.f. E and internal impedance Z_g (fig. 1). The vector r in the diagram must then be rotated through an angle $4\pi f l/v$. If $Z_g = \zeta$, the voltage $V(l)$ on the input terminals of the cable is:

$$V(l) = \frac{Z_l}{Z_l + \zeta} E = \frac{\frac{1}{2}Z_l + \frac{1}{2}\zeta + \frac{1}{2}Z_l - \frac{1}{2}\zeta}{Z_l + \zeta} E$$
$$= \tfrac{1}{2}(1 + r_l)E, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

where, from (5), $r_l = (Z_l - \zeta)/(Z_l + \zeta)$ is the reflection factor at the input of the cable (where $y = l$). For r_l, from (8) we may write

$$r_l = re^{-j4\pi f l/v},$$

therefore

$$V(l) = \tfrac{1}{2}E(1 + |r|\, e^{j\varphi} \cdot e^{-4\pi f l/v}).$$

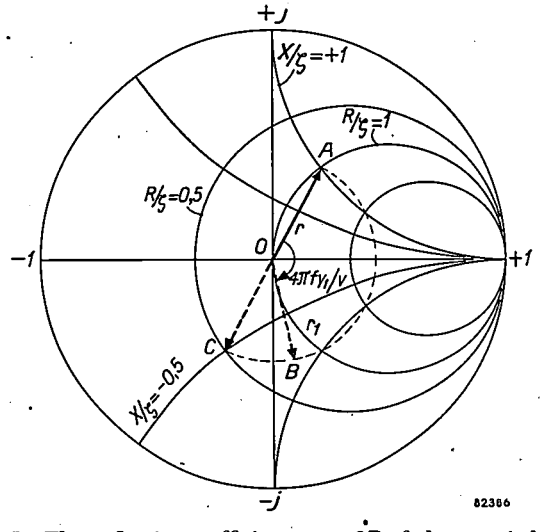


Fig. 5. The reflection coefficient $r_1 = OB$ of the remainder of the transmission line, shown in fig. 4*b*, is determined by rotating vector $r = OA$ in the Smith chart through an angle $4\pi f y_1/v$.

It is seen from (9) that vector PQ, representing voltage $V(l)$ in the diagram (see *fig. 6a*), arises from the summation of constant vector PO ($= \frac{1}{2}E$) and vector OQ ($= \frac{1}{2}r_l E$). With a suitable choice of scale, $\frac{1}{2}E$ can be made equal to the radius of the diagram.

Any increase in the frequency f will be indicated by a rotation of Q around the circle with radius $\frac{1}{2}|r_l|E$, from which the dependence of the input voltage $|V(l)|$ on the frequency can be deduced (fig. 6*b*).

It will be seen from the above that the amplitude of the outgoing wave ($\frac{1}{2}E$) does not depend upon the reflection taking place at the end. However, it should be noted that this does not hold good when $Z_g \neq \zeta$, since the wave reflected at the end is then reflected a second time by the voltage source, and so contributes to the outgoing wave. In this case the calculations are complex, and are in fact entirely beyond the scope of this article. However, one or two results will be referred to later.

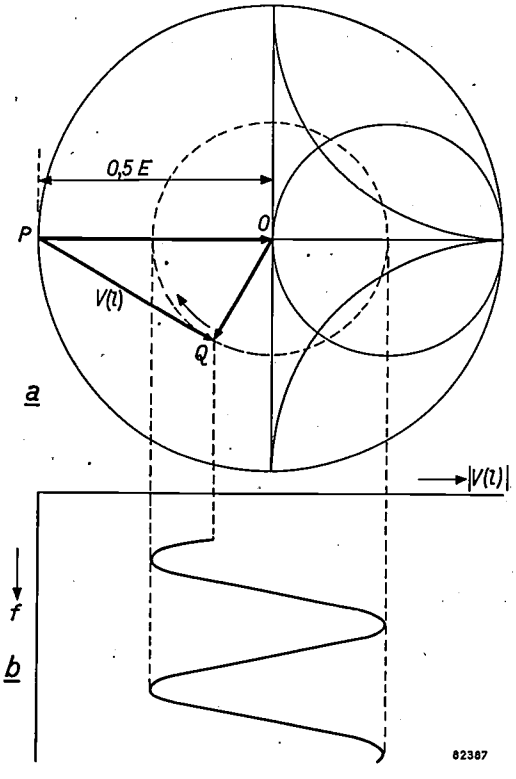


Fig. 6. *a*) $V(l) = \frac{1}{2}(1 + r_l)E$ as demonstrated in the Smith chart. When the frequency increases, Q moves in the clockwise direction around the dotted circle with radius $\frac{1}{2}|r_l|E$. *b*) The modulus of $V(l)$, corresponding to line PQ in (*a*), plotted against the frequency f.

The measurement of reflections

The method described in previous articles [2] [3] involved a fixed frequency and a movable detector. Here, a variable frequency and a stationary detector are employed.

The modified method now to be described is specially designed to facilitate matching over a wide range of frequencies. Other variants of the method more suitable where accuracy is preferred to ready interpretation and speed, will be discussed at the end of the this article.

A block diagram of the measuring equipment is shown in *fig.* 7.

Here, a transmission line whose characteristic impedance is known, is connected, with a detector in parallel, to a "wobbulator", that is, a signal generator whose frequency varies periodically, in this case with the mains frequency. The transmission line is either a coaxial or an unshielded twin cable, many times longer than the mean wavelength. It is terminated with the unknown impedance Z. The equipment also includes an oscilloscope. The time-base voltage of this instrument is synchronized to the frequency modulation of the wobbulator, so that the abscissa of the oscillogram is the frequency co-ordinate. The detector produces a D.C. voltage fluctuating at low frequency; the A.C. component of this voltage, after amplification, produces the vertical deflection. A diagram to show the type of oscillogram obtained with this system is given in fig. 6*b*.

With a circuit in which the signal to produce the vertical deflection is present only during the forward trace, the horizontal axis of the oscillogram will be described during the return trace.

The oscillograms obtained in one or two simple cases will now be described.

a) The impedance Z is a pure resistance R. Then:

$$r = \frac{R - \zeta}{R + \zeta},$$

Hence $\varphi = 0$ when $R > \zeta$, and $\varphi = \pi$ when $R < \zeta$. Accordingly, the head of vector r is always on the real axis, at -1 when $R = 0$, at $-\frac{1}{2}$ when $R = \frac{1}{3}\zeta$, at 0 when $R = \zeta$, at $+\frac{1}{2}$ when $R = 3\zeta$ and at $+1$ when $R = \infty$ (see *fig.* 8*a*).

When the frequency varies, point Q in diagram 6*a* describes a circle about 0, the radius of this circle being governed by R. The length of line PQ is then the modulus of vector $V(l)$. The oscillogram of $|V(l)|$ as a function of $(4\pi l/v)f$ for the above-mentioned values of R will then be as shown in fig. 8*b*.

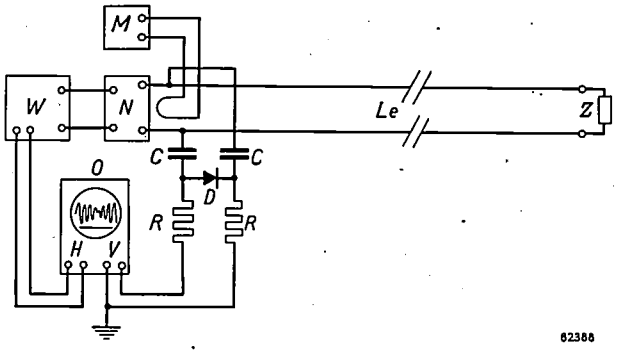


Fig. 7. Circuit employed to determine the reflection coefficient of an unknown impedance (Z) terminating the line Le (ζ = 300 ohms). W wobbulator, N matching network, V detector, O oscilloscope (horizontal deflection synchronized with the frequency variation of the wobbulator, vertical deflection proportional to the A.C. output voltage of the detector), M marker (described later), C isolating capacitors (100 pF), R resistors (10,000 ohms), to prevent H.F. voltage from entering the oscilloscope.

b) The impedance Z is pure a reactance jX. Then:

$$r = \frac{jX - \zeta}{jX + \zeta}.$$

Hence

$$|r| = 1, \qquad \varphi = \pi - 2\tan^{-1}\frac{X}{\zeta}.$$

The fact that $|r| = 1$, shows that in the case of a purely reactive terminating impedance the head of vector r is always on the periphery of the Smith chart (fig. 8*c*). Fig. 8*d* represents the oscillogram of $|V(l)|$ as a function of $(4\pi l/v)f$ for three values of X. Since only the argument (φ) of r varies with X, a change of X merely produces a horizontal displacement of the oscillogram.

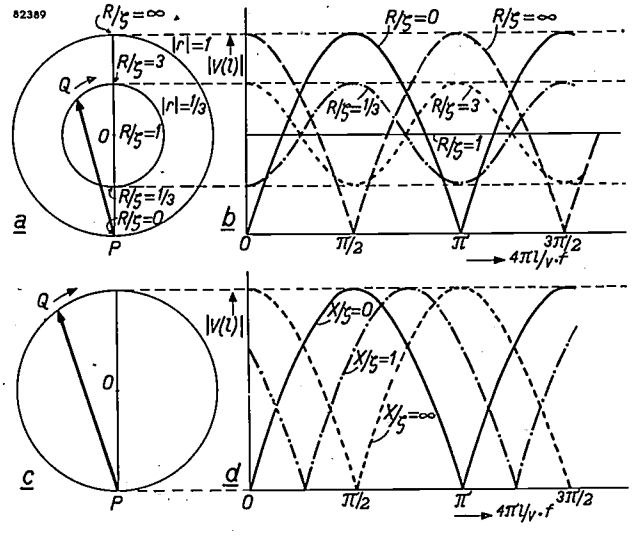


Fig. 8. *a-b*) The terminating impedance (fig. 7) is a pure resistance R. In (*a*), point Q lies on the circle with radius $\frac{1}{2}$ when $R/\zeta = 1/3$ or $R/\zeta = 3$, and travels round this circle as the frequency varies: (*b*) shows length $PQ = |V(l)|$ plotted against f (dotted line and chain-dotted line). When $R = 0$ or $R = \infty$, Q lies on the circle with radius unity; $|V(l)|$ then varies in accordance with the curve drawn as a full line, or with that drawn as a broken line. When $R/\zeta = 1$, Q coincides with O and $|V(l)|$ is constant (no reflection).
c-d) The terminating impedance (fig. 7) is a pure reactance jX. Q lies on the circle with radius unity whatever the value of X; a variation of X merely produces a horizontal displacement of $|V(l)|$ as a function of f.

Determination of the reflection coefficient from the oscillograms

The determination of the reflection coefficient r from the oscillograms is based on the following considerations:

a) From (9), $|V(l)|$ fluctuates between $\frac{1}{2}E(1 + |r|)$ and $\frac{1}{2}E(1 - |r|)$. Accordingly, the oscillogram shows a ripple whose amplitude a_r (peak to peak, *fig. 9a*) is proportional to the variation of $|V(l)|$; hence:

$$a_r = g \cdot \tfrac{1}{2}E\{(1 + |r|) - (1 - |r|)\} = gE|r|.$$

The proportionality factor (g) is determined by introducing a known reflection coefficient, e.g. total reflection ($|r| = 1$, fig. 9*b*), say, by leaving the end of the line open. The ripple amplitude (a_0) is then gE, so that:

$$|r| = \frac{a_r}{a_0} \qquad (10)$$

Accordingly, the ratio of the two ripple amplitudes is equal to the modulus of the reflection coefficient

b) The oscillogram is periodic. In the event of an increase $(\Delta f)_0$ in f, such that $4\pi(\Delta f)_0 l/v = 2\pi$, so that

$$(\Delta f)_0 = \frac{v}{2l}, \qquad (11)$$

$|V(l)|$ will return to the original value. Given a long transmission line, then, the frequency variation will be small.

c) The interval between the minima in fig. 8*a* and *b* is $(\Delta f)_r = (\varphi/2\pi)(\Delta f)_0$. Hence

$$\varphi = \frac{(\Delta f)_r}{(\Delta f)_0} 2\pi. \qquad (12)$$

Total reflection can be obtained not only by open-circuiting the end of the line, but also by short-circuiting it. In the latter case, π should be added to the result of (12).

It is seen, then, that the ratio of the two differences in frequency to be read from the oscillograms, corresponds to the argument of the reflection coefficient. Hence the reflection coefficient itself can be fully established by means of (10) and (12) combined.

The derivation of the unknown impedance, Z, from the Smith chart with the aid of $|r|$ and φ has already been described.

However, the method that will now be considered is more accurate than the mere reading of $(\Delta f)_r$ and $(\Delta f)_0$ from oscillograms. An auxiliary oscillator

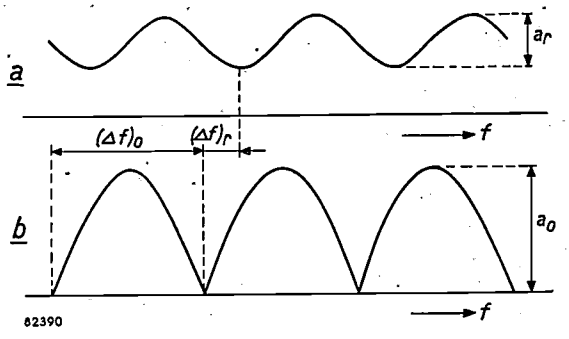


Fig. 9. Oscillograms obtained with the circuit shown in fig. 7, (*a*) employing an arbitrary terminating impedance. (*b*) with the end of the test line open. The ratio of the ripple amplitudes (a_r and a_0) is the modulus of the reflection coefficient, and 2π times the ratio of the differences in frequency $(\Delta f)_r$ and $(\Delta f)_0$ is the argument of this coefficient (φ).

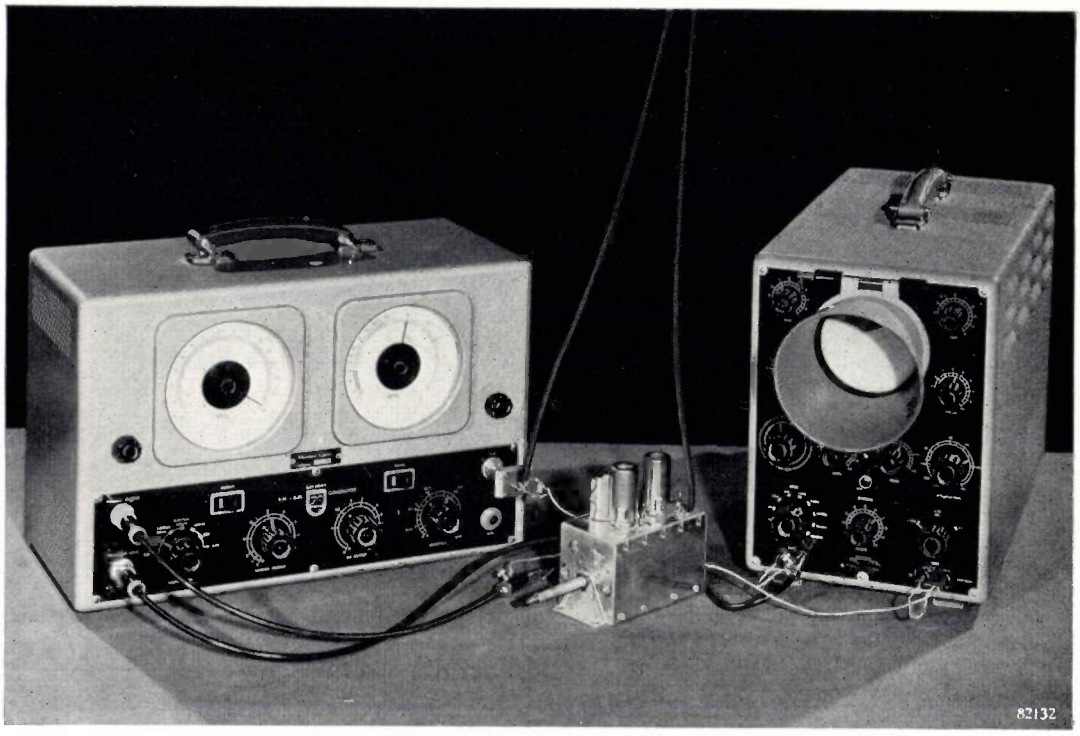


Fig. 10. Equipment for the measurement of reflections. From left to right: the wobbulator (with built-in marker), type GM 2889, the object under test (here a channel-selector for a TV receiver) and the cathode-ray oscilloscope. Note the input and output ends of the 60 metre unshielded twin lead employed as the transmission line.

(M in fig. 7), with facilities for varying and reading the frequency accurately, is coupled loosely to the measuring system. This oscillator is then tuned to a frequency within the frequency variation of the wobbulator. Each time the wobbulator frequency passes that of the auxiliary oscillator, a beat occurs producing a small irregularity in the line in the oscillogram (hence the auxiliary oscillator is termed a "marker"). The marker enables $(\Delta f)_r$ and $(\Delta f)_0$ to be determined very accurately as differences between consecutive adjustments of the marker.

The measuring equipment is shown in *fig. 10*.

Effect of attenuation

The impedance to be measured, Z, is generally dependent on the frequency; hence $|r|$ and φ both vary with the frequency. Employing the method described here, it is necessary to vary the frequency over a range comprising several intervals $(\Delta f)_0$, and for the sake of accuracy, $|r|$ must be prevented from varying unduly within any one of these intervals. This condition is easier to satisfy the smaller the interval. From (11), however, it is necessary to increase the length (l) of the transmission line according as $(\Delta f)_0$ decreases, and the attenuation, so far ignored, increases with the length of the line.

We must therefore now consider the effect of attenuation on the measurements.

If the attenuation is α nepers per metre, the amplitude of the reflected wave on reaching the detector will be a factor of $e^{-2\alpha l}$ lower than it would would be without attenuation. Owing to this relatively lower amplitude, the voltage $|V(l)|$ will invariably remain above zero, even in the case of total reflection (compare *fig. 11a*, relating to $\alpha = 0$, with fig. 11*b*, which refers to $\alpha > 0$). Accordingly, it is necessary to multiply $|r|E$ and E, in the expressions of a_r and a_0, respectively, by $e^{-2\alpha l}$; at the same time, the ratio a_r/a_0, which, from

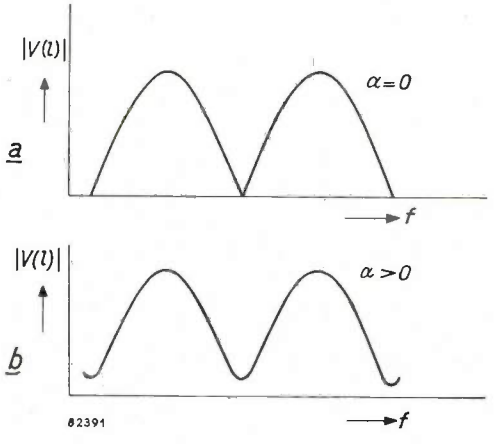


Fig. 11. $|V(l)|$, plotted against f in the case of total reflection *a*) with linear deflection, but without attenuation, *b*) with linear deflection and attenuation. In case (*b*), $|V(l)|$ remains above the zero line.

(10), is equal to $|r|$, remains unchanged. In principle then, attenuation will not give rise to errors in the measurement, at least as regards $|r|$.

However, minor irregularities are invariably superimposed on the ripple produced by reflection. They may arise partly from the fact that the generator voltage varies to a certain extent with frequency, and partly from unavoidable non-uniformity in the transmission line (for example, in the case of unshielded twin lead, the supports; the effect of points of contact with the supports are minimized by suspending the cable from thin cords. The effect of insulators will be discussed later).

Such irregularities affect the result all the more, the stronger the attenuation; this imposes a limit on the effective length of the cable. A suitable length is 60 metres.

Other effects

Effect of mismatch between generator and transmission line

So far it has been assumed that the internal impedance (Z_g) of the generator (wobbulator) matches the characteristic impedance (ζ) of the measuring system. Let us now consider the variation of voltage $|V(l)|$ as a function of frequency when Z_g does not match ζ. The actual calculation of this variation is so complex that it is beyond the scope of this article, but one or two of the results are shown in *fig. 12*.

Fig. 12*a* depicts $|V(l)|$ plotted against f for a real value of Z_g (viz. 2ζ), and fig. 12*b* shows a similar curve for a complex value of this impedance ($Z_g = \zeta/(1-j)$), for $|r| = 0.33$, 0.80 and 1.0 in both cases.

When Z_g is purely resistive (fig. 12*a*), the positions of the maxima and minima do not depend very much upon $|r|$, and the value, represented by $|r'|$, deduced from the ripple ratio a_r/a_0, does not differ appreciably from the correct value $|r|$ (fig. 12*a* gives 0.30 and 0.77 instead of 0.33 and 0.80, respectively).

However, as soon as Z_g acquires a reactive component, the maxima and (to a smaller extent) the minima are displaced. This is owing to the fact that in the case of, say, a capacitative Z_g, the voltage $|V(l)|$ does not reach a maximum when the input impedance of the cable is real, but does do so in the case of an inductive input impedance, where a build-up takes place.

To calculate φ from equation (12), then, it is most convenient to deduce the difference in frequency from the interval between minima, since they are more stable than the maxima. The effect of the displacement can be further reduced by adding to the complex impedance Z_g a reactance such that the two combined assume a real value (the capacitative impedance of the detector in parallel with Z_g is always a part of Z_g).

Effect of the detector characteristic

Fig. 13 shows the relationship between the amplitude (V) of the high-frequency voltage, and the detected D.C. voltage (U), in the case of a germanium diode [7]) (this being more suitable than a vacuum diode by virtue of the fact that its capacitance is lower). Usually, owing to the non-linearity of this characteristic, the variation of voltage U does not entirely correspond to the changes in V.

With the method considered here, the oscilloscope is employed to observe the fluctuation of U, instead of $V(l)$.

[7]) See Philips tech. Rev. **16**, 225-232, 1954/55 (No. 8).

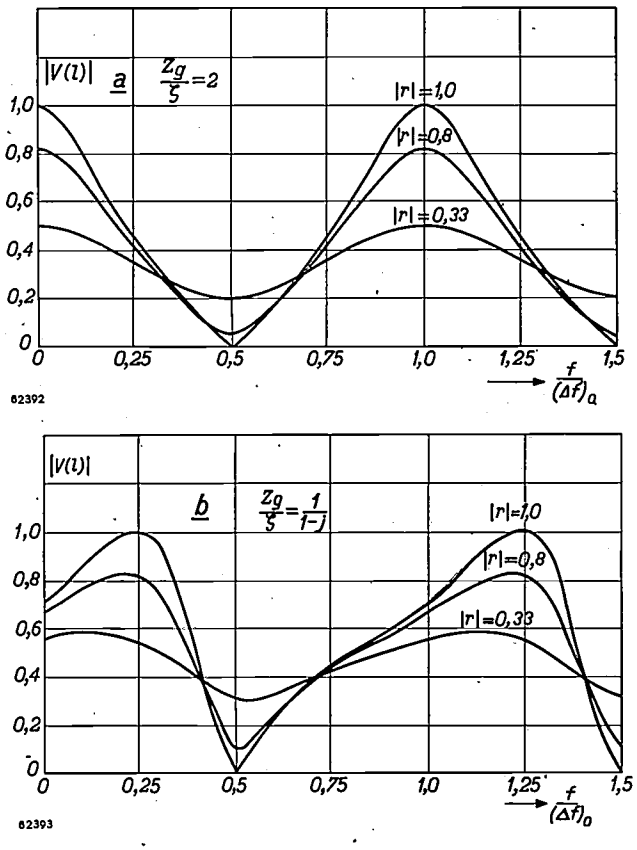


Fig. 12. $|V(l)|$ plotted against f when Z_g does not equal ζ. Curves (*a*) refer to $Z_g = 2\zeta$ (real), and curves (*b*) to $Z_g = \zeta/(1-j)$; in both cases $|r|$ has the values 0.33, 0.80 and 1.0. In (*a*) the position of the maxima and minima is virtually independent of $|r|$. In (*b*), the position of the maxima is largely determined by $|r|$; for the minima this is not so. It is assumed that E is unity.

Here, then, the ratio of the ripple amplitudes of U as derived from the oscillogram will differ from that of the ripple amplitudes of $|V(l)|$, which is required for the determination of $|r|$ from (10).

The error arising from this difference in ratio has been calculated in one or two cases. With linear detection, and with

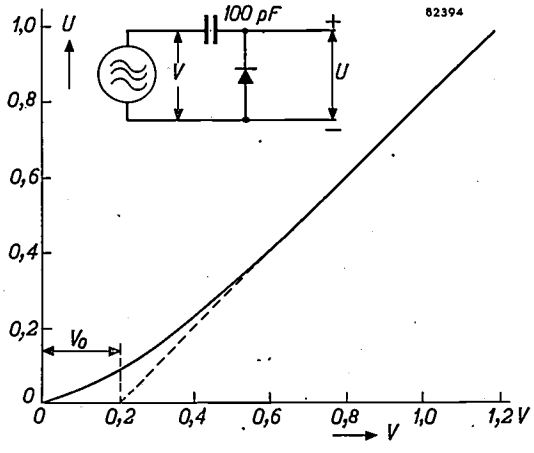


Fig. 13. Relationship between the detected D.C. voltage (U) and the amplitude (V) of the high-frequency A.C. voltage, for a given germanium diode. The value (V_0) at which the straight part of the characteristic (produced) would cut the V axis is about 0.2 V.

square-law detection there is no error. These cases correspond roughly to $E \gg V_0$ and $E \ll V_0$, respectively (for V_0 see fig 13). However, the output voltages of most wobbulators (0.1-0.5 V) are of the same order of magnitude as V_0; hence we find in the characteristic of fig. 13 a deviation corresponding to line b in *fig. 14*.

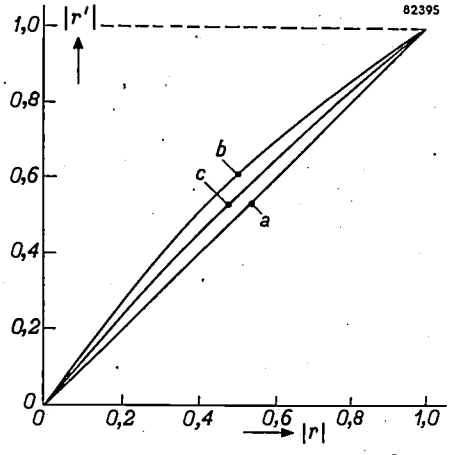


Fig. 14. Given the detection characteristic shown in fig. 13, and with E of the same order of magnitude as V_0, the measured value of the reflection coefficient will be too high ($|r|'$); note the deviation of curve b with respect to the straight line at an angle of 45°. Attenuation reduces this error, as will be seen from the curve (c) referring to $e^{-2al} = 0.8$.

Accordingly, it is expedient to increase the attenuation of the line, since with total reflection, the minima of $|V(l)|$ then remain above zero (fig. 11b) and therefore cover a relatively smaller portion of the bend in the detection characteristic. If e^{-2al} is say, 0.8, the deviations will be halved (curve c in fig. 14).

Applications

Matching a tuned circuit to a cable

The impedance Z of a parallel circuit (L, C, and R in parallel) may be written as:

$$Z = \frac{R}{1 + j\beta_0 Q},$$

where $\beta_0 = \frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}$ (with $\omega_0^2 = 1/LC$) is the detuning (here indicated by β_0 to distinguish it from the phase constant β), and Q the quality factor, i.e. $R\sqrt{C/L}$. From (5), when the circuit is connected to the transmission line, the reflection coefficient is:

$$r = \frac{\frac{R}{1 + j\beta_0 Q} - \zeta}{\frac{R}{1 + j\beta_0 Q} + \zeta}.$$

Hence the modulus of r is:

$$|r| = \sqrt{\frac{(R - \zeta)^2 + \beta_0^2 Q^2 \zeta^2}{(R + \zeta)^2 + \beta_0^2 Q^2 \zeta^2}}. \quad . \quad . \quad (13)$$

If the frequency be so increased or reduced that ω passes resonance (ω_0), β_0^2 will drop to zero, and then rise again: from (13), $|r|$ has a minimum $(R - \zeta)/(R + \zeta)$ at resonance. When circuit and cable are matched ($R = \zeta$), this minimum is zero.

As we have already seen, the voltage $V(l)$ at the input of the transmission line is equivalent to the vector sum of the stationary vector $\frac{1}{2}E$ and the rotating vector $|r| \cdot \frac{1}{2}E$. As we have seen above, $|r|$ varies during the frequency modulation cycle, that is, decreases to begin with, but increases when the frequency goes beyond resonance. In response to this variation, the end of vector $V(l)$ first describes an inward spiral, and then an outward spiral. Accordingly, the oscillogram shows, from left to right, an initial decrease in amplitude followed by an increase. When the minimum is zero, line and circuit are matched; this enables us to see at a glance whether they are matched or not.

In the case of a mismatch, matching can be effected by connecting the line to a tapping in the circuit (*fig. 15a*). It is then necessary to substitute R/τ^2 for R in formula (13), τ being the transformation ratio. Fig. 15 depicts $|r|$ plotted against $\beta_0 Q$ for various values of τ; matching occurs where $\tau = \tau_{opt} = \sqrt{R/\zeta}$.

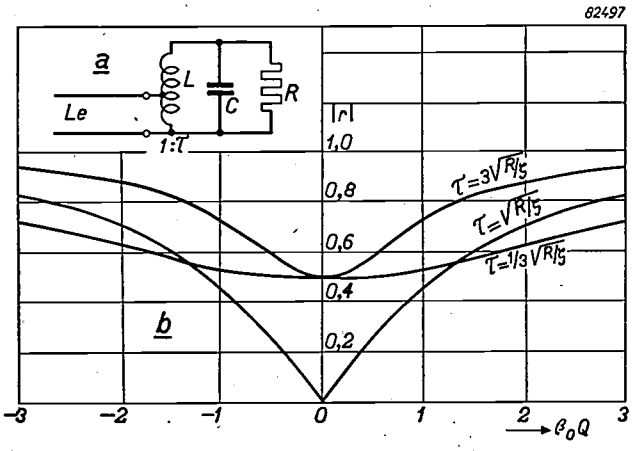


Fig. 15. *a*) The impedance of a resonant circuit (L, C, R) can be matched to that of the transmission line (Le) by employing a suitable transformation ratio (τ).
b) Relationship between $|r|$ and the detuning β_0 (the quantity actually plotted is $\beta_0 Q$), for various values of the transformation ratio (τ). The optimum ratio (τ_{opt}) is $\sqrt{R/\zeta}$.

Fig. 16 shows the results of experiments with three different tappings on the coil of a circuit tuned to about 560 Mc/s. Fig. 16b, where the minimum amplitude is about zero, is the closest approximation to ideal matching. The upper three envelopes in fig. 16 correspond to the curves shown in fig. 15b [8]).

[8]) It should be noted that although this variation is related to the quality factor Q, the value of Q cannot be deduced from it as readily as from a resonance curve.

Both components of the impedance are derived from the measured values of the reflection coefficient. It can be shown (although we must omit the proof) that in the Smith chart the measuring points

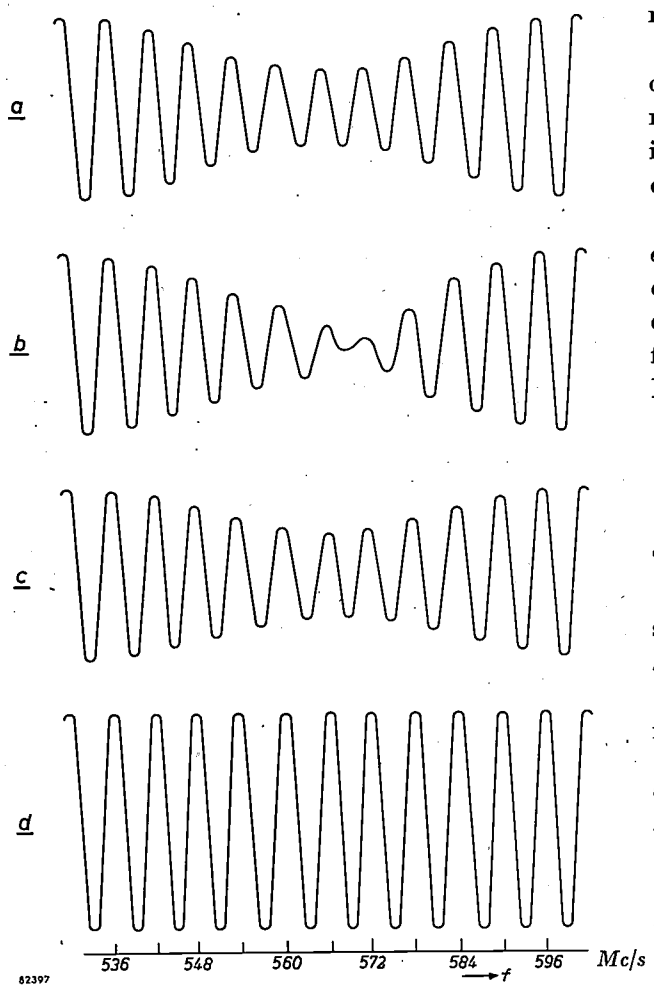


Fig. 16. Experimental results obtained in a case similar to that illustrated in fig. 15*a*. Here, the circuit is tuned to about 560 Mc/s.
a) τ too high, *b*) matching virtually correct, *c*) τ too low. In case (*d*), the end of the transmission line is short-circuited.

corresponding to the different frequencies should form a circle. The extent to which this condition is satisfied will be seen from *fig. 17*.

Measuring cable characteristics

The method considered here is also very effective as a means of measuring cable characteristics.

Firstly, let us consider the characteristic impedance ζ. This can be measured by terminating the cable with such a resistance that the oscillogram has a constant amplitude. Although in this case the generator impedance Z_g cannot be matched to ζ (this being unknown), it does not detract from the accuracy of the measurement, since, provided that $R = \zeta$, the oscillogram will show a constant amplitude even if $Z_g \neq \zeta$. At the frequencies involved, continuously variable resistors of the required order of magnitude (some tens or hundreds of ohms) are not sufficiently free of self-inductance and capacitance; therefore, a number of fixed resistors of known value are employed instead.

The reflection coefficient r, is measured with the cable connected to each resistor in turn, and the resistance corresponding to $|r| = 0$ is then determined by interpolation; this resistance equals the characteristic impedance.

Another cable-constant, $k = v/c = 1/\sqrt{\varepsilon_{rel}}$ (where c is the velocity of light and ε_{rel} the relative dielectric constant), may be defined as the factor expressing the wavelength in in the cable as a fraction of the corresponding wavelength in air. From (11) we have:

$$k = \frac{2l(\Delta f)_0}{c}.$$

$(\Delta f)_0$ can be determined accurately by means of the "marker" already described.

By this method it was found that for coaxial and shielded twin cables, with solid polythene insulation $k = 0.65$. For unshielded ribbon type twin lead (likewise insulated with polythene), of characteristic impedance $\zeta \doteq 300$ ohms, a usual value is $k = 0.79$.

Lastly, let us consider the attenuation in a cable. The amplitude of the oscillogram is proportional to $E(1 - e^{-2al})$ when the end of the cable is short-circuited, and to E when the cable is omitted altogether. From these amplitudes, a can be calculated.

Determining the effect of terminal insulators

TV and F.M. receivers are often connected to a roof-aerial by means of an unshielded polythene twin lead, which must be secured at various points by means of insulators. The insulators impair the uniformity of the cable and so cause reflections, which may affect the reception of the signal. Even a relatively small number of such irregularities may give rise to a most confused situation; hence reflections should be minimized as far as possible.

The measurements that will now be described illustrate a simple case. A long ribbon type twin lead is terminated with a resistor matching it as far as possible and thus ensuring a low reflection coefficient. Now, curve *a* in *fig. 18* shows measured values of $|r|$ plotted against the frequency (f) up to about 1000 Mc/s. An insulator is attached to the cable, close to the terminating resistor. Measured values of $|r|$ are again plotted as a function of f (the total reflection required for this purpose being procured

by short-circuiting the cable just beyond the insulator). The result shows a considerable increase in the reflection coefficient (curve *b* in fig. 18). In the case considered it was shown that the effect of the insulator was equivalent to a capacitance of 0.13 pF between the cable cores at the point of attachement.

Another insulator is now fixed to the cable, about 7.5 cm from the first. This roughly doubles the reflection coefficient at the lower frequencies (about 100 to 300 Mc/s, curve *c* in fig. 18), but reduces it considerably at frequencies in the region of 800 Mc/s. This is understandable in view of the fact that at 800 Mc/s the wavelength in the cable is $kc/f = 0.8 \times 3 \times 10^8/(800 \times 10^6) = 0.30$ m; hence the distance travelled by the wave from one insulator to the other and back again is precisely half a wavelength and the waves will therefore cancel one another.

Variants of the method described

The wobbulator and oscilloscope method is very quick and convenient in that it provides an immed-

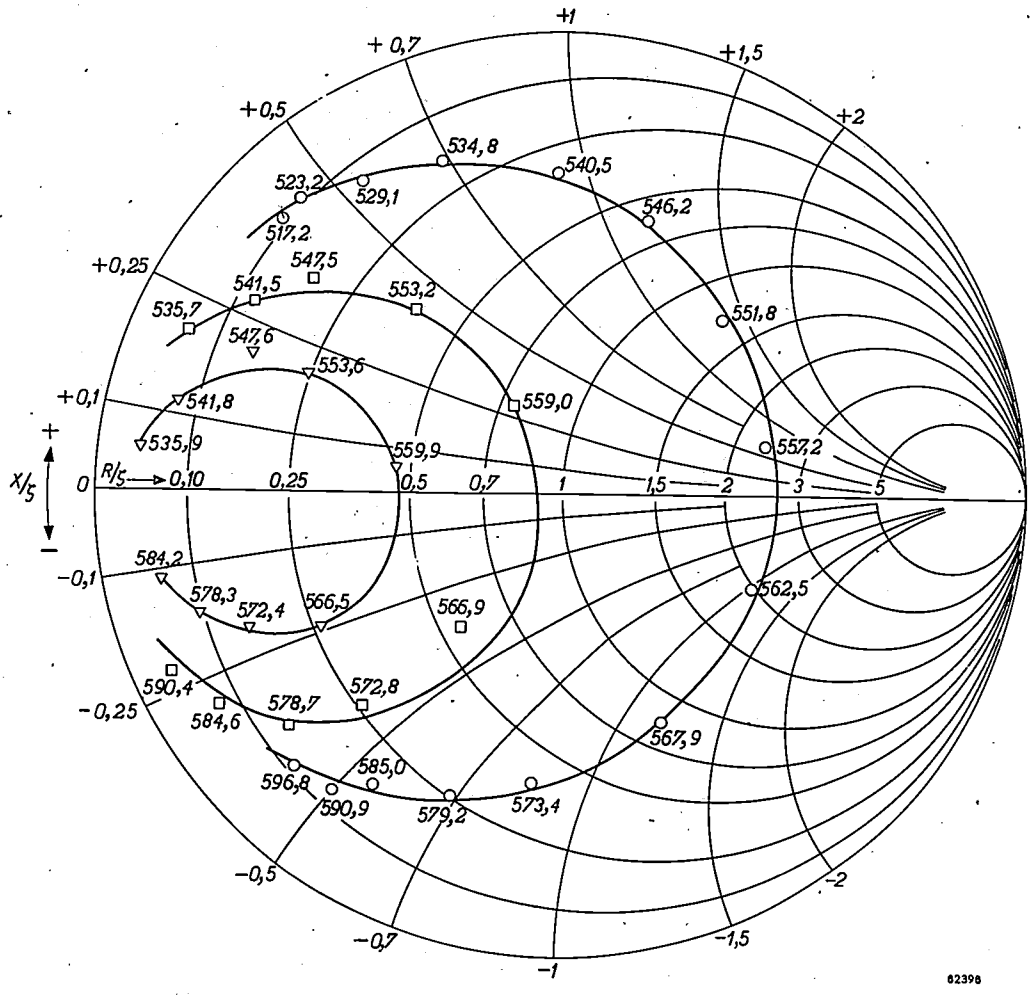


Fig. 17. Impedances as measured in the cases illustrated in fig. 16*a*, *b* and *c*, plotted in Smith's diagram. Theoretically the curves referring to a constant transformation ratio (τ) should be circles. The appropriate frequency (in Mc/s) is indicated at each point.

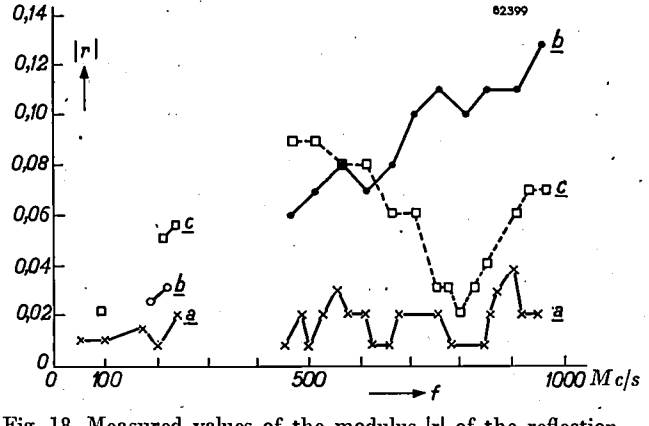


Fig. 18. Measured values of the modulus $|r|$ of the reflection coefficient in a unshielded polythere twin lead with $\zeta = 300$ obms, plotted against the frequency *f*. The cable is matched fairly well by a resistance. *a*) Without deliberately introduced irregularities, *b*) with one insulator close to the terminating resistor, *c*) with another insulator about 7.5 cm from the first.

iate indication of the effect of any modification to the object under test. In fact, this method is most suitable for experimental measurements covering a wide range of frequencies.

On the other hand, the results plotted from an oscillogram are never as accurate as those obtained with direct-reading instruments, Where accurate numerical results are required, the following procedure may be adopted. A simple variable-frequency oscillator is substituted for the wobbulator and the D.C. current from the detector is measured with a microammeter. The frequencies associated with the maximum and minimum current values are determined, and these values themselves are read from the meter. The measurement is then carried out once more with the line either open, or short-circuited. Lastly, $|r|$ and φ are calculated with the aid of formulae (10) and (12), as before.

Better still is to employ an oscillator connected to a calibrated attenuator. The latter is then so adjusted as to maintain the D.C. detector current constant (with the object of avoiding any variation in the detector load, so that the detector characteristic does not affect the measurement). The ripple amplitudes are then computed from the attenuator readings.

Another refinement is to modulate the oscillator in *amplitude* at a low frequency. The detector then produces a low-frequency voltage, which can be measured with a very sensitive electronic voltmeter. This method enables a low input voltage to be employed; any valves in the object under test will not then be overloaded.

Summary. An impedance can be measured by employing it as the terminating impedance of a transmission line. The transmission line in the form of a cable about 60 m long, is fed from a variable-frequency oscillator. A detector (germanium diode) with an indicating instrument is connected to the line input. The oscillator frequencies corresponding to maximum and minimum deflections in the indicator are then determined, once with the line terminated with the unknown impedance, and once with the end of the line either short-circuited or open (total reflection). The argument and modulus of the (complex) reflection coefficient are then calculated from these frequencies and the associated readings, and used to plot the unknown impedance in the Smith chart.

One of the practical forms of this method uses a frequency-modulated oscillator — that is, a wobbulator — and an oscilloscope, whose horizontal deflection is synchronized with the frequency modulation. The oscillograms then provide data concerning the reflection coefficient, from which the impedance can be deduced; moreover, they show at once whether the impedance and the line are matched or not.

One or two applications are discussed, e.g. the matching of a tuned circuit to a cable, the measurement of various cable characteristics, and the determination of the effect of irregularities arising from insulators on an unscreened twin lead.

The wobbulator and oscillograph method is most suitable for quick experimental measurements covering a wide range of frequencies. For more accurate numerical results it is better to substitute a variable oscillator (preferably amplitude-modulated), with a calibrated attenuator, for the wobbulator, and a D.C. microammeter (or electronic millivoltmeter) for the oscilloscope.

ENTROPY IN SCIENCE AND TECHNOLOGY

III. EXAMPLES AND APPLICATIONS (continued)

by J. D. FAST.

536.75

As in the previous article (II), the author demonstrates with examples the important part played by entropy in widely differing fields of science and technology. This article deals with the theory of the specific heat of solids, the thermodynamic equilibrium of certain lattice defects in solids and their relationship with diffusion phenomena, and the equilibria of mixtures, both in a solid form (e.g. alloys) and in the form of liquid solutions, including the extreme case of solutions of polymers. The statistical aspects of rubber elasticity are also discussed, a phenomenon that can be completely described as an entropy-effect.

The first article of this series [1]), referred to below as I, gave some general observations on the concept of entropy; the second [2]), referred to as II, presented some applications of this concept in chemistry, physics and technology. In the present, third article, some further examples are discussed. A fourth article, to be published later, will be of a somewhat different nature, and will discuss the application of the concept of entropy to information theory.

The specific heat of an Einstein solid

In I (page 262) we discussed a simplified model of a solid, in which the identical atoms behave as linear harmonic oscillators, executing their vibrations around fixed centres practically independently of one another, and capable of absorbing equal quanta of a value $h\nu$. Suppose that this solid, consisting of N atoms, absorbs a number q of energy quanta $h\nu$ (starting from absolute zero). This energy can be distributed among the oscillators in a great number of ways (it is assumed that $N \gg 1$ and $q \gg 1$). Each distribution represents one microstate and the total number of micro-states m is given, according to (I, 9) by:

$$m = \frac{(q+N-1)!}{q!\,(N-1)!}.$$

For the sake of convenience it was not taken into account when deriving this formula that an atom in an Einstein solid has to be assigned *three* degrees of vibrational freedom. An actual crystal of N atoms in this model can be regarded as a system of $3N$ linear oscillators. Since real crystals as a rule contain more than 10^{19} or 10^{20} atoms, unity is negligible compared to N, and the formula becomes:

$$m = \frac{(q+3N)!}{q!\,(3N)!} \quad \ldots\ldots \text{(III, 1)}$$

The energy of the crystal, after q vibrational quanta have been absorbed, has risen to an amount

$$U = q\,h\nu \quad \ldots\ldots\ldots \text{(III, 2)}$$

above the zero-point energy, whilst the entropy, according to (I, 12) and (III, 1) and using Stirling's formula in the approximation (I, 3) is given by

$$S = k\{(q+3N)\ln(q+3N) - q\ln q - 3N\ln 3N\}. \quad \text{(III, 3)}$$

The Helmholtz free energy, $F = U - TS$, may therefore be written as

$$F(q) = qh\nu - kT\{(q+3N)\ln(q+3N) - q\ln q - 3N\ln 3N\}. \quad \text{(III, 4)}$$

Until now the number of absorbed quanta has been assumed to be known. In fact it is usually not primarily this number of quanta which is known, but the temperature to which the crystal is heated by bringing it into thermal contact with surroundings at constant temperature. What we wish to know, therefore, is the number of quanta q present at a given temperature T in the Einstein solid. The knowledge of q as a function of T will enable us to calculate the specific heat at any temperature.

If the solid could entirely submit to its "striving" towards a minimum value of of the energy, it would not absorb any quanta at all. Conversely, if it could entirely submit to its "striving" towards a maximum value of the entropy, the number of quanta absorbed would continue to increase. The compromise (the state of equilibrium) lies, according to I, at the point where the Helmholtz free energy is a

[1]) J. D. Fast, Entropy in science and technology, I, The concept of entropy, Philips tech. Rev. **15**, 258-269, 1954/55.
[2]) J. D. Fast, Entropy in science and technology, II, Examples and applications, Philips techn. Rev. **16**, 298-308, 1954/55.

minimum, i.e. where

$$\frac{dF(q)}{dq} = 0.$$

Differentiating (III, 4) and equating to zero,

$$h\nu - kT \ln \frac{q+3N}{q} = 0$$

or, after re-arranging,

$$q = \frac{3N}{e^{h\nu/kT}-1}. \quad \ldots \quad \text{(III, 5)}$$

The specific heat at constant volume is given by:

$$c_v = \frac{dQ}{dT} = \frac{dU}{dT} = \frac{d(qh\nu)}{dT}.$$

Substituting for q from (III, 5) gives the relationship sought:

$$c_v = 3\,kN\left(\frac{h\nu}{kT}\right)^2 \frac{e^{h\nu/kT}}{(e^{h\nu/kT}-1)^2}. \quad \text{(III, 6)}$$

This formula was deduced by Einstein in 1907. For the part of the internal energy that varies with temperature we can write, according to (III, 2) and (III, 5):

$$U = \frac{3N\,h\nu}{e^{h\nu/kT}-1}. \quad \ldots \quad \text{(III, 7)}$$

If kT is very much greater than $h\nu$, then

$$e^{h\nu/kT} \approx 1 + h\nu/kT$$

and therefore,

$$U \approx 3\,NkT,$$

$$c_v \approx 3\,Nk.$$

At relatively high temperatures, therefore, the specific heat of an Einstein solid reaches a constant value which is not only independent of the temperature but also of ν. If this idealized solid represented the behaviour of actual (elementary) solids, then the latter would all have the same specific heat per gram-atom at not too low temperatures. This would amount to $c_v = 3\,N_0k = 3R$, in which N_0 represents Avogradro's number and R the gas constant. According to the experimentally established law of Dulong and Petit, many solid elements do in fact have an approximately equal specific heat of about 6 cal per degree per gram-atom (R = approx. 2 cal per degree) at not too low temperatures It is also in agreement with observation that (III, 6) requires that c_v, at decreasing temperature, finally approaches zero; cf. *fig. 1.* which represents the relationship between c_v/R and $kT/h\nu$ according to (III, 6).

To a first approximation, Einstein's formula thus provides a good expression for the thermal behaviour of the solids considered here. The form of the experimental $c_v(T)$ curves does not agree in detail with the formula, however. The largest

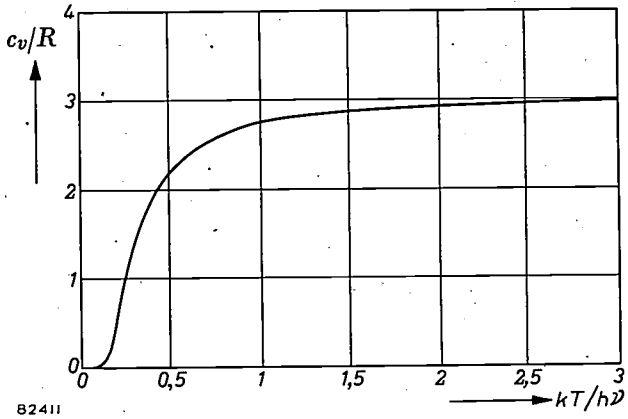


Fig. 1. c_v/R as a function of $kT/h\nu$ for an Einstein solid (formula (III, 6)).

deviations occur at very low temperatures. This is due to the fact that a real crystal cannot, in fact, be considered as an asssembly of nearly independent oscillators of equal frequency as assumed in the Einstein derivation. A crystal has a great number of vibration modes at widely divergent frequencies; they may not be regarded as the vibrations of individual atoms, for they are intrinsic to the lattice as a whole: they can be pictured as a pattern of standing waves in the crystal. A theory based on this model and in better agreement with experiment has been formulated by Debye. We shall not enter into this theory in view of the fact that, considered purely thermodynamically, it covers no fresh ground. One of its consequences, which is in good agreement with experiment, may be mentioned: at low temperatures the Stefan-Boltzmann law (cf. II) is applicable to the vibrations in a solid. According to this law, the energy is proportional to T^4, so that the specific heat is proportional to T^3. This consequence is due to the fact that both the radiation inside a hollow body and the vibrations in a solid can be considered as a broad spectrum of modes of vibration. It is true that the hollow space has an infinite number of modes, whereas the number of modes in the solid is restricted by the number of atoms N (it amounts to $3N$). At low temperatures, however, this difference is irrelevant since then, according to quantum mechanics, the high-frequency modes play no part.

We have already seen that to consider a solid as an assembly of independent oscillators of equal frequency is rather an inexact model. A diatomic or polyatomic gas, on the other hand, conforms very

satisfactorily to this model. In a diatomic gas each molecule can vibrate in such a manner that the atoms oscillate along their line of centres. (Translational and rotational degrees of freedom also contribute to the specific heat of a gas, but will not be discussed here.) Exchanges of energy are brought about by collisions between the gas molecules. Apart from this the vibrations are mutually independent. Formula (III, 6), with the omission of the factor 3, is therefore applicable with considerably greater accuracy to the vibrations of a diatomic gas than to those of a solid. The specific-heat values calculated on this basis provide the necessary corrections to give validity to the calculation of chemical equilibria (see the beginning of II) even at very high temperatures.

Vacancies and diffusion in solids

Atomic diffusion plays an important part in several processes occurring in solids. Diffusion in metals and alloys in particular has been the subject of extensive research during the last few decades. This has revealed that the presence of defects in the periodic crystal lattice, particularly vacancies or interstitial atoms, is essential for the occurrence of diffusion.

As we have seen in the foregoing discussion of an Einstein solid, the vibrations of the atoms around their positions of equilibrium become more and more violent as the temperature rises. Even before the melting point is reached, a fraction of the atoms will possess sufficient energy to leave their lattice positions completely. These atoms form new lattice planes on the outside of the crystal, since the spaces between the other atoms, the interstices, are too small to accommodate them. Starting, then, with a perfect crystal (no defects), any vacancies which occur in the crystal as the temperature rises must originate in the boundary layer. One can imagine that a few atoms in the boundary layer leave their original positions and occupy new positions on the surface (*fig. 2*). Atoms from deeper layers can subsequently jump into the newly created vacancies, and so on. We may equally well say that the vacancies arise at the surface and subsequently diffuse to the interior. The fact that at high temperatures lattice defects are bound to arise even in the state of equilibrium, is due to the fact that their occurrence represents an increase in the entropy. Disregarding the extremely small macroscopic volume changes occurring with the formation of the vacancies, it can be said that at a certain temperature T vacancies will continue to be formed until ultimately the Helmholtz free energy $F = U - TS$ has reached its minimum value. Although the internal energy U will rise by an amount ΔU due to the introduction of vacancies, it is nevertheless possible that under certain conditions the value of the free energy will drop, viz. if $T\Delta S > \Delta U$.

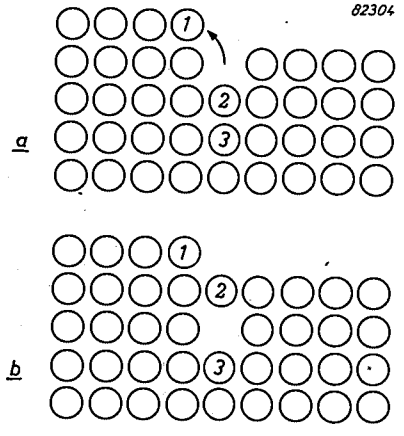


Fig. 2. The vacancies in a solid are assumed to form at the outer surface (*a*) and subsequently diffuse to the interior (*b*), this being equivalent to the outward diffusion of atoms (*1*, *2*, *3*).

If the energy ε required for forming a vacancy is known, the percentage of vacancies can be evaluated for any temperature. This is done as follows.

If a perfect crystal of N atoms changes into an imperfect state with n vacancies, then the corresponding increase in entropy, $\Delta S = k \ln m$, is given by the number of different ways m in which N identical atoms can be accommodated in a lattice with $(N + n)$ available positions. This number is

$$m = \frac{(N+n)!}{N!\,n!}. \quad \ldots . \quad \text{(III, 8)}$$

Using Stirling's formula in the approximation (I, 3), we find:

$$\Delta S = k\{(N + n) \ln (N + n) - N \ln N - n \ln n\}.$$

As long as the concentration of the vacancies is so small that their interaction is negligible, the increase in the energy is determined by $\Delta U = n\varepsilon$ and thus the change of the free energy by

$$\Delta F = \Delta U - T\Delta S = \\ = n\varepsilon - kT\{(N + n) \ln (N + n) - N \ln N - n \ln n\}. \quad \text{(III, 9)}$$

In the equilibrium state, the free energy is a minimum; thus $\partial F/\partial n = 0$, or, from (III, 9):

$$\varepsilon - kT\{\ln (N + n) - \ln n\} = 0,$$

i.e.

$$\frac{n}{N+n} = e^{-\varepsilon/kT}.$$

Since n is negligible with respect to N, this becomes:

$$\frac{n}{N} = e^{-\varepsilon/kT} = e^{-E/RT}, \quad . \quad . \quad . \quad (III, 10)$$

where E is the energy required to make 6×10^{23} (Avogadro's number) vacancies. According to the calculations of Huntington and Seitz [3]) ε has the value of approximately 1 eV for copper, i. e. $E \approx$ 23 000 cal per "gram-atom of vacancies". For 1000 °K, therefore, we arive at:

$$\frac{n}{N} = e^{-23\,000/2000} = 10^{-5}.$$

In this case one in each 100 000 lattice positions would be unoccupied. This corresponds to an average distance between adjacent vacancies of the order of thirty interatomic distances.

The additional contribution to the internal energy owing to the formation of vacancies is accompanied by an additional contribution to the specific heat, given by:

$$\frac{dAU}{dT} = \frac{dn\varepsilon}{dT} = \frac{d(Nee^{-\varepsilon/kT})}{dT} = R\left(\frac{\varepsilon}{kT}\right)^2 e^{-\varepsilon/kT} \quad \text{per gram-atom.}$$

The above is based on the assumption that the increase in the entropy during the transition from the ideal to the disturbed arrangement is exclusively based on the number of possible distributions m, given by formula (III, 8). In reality the vibration frequencies of the atoms near the vacancies also change; this contributes to the entropy.

This may be explained as follows. An atom adjacent to a vacancy is bound more weakly than an atom completely surrounded by similar atoms. Consequently it will have a lower vibration frequency, at least in the direction of the vacancy. Since, from (III, 3) and (III, 5).

$$S \approx 3\,Nk \ln \frac{kT}{h\nu} \quad \text{for} \quad kT \gg h\nu, \quad . \quad . \quad (III, 11)$$

is applicable to an Einstein solid, the introduction of vacancies causes not only the entropy increase as evaluated in (III, 8), but also an increase in the vibrational entropy. The concentration of the vacancies can, therefore, be considerably greater than the value calculated above.

At high temperatures the vacancies move at random through the lattice. A displacement of a vacancy over one interatomic distance is brought about if an adjacent atom jumps into the vacant site. For such a jump an amount of energy, at least equivalent to the activation energy q is required. This energy will, as a rule, be large compared to the mean thermal energy of the atoms. The fraction of the total number of atoms which possesses at least the activation energy q is given by

$$f = e^{-q/kT} = e^{-Q/RT},$$

where $Q = N_0 q$ is the "activation energy per gram-atom". The diffusion constant D of the *vacancies* will be approximately proportional to this factor:

$$D = D_0\, e^{-Q/RT}, \quad . \quad . \quad . \quad . \quad (III, 12)$$

where D_0 is a constant which is independent or only slightly dependent on the temperature.

Primarily, however, we are often not so much interested in the diffusion of the vacancies as in the diffusion of the atoms of the pure metal, i.e. the self-diffusion, or the diffusion of foreign atoms occupying the lattice sites (substitutional atoms). The self-diffusion can be studied with the aid of certain radioactive isotopes of the atoms of the pure metal. The diffusion constant for this type of diffusion should be smaller than that for the vacancy diffusion, because the atoms are capable of jumping only at that moment when there happens to be a neighbouring vacancy. The number of times per second that this condition occurs is proportional to the diffusion constant of the vacancies and to their relative number, and hence to (III, 12) and (III, 10).

According to this reasoning, the diffusion coefficient will be given by an expression of the form:

$$D = D_0'\, e^{-Q/RT}\, e^{-E/RT}$$

or

$$D = D_0'\, e^{-W/RT}, \quad . \quad . \quad . \quad . \quad (III, 13)$$

where

$$W = Q + E \quad . \quad . \quad . \quad . \quad . \quad (III, 14)$$

and D_0' is a constant which is independent or only slightly dependent upon the temperature.

The foregoing also applies to substitutional atoms, provided that the properties and size of these atoms differ so little from those of the solvent that the vacancies have no preference for either type of atom, and that the necessary activation energy is the same for both.

Alloys

Entropy of mixing and energy of mixing

An enormous number of different alloys with a great diversity of properties can be made, owing to the almost limitless combinations which are possible: many metallic elements can be combined in several ways into binary, ternary, etc. alloys, the relative quantities in each combination can be varied, and each alloy can be subjected to a great variety of heat and mechanical treatments.

The central problem in the metallurgy of alloys is that of the solubility of the different metals, more particularly the solubility in the solid state. If we

[3]) H. B. Huntington and F. Seitz, Phys. Rev. 61, 315-325, 1942.

confine ourselves to the alloys of two metals (binary alloys), we find that the majority of combinations, no matter in what proportion they are mixed, form a homogeneous mixture in the liquid state, whereas in the solid state they have only a limited miscibility. Miscibility in all proportions in only possible in the solid state if both metals have the same crystal structure and if their atoms show only slight differences in size and electron configuration.

Homogeneous mixing is in nearly all cases accompanied by an increase of the entropy, i.e. it is nearly always promoted by the entropy effect. The separation of the mixture into two phases can only occur if also the energy increases during the homogeneous mixing. If the energy does not change or even decreases in the course of mixing, then only one homogeneous phase can exist in the state of equilibrium. The entropy increase occurring when two metals A and B are homogeneously mixed, can again be directly established by a statistical reasoning, provided that all configurations are equally probable. If we consider a total number of N metal atoms, of which Nx are of type B and hence $N(1-x)$ are of type A, then the number of possible ways in which these can be distributed among the available lattice points is given by:

$$m = \frac{N!}{(Nx)!\{N(1-x)\}!} \quad \text{(III, 15)}$$

The increase in entropy that accompanies the mixing i.e. the entropy of mixing $\Delta S = k \ln m$, thus becomes:

$$\Delta S = Nk\{-x \ln x - (1-x)\ln(1-x)\}. \quad \text{(III, 16)}$$

Strictly speaking, this expression only gives the entropy of mixing at 0 °K. If it is assumed that the vibrational entropy (cf. III, 11) changes little or not all in the course of mixing, (III, 16) is also applicable at other temperatures.

To find a mathematical expression for the energy of mixing (also called the heat of solution) we start from the over-simplified assumption that the internal energy at zero temperature can be written as the sum of the binding energies of nearest neighbours only. As a consequence of this only three interaction energies ε_{AA}, ε_{BB} and ε_{AB} between neighbouring pairs A-A, B-B and A-B will occur in the terms of this sum [4]). In order to do the summation we have to evaluate the numbers of the three types of neighbour configurations.

The number of nearest neighbours to any one atom in alloy is denoted by z ($z = 8$ for the body-centred cubic structure and $z = 12$ for the close-packed cubic and hexagonal structures). If the atoms of the type A and B are distributed at random among the lattice points, then the chance of an atom A occupying any given lattice point is $(1-x)$ whilst that of an atom B is x. The probability of finding the combination A-A in two adjacent positions is further given by $(1-x)^2$, that of finding B-B by x^2 and that of finding either A-B or B-A by $2x(1-x)$. In total there are $Nz/2$ bonds between the N atoms of the alloy; in this particular case they are distributed as follows:

$$\frac{Nz}{2}(1-x)^2 \quad A\text{-}A \text{ bonds},$$

$$\frac{Nz}{2}x^2 \quad B\text{-}B \text{ bonds},$$

$$Nz\,x\,(1-x) \quad A\text{-}B \text{ bonds}.$$

The energy of the alloy at 0 °K is therefore given by:

$$U_{AB} = \frac{Nz\varepsilon_{AA}}{2}(1-x)^2 + \frac{Nz\varepsilon_{BB}}{2}x^2 + Nz\varepsilon_{AB}x\,(1-x). \quad \text{(III, 17)}$$

For pure A ($x = 0$) the formula gives:

$$U_{AA} = \frac{Nz\,\varepsilon_{BB}}{2};$$

for pure B ($x = 1$):

$$U_{BB} = \frac{Nz\varepsilon_{BB}}{2}.$$

For a *heterogeneous mixture* of the pure metals, the internal energy is given by:

$$\frac{Nz\varepsilon_{AA}}{2}(1-x) + \frac{Nz\,\varepsilon_{BB}}{2}x. \quad \text{(III, 18)}$$

The heat of solution at 0 °K is given by the difference between (III, 17) and (III, 18):

$$\Delta U = x(1-x)Nz\left\{\varepsilon_{AB} - \frac{\varepsilon_{AA}+\varepsilon_{BB}}{2}\right\}. \quad \text{(III, 19)}$$

As a rule the vibrational energy will not have much influence during the mixing process, so that (III, 19) is also valid at higher temperatures.

Since $x(1-x)Nz$ is always positive, the sign of ΔU is determined by that of $\varepsilon_{AB} - \frac{1}{2}(\varepsilon_{AA} + \varepsilon_{BB})$.

Solution, ordering, segregation and precipitation

If in the foregoing ε_{AB} were to equal the mean of ε_{AA} and ε_{BB} (which will probably never be exactly

[4]) We would ultimately find the same formula (III, 19) — still to be derived — if we also took into account, the interaction between *more remote* pairs. The energies ε_{AA}, ε_{BB} and ε_{AB} would then represent the sums of the interaction energies between pairs A-A, B-B and A-B. Fundamentally the formula would only change if we considered the interaction between e.g. sets of three or four atoms.

true), then $\Delta U = 0$. As regards energy, there is no preference whatsoever for a specific bond, and the entropy effect causes a homogeneous mixing in the state of equilibrium.

If ε_{AB} is more negative than the mean of ε_{AA} and ε_{BB}, then ΔU is negative. In that case there is a stronger attraction between the atoms A and B than between atoms of the same type. During the mixing, therefore, heat is liberated. The entropy effect is here supported by the energy effect and the homogeneous solution, just as in the previous case, is the thermodynamically stable condition. Contrary to the previous case, however, there here occurs (at a suitable atomic ratio) a tendency towards *ordering* in the sense that each A atom is preferentially surrounded by B atoms and each B atom by A atoms At temperatures such that the absolute value of $\varepsilon_{AB} - \frac{1}{2}(\varepsilon_{AA} + \varepsilon_{BB})$ is large compared to kT, the atoms can submit to this ordering provided that a perceptible diffusion can still occur at this temperature.

A typical example of an alloy in which ordering occurs, is β-brass. This phase occurs in the system copper-zinc. The stability range extends from about 40 to 50% zinc. The system has a body-centred cubic structure. At high temperatures the entropy-effect prevails, so that the copper and zinc atoms are distributed at random among the lattice points. At lower temperatures (below approx. 450 °C) the energy effect predominates, so that ordering occurs. Roughly speaking, each Zn-atom surrounds itself by eight Cu-atoms and each Cu-atom by eight Zn-atoms (*fig. 3*). This ordered structure is designated β'-brass.

If ε_{AB} is less negative than the mean of ε_{AA} and ε_{BB}, then ΔU is positive. The attraction between identical atoms is then stronger than the attraction between different atoms. During isothermal mixing, the internal energy increases, i.e. heat is absorbed. The entropy effect promotes the mixing, but the energy effect promotes separation. At low temperatures the latter effect prevails and the system consists of two phases: virtually pure A and virtually pure B. At rising temperature, however, the entropy effect causes increasing mutual solubility. For a system satisfying the conditions of the previous section the mutual solubility as a function of the temperature can be evaluated as follows.

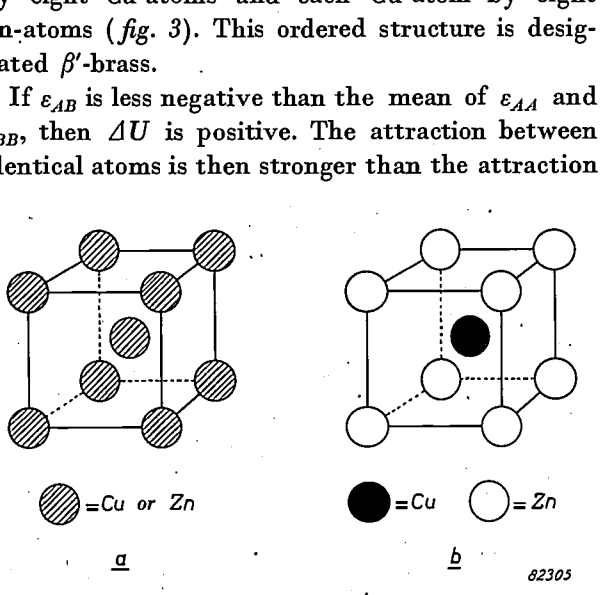


Fig. 3. The unit cell of brass with the composition 50% Cu + 50% Zn.
a) At high temperature (β-brass), the atoms are distributed at random among the lattice points.
b) At lower temperatures (β'-brass), the atoms are ordered, as this is more favourable from the point of view of the internal energy.

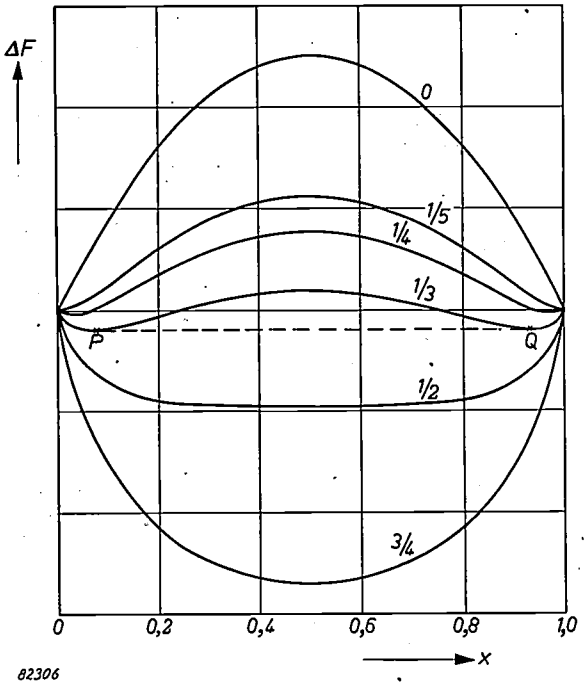


Fig. 4. Free energy of mixing ΔF, according to (III, 20) and (III, 21), as a function of the atomic fraction x for various values of kT/C. At low temperatures, segregation occurs as this is more favourable from the point of view of internal energy. For a given temperature, the compositions of the two co-existing phases are defined by the points of contact P and Q of the (dashed) double tangent to the ΔF-curve.

The increase in the free energy due to homogeneous mixing is given, according to (III, 16) and (III, 19) by

$$\Delta F = NCx(1-x) + NkT\{x \ln x + (1-x) \ln (1-x)\}, \quad \text{(III, 20)}$$

where

$$C = z\left\{\varepsilon_{AB} - \frac{\varepsilon_{AA} + \varepsilon_{BB}}{2}\right\} \quad \text{(III, 21)}$$

Fig. 4 shows ΔF as a function of x for various values of the ratio kT/C. The curve for $T = 0$ ($kT/C = 0$) is obviously identical to the ΔU-curve[5]).

[5]) It will be clear that the expressions (III, 20), (III, 21) are only justified for small values of C/kT. In fact, if C differs from zero, some ordering will immediately occur, causing the entropy of mixing to become smaller than that relating to a random distribution; the energy of mixing will also change.

Throughout a wide temperature range the ΔF-curves show two minima, which approach each other as the temperature rises. The ΔF-curves are symmetrical with respect to the vertical axis $x = 0.5$, which is evidently due to the symmetrical nature of x and $(1-x)$ in (III, 20). The double tangent PQ to a ΔF-curve is thus parallel to the abscissa. The free energy has a minimum in the state of equilibrium, i.e. the homogeneous mixtures (solutions) represented by the branches of the curve to the left of P and to the right of Q, are stable. The portion of the ΔF-curves situated between P and Q, on the other hand, represent non-stable solutions. By a separation into two phases with compositions P and Q, the free energy of a homogeneous alloy can be lowered to a point on the straight line PQ. The points P and Q, therefore, indicate the solubility values sought. The position of these points can be found by putting the differential coefficient of (III, 20) equal to zero:

$$\frac{d(\Delta F)}{dx} = NC\,(1-2x) + NkT\{\ln x - \ln(1-x)\} = 0, \qquad \text{(III, 22)}$$

from which it follows:

$$\frac{x}{1-x} = e^{-C(1-2x)/kT}. \qquad \text{(III, 23)}$$

For the construction of the curve representing the solubility as a function of the temperature it is more convenient to put the relationship between x and T in the form

$$T = \frac{C}{k}\,\frac{(1-2x)}{\ln\{(1-x)/x\}}. \qquad \text{(III, 24)}$$

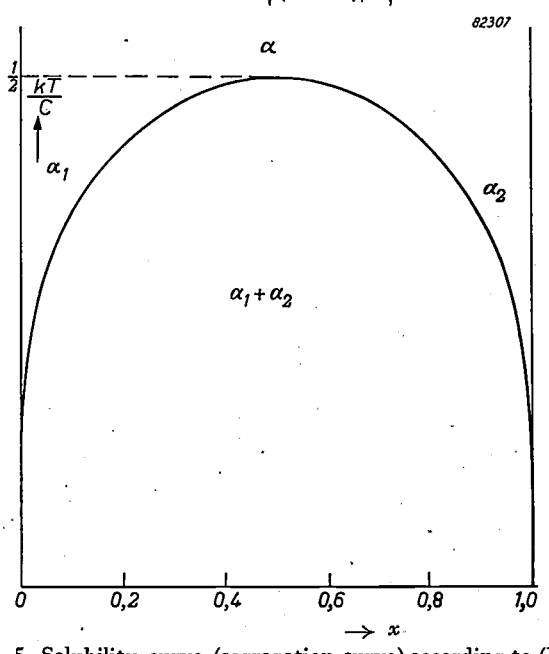


Fig. 5. Solubility curve (segregation curve) according to (III, 23) or (III, 24), giving the relationship between kT/C and the composition (P and Q in fig. 4) of the co-existing phases α_1, α_2.

Fig. 5 gives the solubility curve according to (III, 23) or (III, 24). The diagram shows that above a certain temperature complete miscibility occurs.

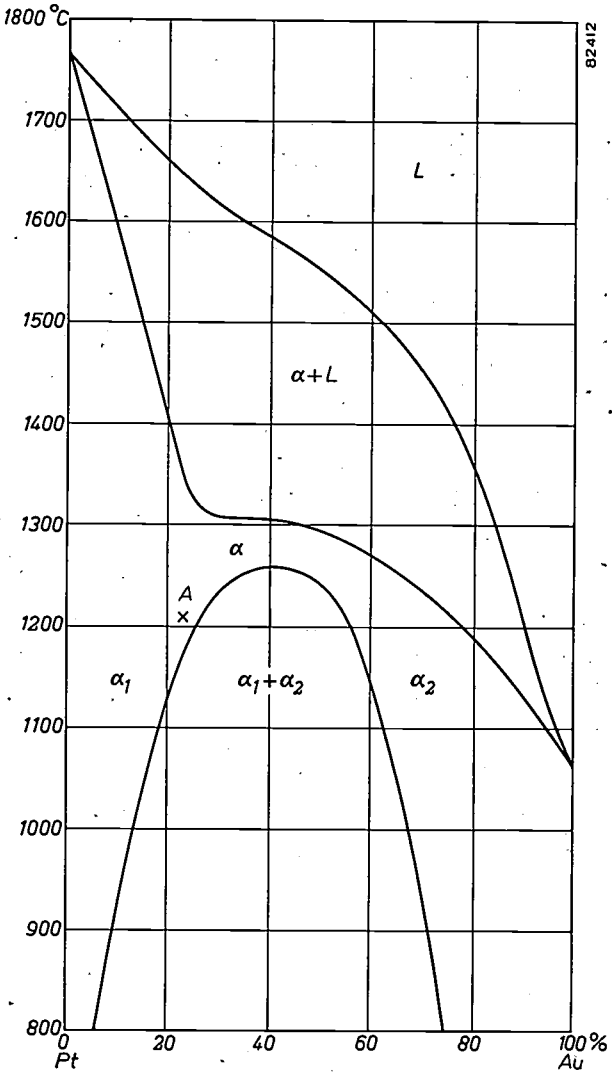


Fig. 6. Phase diagram of the system platinum-gold, which, even in the solid state, shows a complete curve of the type of fig. 5. α, α_1 and α_2 indicate solid phases; the liquid phase is denoted by L.

This "critical" temperature can be evaluated as follows. A ΔF-curve with two minima will also have two points of inflection, at which we have

$$\frac{d^2(\Delta F)}{dx^2} = -2NC + NkT\left(\frac{1}{x} + \frac{1}{1-x}\right) = 0,$$

whence

$$2C = kT\left(\frac{1}{x} + \frac{1}{1-x}\right). \qquad \text{(III, 25)}$$

At the temperature rises, the points of inflection approach each other to coincide ultimately at $x = 1/2$, when the critical temperature T_k is reached. From (III, 25) it follows:

$$T_k = \frac{C}{2k}. \qquad \text{(III, 26)}$$

Actual solubility curves never show the symmetrical form of fig. 5. This is mainly due to the fact that the assumption made above, viz. that the energy may be taken as the sum of interactions between pairs only, involves a certain approximation [4]). A complete solubility curve in the sense described above occurs, e.g., in the system platinum-gold (*fig. 6*). In most cases of binary systems with a positive ΔU-value, however, melting phenomena occur long before the critical mixing temperature is reached. An example of this is the system silver-copper (*fig. 7*).

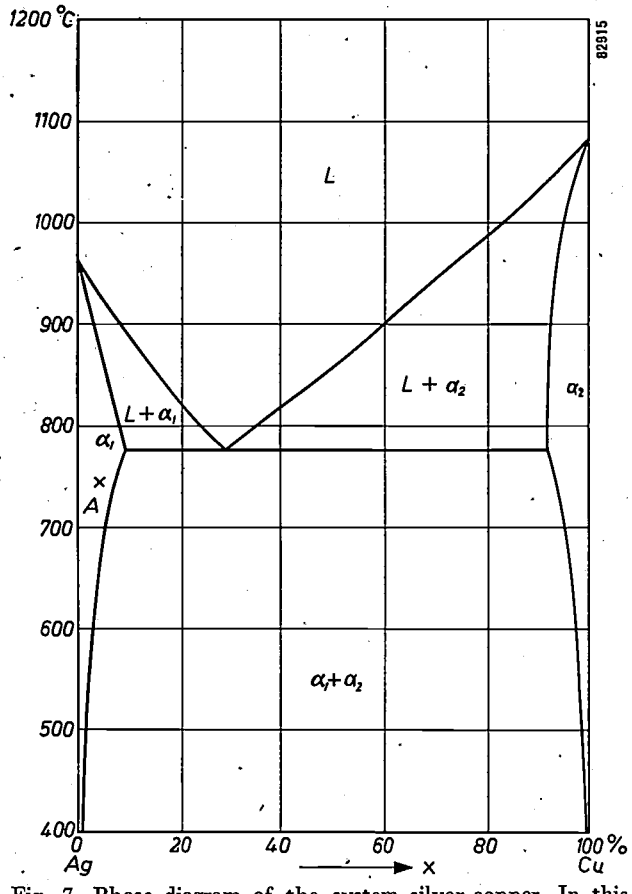


Fig. 7. Phase diagram of the system silver-copper. In this case melting phenomena occur long before the critical mixing temperature is reached.

The decreasing solubility at decreasing temperature occurring in systems of the type discussed here, is often utilized industrially, where it is known as precipitation hardening. This is done by rapidly cooling an alloy from the homogeneous region (e.g. from point *A* in figs 6 and 7) down to a temperature at which there are two phases in the state of equilibrium and at which virtually no diffusion occurs. Owing to the rapid cooling the thermodynamically required separation cannot occur and a super-saturated homogeneous solution is obtained. By subsequently heating the alloy at a suitable, not excessively high temperature, incipient precipitation occurs, which is accompanied by a substantial increase in the hardness.

Rubber elasticity

Rubber and rubber-like materials belong to the class of substances known as high polymers. Their molecules take the form of long chains, the links of which consists of groups of relatively few atoms (monomers). There may be several hundreds of identical monomers in the molecule. These monomeric groups are bound together by real chemical bonds. The atoms in these molecules nevertheless have a certain mobility with respect to each other in the sense that one bond can rotate around the other at a constant angle. In *fig. 8* this is shown schematically for a simple case, the black spots representing atoms of the chain. Owing, to this

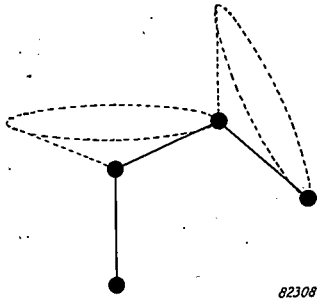


Fig. 8. Macro-molecules often possess a great flexibility owing to the fact that the chemical bonds, though of constant valence angle, can occupy different positions with respect to each other.

rotational freedom, the chain molecules can assume an enormous number of different forms. Against the one chance of the molecule occurring in the form of a stretched chain there exist countless other possibilities of each molecule occurring in a coiled or tangled form (see *fig. 9*).

The entropy of a high-polymer substance will be greater when the chain molecules have an irregularly coiled form than when they are stretched out and

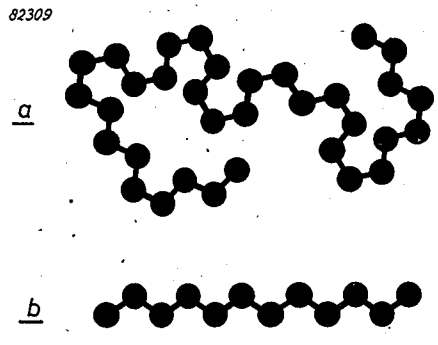
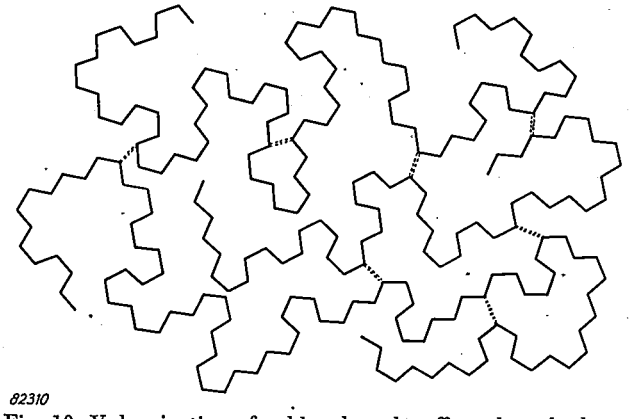


Fig. 9. Part of a molecular chain: *a*) coiled at random, *b*) stretched. For simplicity case (*a*) is depicted in a flat plane (the same applies to fig.10).

lie in parallel arrangement: this follows from the relationship $S = k \ln m$, discussed in I (m = number of micro-states). In some polymers, e.g. cellulose, the latter condition is so much more favourable energetically than the former, that the tendency towards a maximum value of the entropy is overcome by the tendency towards a minimum value of the energy, so that the ordered (stretched) condition prevails. In rubber the difference in energy between the two conditions is only very slight, so that the entropy effect prevails and the molecules have a coiled form. By exerting a tensile stress in a rubber rod, the chains can be stretched, during which considerable elongation of the whole rod occurs. After the stress is removed, the thermal motion of the links of the chain restore the chains into the coiled and thus shortened form. The fact that also the rod as a whole returns to its initial form is explained by the presence of cross-links between the polymer chains. These links are established during the vulcanizing process of the rubber, which gives rise to a wide-meshed network (*fig. 10*). In normal cases about 1% of the monomeric groups are linked by cross-links in this process.



82310

Fig. 10. Vulcanization of rubber has the effect that the long rubber molecules (——) are bound together by short cross-links (═══). These cross-links may consist e.g. of two sulphur atoms: — S — S —.

Some of the remarkable thermal properties of rubber can be explained by the foregoing, e.g. the negative coefficient of expansion of strongly stretched rubber. The contraction at rising temperature and constant load is caused by the increasing thermal agitation of the links of the chain molecules, i.e. the increased tendency of the molecules to assume their shortened, coiled form. The parallel alignment of the molecules of natural rubber under considerable stress is a kind of crystallization phenomena. The crystalline nature of the rubber in this condition can be established by X-ray diffraction. After removal of the load the crystalline parts "melt" again.

For a thermodynamical treatment of the subject we start from the first law

$$\mathrm{d}U = \mathrm{d}Q + \mathrm{d}W, \quad \ldots \quad \text{(III, 27)}$$

where $\mathrm{d}Q$ represents the heat supplied to the system and $\mathrm{d}W$ the work exerted upon it. Suppose that the system is a rubber rod of unit diameter and length l, and it is stretched to a length $l + \mathrm{d}l$. If we disregard the very small amount of energy spent to bring about the small volume change, $\mathrm{d}W$ is given by:

$$\mathrm{d}W = \sigma\,\mathrm{d}l,$$

where σ is the applied stress. If the stretching takes place reversibly, $\mathrm{d}Q = T\mathrm{d}S$ (if S is the entropy of the rubber rod); σ is the equilibrium value of the stress in that case. Equation (III, 27) can now be written as:

$$\mathrm{d}U = T\mathrm{d}S + \sigma\,\mathrm{d}l \quad \ldots \quad \text{(III, 28)}$$

At constant temperature, therefore, we have

$$\mathrm{d}(U - TS) = \mathrm{d}F = \sigma\,\mathrm{d}l,$$

and hence

$$\sigma = \frac{\partial F}{\partial l} = \frac{\partial U}{\partial l} - T\frac{\partial S}{\partial l}. \quad \ldots \quad \text{(III, 29)}$$

For many types of rubber it has been found experimentally that in the case of a large, constant elongation Δl, σ is approximately proportional to the absolute temperature T. This confirms the statement made above that $\partial U/\partial l$ (the increase of the internal energy with the elongation and therefore with the uncoiling of the polymer chain) is very small for this class of substances. This means the the elastic properties of rubberlike substances are a result merely of the decreasing entropy, during stretching and not on the accompanying small change in the energy. For the majority of solids, e.g. metals, the reverse is true. These materials become cool slightly when adiabatically and elastically stretched, whereas rubber heats up slightly under the same conditions. This rise in temperature can be partly explained by the fact that the total entropy does not change with a reversible adiabatic process, because $\mathrm{d}S = \mathrm{d}Q_{\mathrm{rev}}/T$. The decrease in the entropy, corresponding to the increasing order during strain, is therefore compensated by an increase in the vibrational entropy of the molecules; this is accompanied by an increase in temperature.

The behaviour of a "perfect" rubber, for which $\partial U/\partial l = 0$ and whose elasticity may be regarded as a pure entropy effect, is completely analogous to that of a perfect gas. For a reversible volume change of a perfect gas,

$$\mathrm{d}U = T\mathrm{d}S - p\,\mathrm{d}V,$$

so that at constant temperature

$$d(U-TS) = dF = -p\,dV,$$

$$p = -\frac{\partial F}{\partial V} = -\frac{\partial U}{\partial V} + T\frac{\partial S}{\partial V}.$$

At constant volume, the pressure of a gas is found to be approximately proportional to the absolute temperature, i.e. the internal energy U of gases is nearly independent of the volume. If the gas is "perfect" $\partial U/\partial V = 0$. The pressure of such a gas can thus also be described as an entropy effect.

Solutions of polymers

We have seen that the entropy of mixing of a homogeneous solid alloy, so far as it depends upon the configuration of the atoms concerned, is given by formula (III, 16). Strictly speaking this formula is only applicable when the *energy* of mixing $\Delta U = 0$. If ΔU differs from zero, the various configurations no longer possess the same probability. In this case the formula provides a sufficient approximation only for temperatures at which kT is considerably greater than $\varepsilon_{AB} - \frac{1}{2}(\varepsilon_{AA} + \varepsilon_{BB})$.

No reference was made above to any difference in size between the atoms or molecules of the mixture. This was not necessary for the *solid* alloys under consideration, in view of the fact that two metals with greatly different atomic radii will not show any appreciable miscibility in the solid state.

Liquid homogeneous mixtures, however, often consist of molecules (or atoms) of widely differing sizes. Solutions of polymers are extreme cases of such mixtures. In the following we shall derive an approximate expression for the entropy of mixing of solutions of this kind.

Let us first consider a liquid mixture of molecules A and B, which have the same shape and size and furthermore show no preference for similar or dissimilar neighbouring molecules ($\Delta U = 0$). In such a mixture, a molecule B may be substituted for a molecule A without any effect on the immediate surroundings. As is known from research on liquids, the immediate surroundings differ little from those in the crystalline state. The arrangement is merely slightly less ordered, so that the order does not extend over such a long range as in the crystalline form. Since we are mainly interested in the bonds between neighbours, it is permissible to use a quasi-crystalline model (*fig. 11*) to represent liquids of the type under consideration. Formula (III, 16) is then applicable without modification. This formula assumes the interchangeability of the molecules A and B. If however the molecules differ considerably in size or shape, then the entropy of mixing, even when the energy of mixing ΔU is zero, is no longer given by (III, 16). For example, a direct interchangeability of molecules of A and B is out of the question if the molecules A are spherical and occupy one site in the diagram of fig. 11 and the molecules B are dumb-bell shaped and occupy two sites (*fig. 12*).

The extreme case of polymers (macro-molecules) dissolved in a micro-molecular solvent can be idealized as follows [6]). It is assumed that a polymer behaves as if it were divided into a large number of

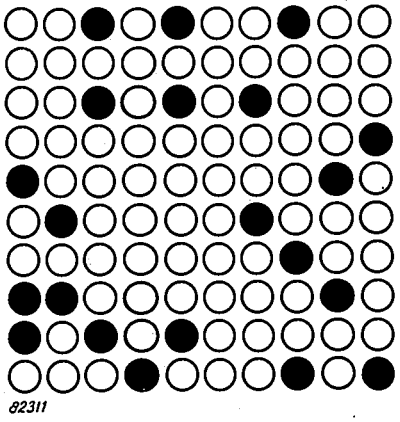


Fig. 11. Schematic representation of a solution of molecules B (black) in a liquid of molecules A (white). The molecules A and B are of about the same size and are statistically distributed in the quasi-lattice.

mobile segments, each having the same size as one molecule of the solvent. It is further assumed that each segment occupies one site in the quasi-lattice and that the adjacent segments of a chain occupy adjacent sites (*fig. 13*). If the solution contains N_1 molecules of solvent and N_2 polymer molecules, each consisting of p segments, then there are $(N_1 + pN_2)$ sites in the sense of figs 11-13. The fractions

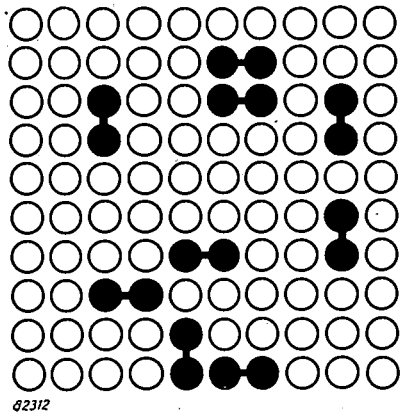


Fig. 12. Schematic representation of a statistical distribution of dumb-bell shaped molecules B (black) in a liquid of spherical molecules A (white). The dumb-bell shaped molecules occupy two adjacent sites of the quasi-lattice.

[6]) Cf. e.g. P. J. Flory, Principles of polymer chemistry, Cornell University Press, New York 1953.

of the total volume occupied by both components are given by

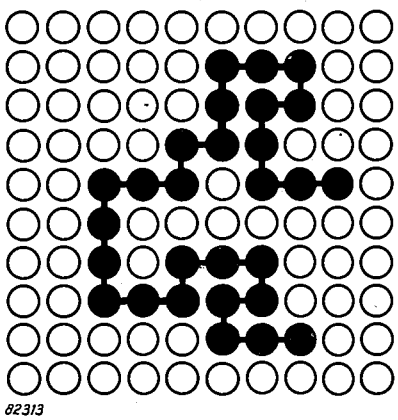
$$\varphi_1 = \frac{N_1}{N_1 + pN_2}, \qquad \varphi_2 = \frac{pN_2}{N_1 + pN_2}. \quad \text{(III, 30)}$$

An expression for the entropy of mixing, in which φ_1 and φ_2 are to play an important part, is now derived as follows. First consider a state in which all $(N_1 + pN_2)$ sites are unoccupied. The first segment of the first polymer molecule can thus be accommodated in $(N_1 + pN_2)$ different ways. Once a certain site hasb een decided upon, the second segment of the same polymer molecule can be accommodated in the quasi-lattice in z different ways, z being the number of adjacent sites of any given site in the lattice (the so-called *co-ordination number*).

Each of the subsequent segments (the 3rd, 4th, . . ., p^{th} of the same polymer-molecule has $(z-1)$ sites are available, because one of the z adjacent sites is already occupied by the first segment. The first polymer molecule can, therefore, be accommodated in

$$(N_1 + pN_2)\, z\, (z-1)^{p-2} \quad . \quad . \quad . \quad \text{(III, 31)}$$

different ways [7]).



82313

Fig. 13. Schematic representation of an arbitrarily coiled polymer molecule (black) in a solvent of the same type as that in figs. 11 and 12.

Let us now assume that the quasi-lattice already accommodates $(k-1)$ polymer molecules. In how many different ways can the k^{th} polymer molecule now be accommodated? The first segment of this molecule can be placed in any one of

$$N_1 + pN_2 - (k-1)\, p = N_1 + p\,(N_2 - k + 1)$$

different sites. The second segment of the k^{th} polymer molecule, unlike that of the first, cannot be accommodated in z different ways, since there is a possibility that one of the adjacent sites is occupied by a segment of one of the $(k-1)$ polymer molecules already accommodated. This possibility can be accounted for with a reasonable accuracy by multiplying z by the ratio of the number of sites still unoccupied to the total number of sites

$$\frac{N_1 + p(N_2 - k + 1)}{N_1 + pN_2}.$$

A similar reduction factor can be applied to each factor $(z-1)$ for the number of ways in which the 3rd, 4th, ... p^{th} segment can be accommodated. The k^{th} polymer molecule can, therefore, be placed in the lattice in

$$\frac{\{N_1 + p(N_2 - k + 1)\}^p z\,(z-1)^{p-2}}{(N_1 + pN_2)^{p-1}} \quad \text{(III, 32)}$$

different ways. For $k = 1$ this expression reduces to the form (III, 31). The total number of ways in which N_2 identical polymer molecules can be accommodated in the lattice is given by the product of the N_2 terms (III, 32) which are obtained by subsequently substituting for k the values 1, 2...., N_2 Since in our case the polymer molecules are actually identical, so that no new situation is created when two of them are interchanged, the resulting number has to be divided by $N_2!$. The total number of possible different arrangements of N_2 identical polymer molecules among the $(N_1 + pN_2)$ sites of our model is therefore

$$m = \frac{\{z(z-1)^{p-2}\}^{N_2}}{N_2!} \times \frac{(N_1+pN_2)^p[(N_1+p(N_2-1)]^p \ldots (N_1+p)^p}{(N_1+pN_2)^{(p-1)N_2}}. \quad \text{(III, 33)}$$

In each of these arrangements a number of sites (N_1) remain unoccupied. Since in each case these sites can be occupied by molecules of the solvent in one way only, the addition of these molecules does not add to the number m. In other words, the expression (III, 33) also gives the number of possible configurations m_{12} of the mixture of N_1 solvent molecules and N_2 polymer molecules. After rearranging, we obtain the expression:

$$m_{12} = \frac{\{z(z-1)^{p-2}\}^{N_2} p^{N_2}}{N_2!\left(\frac{N_1}{p} + N_2\right)^{(p-1)N_2}} \frac{\left\{\left(\frac{N_1}{p} + N_2\right)!\right\}^p}{\left\{\left(\frac{N_1}{p}\right)!\right\}^p}. \quad \text{(III, 34)}$$

The entropy of mixing is given by

$$\Delta S = k \ln m_{12} - k \ln m_1 - k \ln m_2, \quad \text{(III, 35)}$$

where m_1 is the number of possible arrangements of N_1 solvent molecules in the solvent itself (N_1 sites), and m_2 for the number of possible arrange-

[7]) Here certain factors have been overlooked. We shall return to these points at the end of this article.

ments of the N_2 polymer molecules in the polymer itself (pN_2 sites).

It is clear that $m_1 = 1$ and hence $k \ln m_1 = 0$. For m_2 this does not apply because of the many possible coiled configurations of the polymer chains. If it is assumed that the polymer molecules in the undiluted state behave in an analogous way as in the solution, then m_2 is found by putting $N_1 = 0$ in (III, 34):

$$m_2 = \frac{\{z(z-1)^{p-2}\}^{N_2} p^{N_2} (N_2!)^p}{N_2!\, N_2^{(p-1)N_2}}. \quad \text{(III, 36)}$$

From the last three formulae it follows that

$$\Delta S = k \ln \left[\frac{\left\{\left(\frac{N_1}{p} + N_2\right)!\right\}^p}{\left\{\left(\frac{N_1}{p}\right)!\right\}^p (N_2!)^p} \frac{N_2^{(p-1)N_2}}{\left(\frac{N_1}{p} + N_2\right)^{(p-1)N_2}} \right].$$

Employing Stirling's formula in the approximation (I, 3), we find:

$$\Delta S = -k N_1 \ln \frac{N_1}{N_1 + pN_2} - k N_2 \ln \frac{pN_2}{N_1 + pN_2}. \quad \text{(III, 37)}$$

If in the previous evaluation we had erroneously considered the polymer molecules as being equivalent to the micromolecules, then the earlier formula (III, 16) would again have been obtained, which in the same notation as (III, 37) has the form:

$$\Delta S = -k N_1 \ln \frac{N_1}{N_1 + N_2} - k N_2 \ln \frac{N_2}{N_1 + N_2}. \quad \text{(III, 38)}$$

This is obtained from (III, 37) by putting $p = 1$.

If, on the other hand, the pN_2 segments of the polymer molecules had been considered as separate molecules, then we would have obtained the relationship

$$\Delta S = -k N_1 \ln \frac{N_1}{N_1 + pN_2} - kpN_2 \ln \frac{pN_2}{N_1 + pN_2}. \quad \text{(III, 39)}$$

The conclusions from (III, 37) will be intermediate between those of (III, 38) and (III, 39).

Using equation (III, 30), we may also write (III, 37) as:

$$\Delta S = -kN_1 \ln \varphi_1 - kN_2 \ln \varphi_2, \quad \text{(III, 40)}$$

where $\varphi_2 = 1 - \varphi_1$.

This formula has a similar form to that of the earlier formula (III, 16) for the entropy of mixing; instead of the mole fractions x_1 and $x_2 = 1 - x_1$, however, here the volume fractions φ_1 and $\varphi_2 = 1 - \varphi_1$ occur.

From the derivation of formula (III, 37) it appears that any symmetry factor may be disregarded in (III, 31), since this factor would occur both in the expression for m_{12} and in that for m_2, so that it would be cancelled out in the expression for ΔS (cf. formula (III, 35)). For the same reason we may overlook the fact that for the 3rd, 4th ..., p^{th} segments of a polymer molecule less than $(z-1)$ sites are available, because a polymer molecule is not as a rule perfectly flexible and, moreover, one or more of the $(z-1)$ sites may be occupied by previous segments.

Several thermodynamical properties of solutions of polymers, such as osmotic pressure and miscibility can be understood by means of formulae(III,37), which also gives us a picture of the difference between this type of solution and that of common micro-molecular solutions.

IGNITION OF "PHOTOFLUX" FLASH-BULBS WITH THE AID OF A CAPACITOR

by J. A. de VRIEND.

771.448.4:621.319.4

In photography by means of "Photoflux" flash-bulbs, the "time to peak" or "time lag", is dependent on the total resistance of the circuit. The use of a higher voltage appreciably reduces this time lag, and the high voltage can be obtained from a capacitor in accordance with a method that is already widely used.

Ignition time, time to peak and circuit resistance

A large number of photographs are nowadays taken with the aid of flash-bulbs such as the "Photoflux".

The simplest form of ignition is obtained from a dry battery with switch (*fig. 1*) [1]), the latter being usually an electrical contact incorporated in the camera and synchronized with the shutter release.

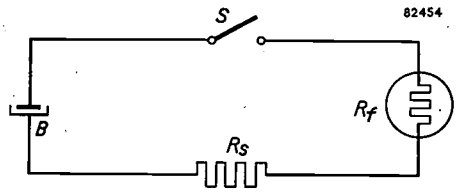


Fig. 1. Ignition of "Photoflux" flash bulb by means of a battery (B). R_f = resistance of igniter wire in the bulb, R_s = external resistance in series with bulb, S = switch.

The maximum emission of light does not occur immediately the circuit is closed; there is even a certain delay before the bulb gives any light at all, but subsequently the intensity rises very steeply (*fig. 2*). The time t_f which precedes the commencement of the light is necessary for the igniter wire in the "Photoflux" bulb to reach the required temperature, after which combustion occurs and light is produced. The time elapsing between the moment of closing the circuit and the flash peak is called the "time to peak" (t_m) in fig. 2).

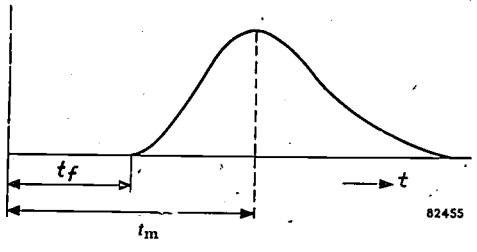


Fig. 2. After the circuit has been closed, a time t_f elapses before the igniter wire reaches the required temperature (heating time), and a time t_m before the flash reaches its maximum intensity (time to peak, or time lag).

Particularly for exposures of less than 1/50 sec, it is required that this time lag shall be roughly constant [2]), viz., according to current standards, 20 milliseconds. To ensure this, the ignition time t_f must not exceed a certain fraction of t_m, for example $t_f < 0.2\ t_m$. It is found that, from the moment the igniter wire reaches the firing temperature, "Photoflux" flash bulbs always attain maximum light emission, or flash peak, in roughly the same space of time. This means that, in fig. 2, $t_m - t_f$ is constant.

The requirement that t_f shall not be greater than $0.2\ t_m$ places a lower limit on the amount of power to be supplied to the igniter wire. If the total electrical resistance in the firing circuit is increased for one reason or another, the power delivered to the igniter wire by the battery is of course reduced. This means a longer ignition time t_f, possibly even beyond the permissible limit.

The total circuit resistance comprises the resistance of the igniter wire itself, the internal resistance of the battery and any other additional resistance that may be in the circuit. Whereas the resistance of the igniter wire is constant by reason of the manufacturing process, that of the battery increases as the battery runs down and extra resistances may be introduced by dirty switch contacts or poor connections. If the leads are longer than usual, this will also increase the resistance.

If a capacitor with resistor in series is added to the circuit shown in fig. 1, a practice that has been more and more widely adopted in recent years, several improvements are assured, the most important of these being that higher circuit resistances can be tolerated, without exceeding the specified firing time. A simple calculation will illustrate the difference between ignition by means of a battery only and that by means of a battery with capacitor.

[1]) For a description of "Photoflux" flash bulbs see Philips tech. Rev. **12**, 185-192, 1950/51 and **15**, 317-321, 1953/54. For a more general review see G. D. Rieck and L. H. Verbeek, "Artificial light and Photography", Philips Technical Library, 1950.

[2]) See first article mentioned in note [1]), p. 191.

Direct battery ignition

Let us assume that the resistance of the igniter wire in the circuit in fig. 1, (R_f) is constant. This is of course not strictly correct, in view of the fairly considerable rise in temperature (and hence also in the resistance), but for a comparison of the two methods of ignition the assumption is permissible and greatly simplifies the calculation.

If R_i is the internal resistance of the battery, R_s the external series resistance, E_b the e.m.f. of the battery, and P_f the power supplied to R_f, we have:

$$P_f = R_f\left(\frac{E_b}{R_f+R_i+R_s}\right)^2 \quad \ldots\ldots (1)$$

The question is now: what should be the minimum value of P_f to ensure that the flash bulb ignites within the specified time?

Disregarding the loss of heat to the surroundings during the very short heating time (only a few millisec) the minimum amount of power will be determined by the maximum permissible heating time t_{fmax} and the required heating energy Q_f (product of thermal capacity and ignition temperature of the igniter wire):

$$P_f \geqq \frac{Q_f}{t_{fmax}}.$$

Care in manufacture ensures that Q_f is practically constant. With (1) we can therefore write:

$$E_b \geqq \left(1+\frac{R_i}{R_f}+\frac{R_s}{R_f}\right)\sqrt{\frac{R_f Q_f}{t_{fmax}}} \quad \ldots\ldots (2)$$

With the following values: $R_f = 2$ ohms, $t_{fmax} = 3.5 \times 10^{-3}$ sec and $Q_f = 3.5 \times 10^{-3}$ joules, equation (2) becomes:

$$E_b \geqq \tfrac{1}{2}\sqrt{2}\,(2+R_s+R_i).$$

This relationship between E_b, R_i and R_s is shown graphically in *fig. 3*. A battery voltage of, say, 4.5 V, with $R_i = 2$ ohms, gives, from the graph, a maximum permissible series resistance of 2.4 Ω. If the series resistance is any higher the maximum heating time will be exceeded.

The longer the heating time, the more significant the heat that is lost to the surroundings, as this in turn increases t_f. In the extreme case, with very high R_s (and/or R_i) these losses will be equal to the applied power before the igniter wire reaches the working temperature, with the result that the bulb cannot ignite at all.

For a higher resistance to be permissible, E_b (fig. 3) must be increased and/or R_i reduced. It should be remembered, however, that there is a relationship between the e.m.f., the internal resistance and the volume of the battery. In practice, batteries having an e.m.f. of at most 4.5 V are generally used. With a higher voltage, either the volume of the battery is too great, or the internal resistance too high. Moreover, R_i increases as the battery runs down, so that the permissible amount of extra series resistance becomes less and less in the course of time.

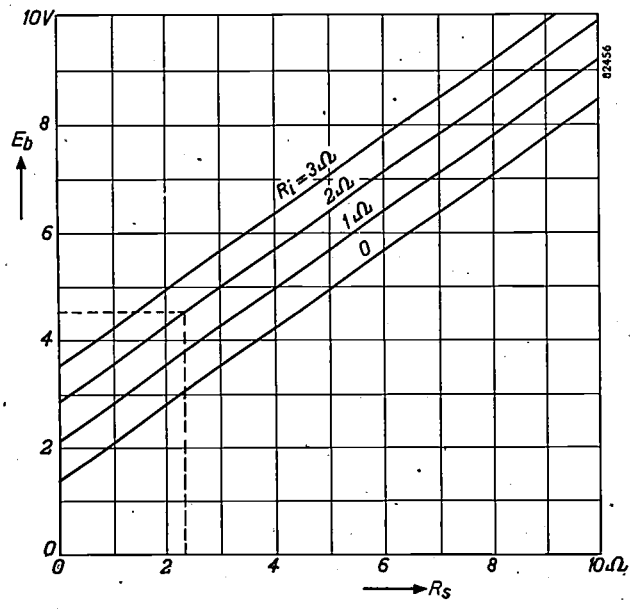


Fig. 3. Calculated values of the maximum permissible series resistance R_s plotted against battery voltage E_b for various battery internal resistances R_i, for direct battery ignition.

There are accordingly many objections to direct battery ignition, but these can largely be eliminated by using a capacitor with series resistor.

Ignition with the aid of a capacitor

The circuit incorporating a capacitor C and series resistor R_v is shown in *fig. 4*.

The internal resistance of the battery does not affect the ignition and therefore no longer causes the power supplied to vary, so that a battery having a high internal resistance can be used, thus making feasible a higher voltage battery, within the limits of a convenient volume.

Another advantage is that the resistance in series with the battery may also be higher, which means such a small load on the battery that its life will be almost as long as with no load at all. The charging current for the capacitor is certainly small, but there should be no objection to this as there is usually sufficient time for recharging (at least a few seconds).

Now, what are the conditions to be imposed on the ignition circuit R_f, R_s and C?

Let E_c and E denote the values of the capacitor voltage at the commencement and end of the discharge time t; then:

$$E = E_c\, e^{-\frac{t}{(R_f+R_s)C}}.$$

The initial energy $\frac{1}{2} C E_c^2$ of the charged capacitor is reduced in a time t by an amount

$$Q_c = \tfrac{1}{2} C\,(E_c^2 - E^2)\,.$$

This energy is dissipated in the resistors R_f and R_s, whereby R_f receives a fraction equal to $R_f/(R_f + R_s)$. Since a heating energy at least equal to Q_f must be delivered in the maximum permissible heating time t_{fmax}.

$$\tfrac{1}{2} C E_c^2 \left[1 - e^{-\frac{2t_{fmax}}{(R_f+R_s)C}}\right] \frac{R_f}{R_f + R_s} \geqq Q_f \quad . \quad . \quad (3)$$

With some re-arrangement and substitution of the values of t_{fmax}, R_f and Q_f we now write:

$$E_c \geqq \sqrt{\frac{0.0035\,(2 + R_s)}{C\left(1 - e^{-\frac{0.007}{(2+R_s)C}}\right)}}.$$

The equality relation of this expression is shown graphically in *fig. 5*, from which we can at once see the advantage of capacitor ignition in comparison with direct battery ignition. Note, for example that at 22.5 V (a practical value) up to 15.7 Ω in series can tolerated with a capacitor of 130 μF. It is also seen that it is important to use the largest possible capacitance compatible with the dimensions of the capacitor. At 22.5 V and with a 700 μF capacitor the permissible series resistance rises to 27 Ω.

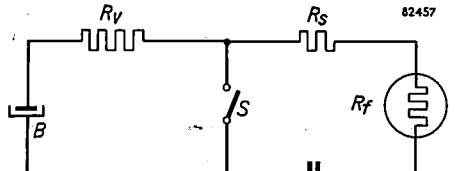


Fig. 4. Ignition with the aid of a capacitor (C). This capacitor is charged by a very small current through the resistance R_f of the "Photoflux" bulb and the high series resistance R_v. When the circuit is closed, C is discharged across R_f. R_s is the extra series resistance present.

In order better to illustrate the curves plotted from equation (3) let us consider two extreme cases.

Writing z as the exponent of e in (3), consider first:

$$z = \frac{2t_{fmax}}{(R_f + R_s)\,C} \ll 1. \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

The term in square brackets in equation (3): $F = 1 - e^{-z}$ can be expanded as a series, viz. $z - \frac{1}{2} z^2 + \ldots$ For a small value of z the first term of this series is sufficient, and (3) then becomes:

$$E_c^2 \geqq \left(1 + \frac{R_s}{R_f}\right)^2 \frac{R_f Q_f}{t_{fmax}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3a)$$

This is the same as equation (2) except that R_i is now missing, which means that a very large capacitor, for which condition (4) is certainly met, can be regarded as a battery without internal resistance. The limiting line $C = \infty$ in fig. 5 is given by equation (3a), and is at the same time the line corresponding to battery ignition when $R_i = 0$.

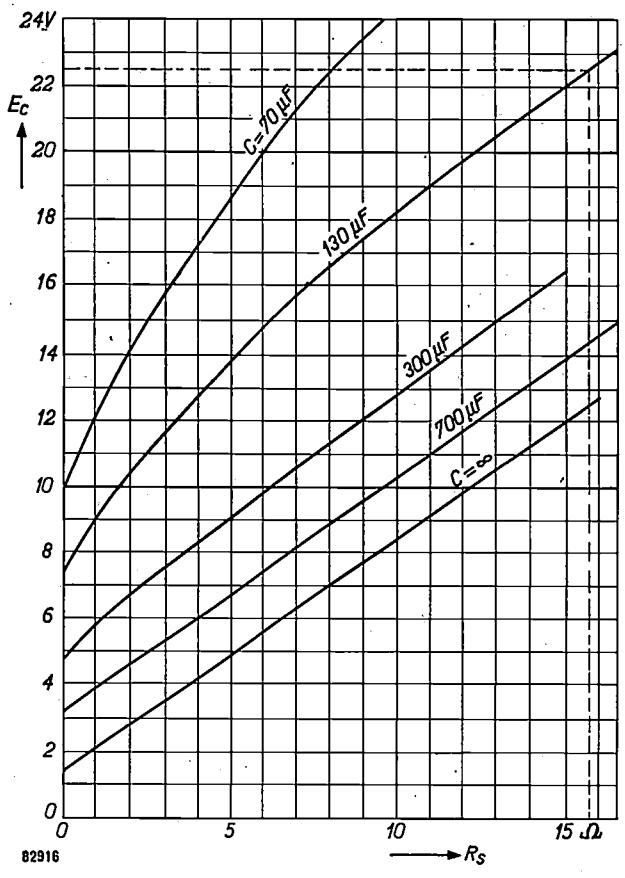


Fig. 5. Calculated values of the maximum permissible resistance R_s in series with the bulb, plotted against the battery voltage E_c for various capacitances C, for capacitor ignition.

The second extreme occurs when e^{-z} is very small (capacitor almost wholly discharged). For instance, with

$$z = \frac{2\,t_{fmax}}{(R_f + R_s)\,C} > 5 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

e^{-z} is <0.01. Formula (3) then takes the form

$$E_c^2 \geqq \frac{2Q_f}{C}\left(1 + \frac{R_s}{R_f}\right). \quad . \quad . \quad . \quad . \quad . \quad . \quad (3a)$$

In that part of the graph where (5) applies, the curves are therefore parabolas in which C occurs as parameter. The slope increases with decreasing C.

Practical results

Fig. 6 illustrates the results of a number of measurements made with direct battery ignition (with an accumulator, i.e. $R_i \approx 0$), and capacitor ignition. For these measurements a known resistance R_s was included in series with the circuit, and the voltage required to raise the igniter wire to the correct temperature in a time t_{fmax} was measured.

The qualitative agreement of figures 3 and 5 with fig. 6 is apparent on examination; quantitatively,

the curves approximate closely to quations (2) and (3). The line for direct battery ignition, with $R_i = 0$, is shown in fig. 6 as $C = \infty$, as this extreme case corresponds to that condition. It deviates more from the calculated curve for $C = \infty$ (dotted) as R_s is smaller. This is owing to the fact that the variation in R_f with increase in temperature represents a greater percentage change in the total resistance at small values of R_s.

In practice, owing to the inevitable amount of spread, the voltages to be employed will be slightly higher than those measured experimentally. Moreover, as the lines in fig. 6 are not very curved, the relationship between E_c, R_s and C can conveniently be represented by straight lines. This gives us *fig.* 7, from which the resistance R_s can be found for given values of E_c and C to yield the specified heating time of 3.5 msec.

In the equally important case of the simultaneous firing of two flash-bulbs, similar calculations are

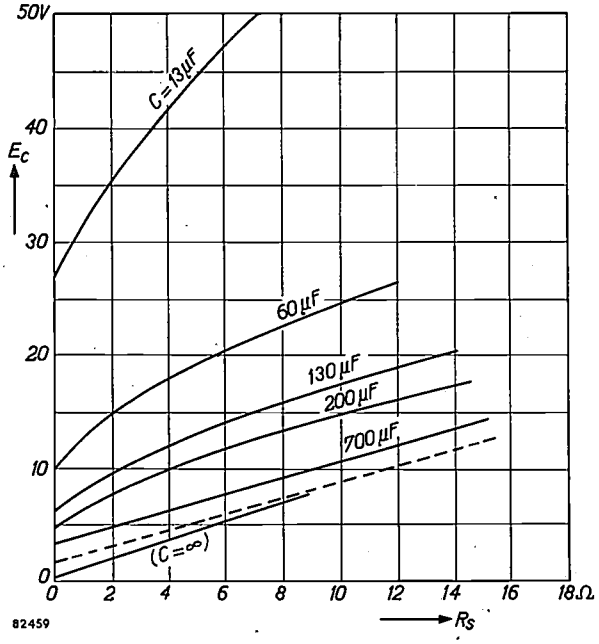


Fig. 6. Maximum permissible series resistance R_s plotted against battery voltage E_c for various capacitances C, as obtained experimentally for capacitor ignition. The line denoted by $C = \infty$ refers to direct battery ignition with $R_i = 0$. The broken line is the calculated curve for the latter case.

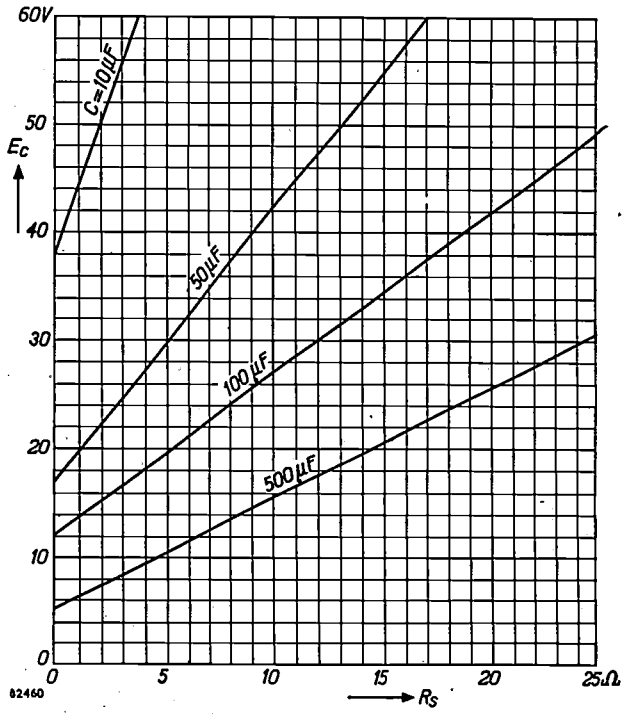


Fig. 7. Capacitor ignition chart for practical use. (Maximum ignition time 3.5 millisec).

applicable. Capacitor ignition is quite practicable provided that the capacitance is high enough, for the discharge is now divided between two circuits (flash bulbs in series are not to be recommended). It will be seen that in this case the series resistor in each circuit may have to conform to a lower limit as well as an upper one, in order to guarantee a sufficiently uniform distribution of the power. However, a discussion of this point is not within the scope of this article.

Summary. When the simplest method of firing a "Photoflux" flash bulb is employed, viz. direct battery ignition, only a small amount of extra resistance in the circuit is permissible if the time lag is to remain sufficiently constant. With higher battery voltages than the customary 4.5 V, the position is better but, as the internal resistance of the battery must be low, again with a view to a constant time lag, the battery would then be too large in size. The well-known method of ignition by means of a capacitor eliminates this difficulty; a voltage of, say 22.5 V can be employed, with a fairly high resistance in series with the bulb. This method is examined quantitatively, and curves are given from which the maximum permissible series resistance for given battery voltage and capacitor value can be read.

DEMONSTRATION OF THE AUSTENITE-PEARLITE TRANSFORMATION BY MEANS OF THE EMISSION ELECTRON MICROSCOPE

621.385.833:669.112.227.32

In normal electron microscopy the object is irradiated with electrons emitted from a filament. The examination of metals and alloys using a normal electron microscope is therefore largely restricted to pictures of very thin replicas of the metal surface. An important advance, permitting the study of metals at high temperatures, is the development of the emission electron microscope in which the object itself acts as the electron source.

The temperature region in which the observations are to be made is mainly determined by the phenomena to be investigated. Unfortunately, most metal surfaces begin to emit electrons in quantity only at very high temperatures. An adequate emission at relatively low temperatures can be obtained by evaporating onto the surface a very thin layer of electropositive atoms, for example barium or caesium, which lower the work function of the surface of the metal. At the same time, the polycrystalline structure of the metal surface is made visible, because the evaporated atoms are absorbed differently according to the lattice orientations of the various surface crystallites. A change in this structure, for example, on recrystallization, can therefore be directly observed.

An emission electron microscope developed by Philips is shown in *fig. 1*. A number of investigations into recrystallization and phase-transitions in metals and alloys have been carried out with this instrument in recent years [1]). The instrument will be fully described in a later issue of this Review. Here it will only be mentioned that its construction follows as far as possible that of the EM 75 kV electron microscope to be described shortly in these pages [2]). Naturally the emission instrument does not have a filament, nor does it have a condenser lens. Further differences in construction include means by which the object is brought to high potential and the special steps taken in view of the higher vacuum required.

As an example of the type of work which may be done with this emission microscope, some photographs are given illustrating the austenite-pearlite transformation in 0.8 % carbon steel. The metal was prepared under high vacuum to minimize the effect of impurities [3]).

Fig. 2a shows the austenitic structure which is stable above 721 °C — a homogeneous solution of carbon in γ-iron; fig. 2*b* shows the way in which the pearlite grows in the austenite. The pearlite is built up of alternate lamellae of α-iron (ferrite) and iron carbide (cementite). The many dark stripes

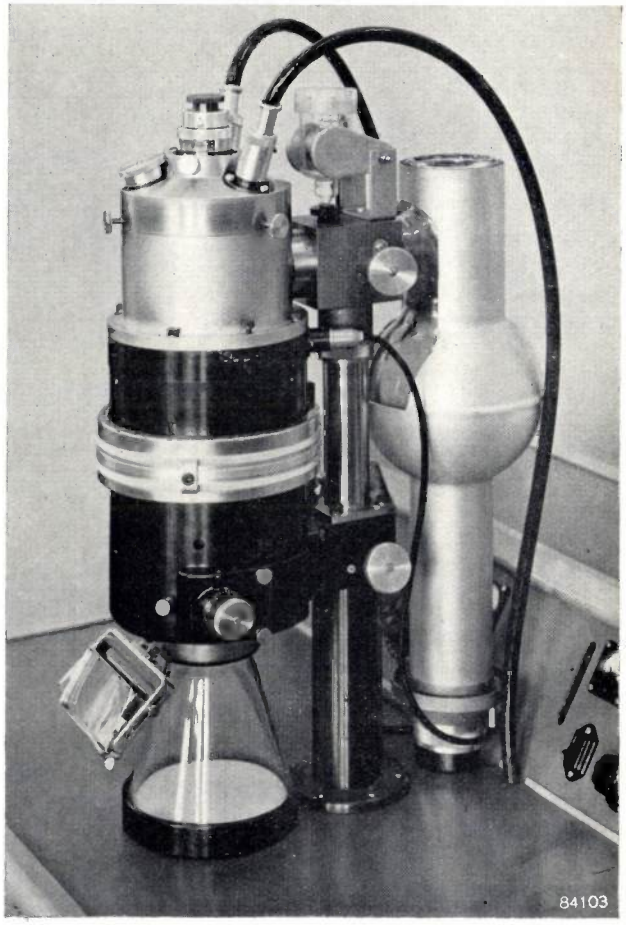
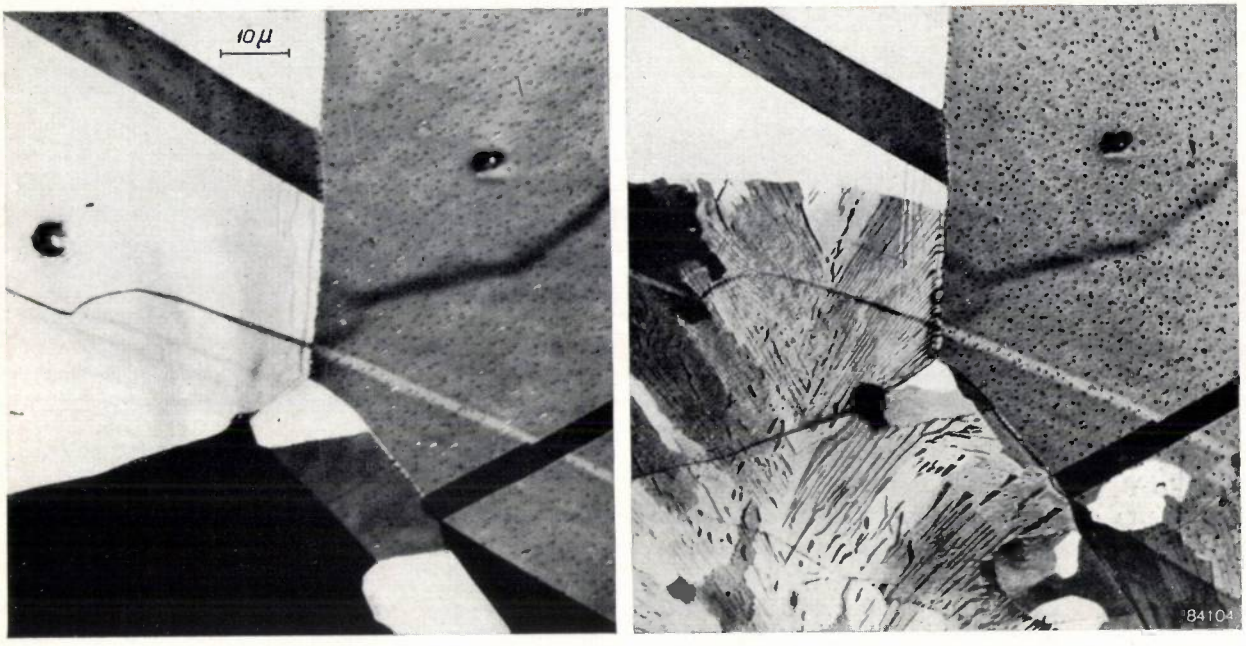


Fig. 1. The Philips emission electron microscope. The photograph shows the body of the instrument and a part of the vacuum system with its liquid air cooler. By turning a milled ring on the microscope body and by vertical displacement of the object, the magnification can be varied from 150 × to 3000 ×. The resolving power is 1000 Å. By means of a built-in camera, a series of 40 pictures can be taken in rapid succession on 35 mm film. If necessary a phenomena can also be filmed by means of a 16 mm cine camera placed under the desk; in this case a transmission fluorescent screen is used and the picture filmed through the glass of the screen.

[1]) G. W. Rathenau and G. Baas, Physica **17**, 117-128, 1951.
G. W. Rathenau, L'état solide, 9e Conseil de Physique Solvay, pp. 55-72, Brussels 1952.
G. W. Rathenau and G. Baas, Métaux **29**, 139-150, 1954 (No. 344).
G. W. Rathenau and G. Baas, Acta Met. **2**, 875-883, 1954 (No. 6).

[2]) A. C. van Dorsten and J. B. le Poole, The EM 75 kV, an electron microscope of simplified construction, to appear shortly in this Review.

[3]) J. D. Fast, A. I. Luteijn and E. Overbosch, Preparation and casting of metals and alloys under high vacuum, Philips tech. Rev. **15**, 114-121, 1953/54.

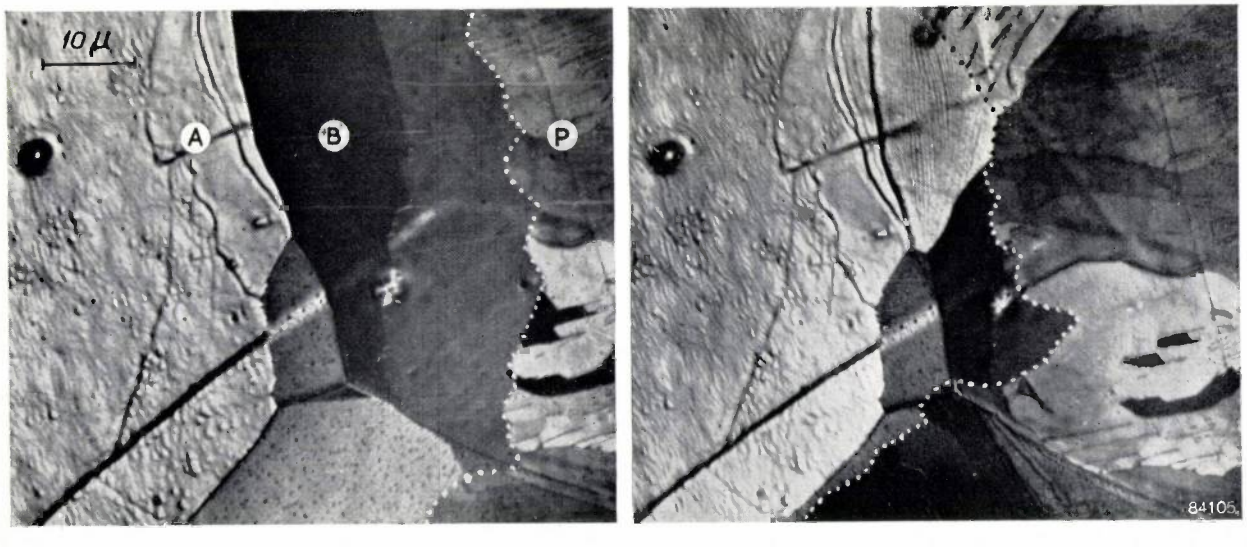


a *b*

Fig. 2. Two pictures from a series made during the conversion of austenite into pearlite in 0.81% carbon steel, at a temperature of about 700 °C. The interval between the two pictures is about 2 minutes.
a) Austenitic structure
b) Austenite partly changed into pearlite.

which can be seen in the pearlite are the cementite lamellae.

Figs 3a and *b* show a phenomenon often observed in the austenite-pearlite transformation in this carbon steel: if the pearlite, growing in the austenite matrix, comes in the neighbourhood of a boundary between two austenite crystals, this boundary tends to meet the approaching pearlite. This can be seen in the



a *b*

Fig. 3. The boundary between the two austenite crystals *A* and *B* moves in the direction of the approaching pearlite *P*. The dotted line shows the boundary between the pearlite and the austenite.
a) Initial situation. *b*) After the boundary has moved.
The presence of the many fine grooves in the part of *B* which has been consumed by *A* shows that the movement of the crystal boundary is a discontinuous process,

photographs at the boundary between the crystals A and B. The boundary between austenite and pearlite is shown with a dotted line.

In the part of crystal A which has grown in this way at the expense of crystal B, one can see many fine grooves in the surface, lying very close together (fig. 3*b*). Their presence shows that the displacement of the crystal boundary is a *discontinuous* processs: if the boundary stays still for a moment, such a groove forms as a result of the surface tension at the crystal boundary: the series of grooves shows the successive positions taken up by the crystal boundary.

G. BAAS.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Administration of the Philips Research Laboratory, Eindhoven, Netherlands.

2140: K. F. Niessen: Magnetic behaviour of some ferrites (Physica **19**, 1127-1132, 1953).

The variation of the Curie temperature of a series of mixed crystals of nickel ferrite and nickel titanate ($Fe_{2-2a}Ni_{1+a}Ti_aO_4$) with $0 < a < 0.5$, has been measured as a function of the content a of titanium, by E. W. Gorter of this Laboratory. From this data, the distribution of the titanium amongst the tetrahedral (A—) and octahedral (B—) sites of the spinel is derived on the assumption that nickel ions always occupy B-sites. The parameters necessary for this calculation were taken from the form of the Curie temperature curve of nickel zinc ferrite, measured as a function of the nickel content by Guillaud and Roux. The partial magnetizations in $Fe_{2-2a}Ni_{1+a}Ti_aO_4$ were taken parallel and antiparallel. In the case $a \ll 1$ the titanium appears to be distributed statistically (i.e. in the same ratio 1 : 2 as the numbers of A— and B— sites), but at an increasing content of titanium this metal appears to have an increasing preference for the A-sites.

2141: J. M. Stevels: Analyse statistique graphique du verre, II (Verres et Réfractaires **7**, 281-286, 1953).

The author shows that for glasses of the system Na_2O-CaO-SiO_2 there is a linear relation between the specific refraction $S = [(n_D^2-1)/(n_D^2+2)]/\varrho$ and the dispersion $D = n_F - n_C$, such that $S =$ constant $\times D + f(R)$. The function $f(R)$ is a function of R only (ratio of the number of oxygen ions to the number of Si^{4+}ions) and is independent of the content of Na^+ and Ca^{++} ions. If qualitative analysis shows that the glass contains only the above-mentioned ions, then the formula makes it possible, by a simple measurement of S and D, to determine the numbers of bridging and non-bridging oxygen ions. In an appendix the author discusses the theoretical basis of the formula.

2142: H. C. Hamaker: "Average confidence" limits for binomial probabilities (Rev. Int. Statist. Inst. **21**, 17-27, 1953).

A detailed investigation of the "confidence limits" of binomial probabilities as generally used, shows that these limits are fixed in such a way that in the most unfavourable circumstances the percentage of correct decisions is never below the confidence level. In industrial practice it should not be assumed that these unfavourable conditions are always present. Thus it is reasonable to modify the limits in such a way that the percentage of correct decisions is on average equal to the confidence level. The "average confidence" limits introduced in the present paper are designed towards this end. They determine a region of "average confidence" which is narrower than that generally used.

2143: F. van Tongerloo: Magnetic and dielectric elements for computers (T. Ned. Radiogenootsch. **18**, 265-285, 1953; in Dutch).

A review of the applications of magnetic and dielectric materials with square hysteresis loops. The first part of the article deals with circuits which depend on a high value of the ratio of remanance to saturation; the second part deals with circuits in which the squareness ratio is important. The article contains no original developments.

2144*: E. J. W. Verwey: New developments in synthetic ceramics (Proc. Int. Symp. on reactivity of solids, Gothenburg 1952, Elanders Bocktryckeri A.B., Göteborg, 1953, pp. 703-715).

Survey of some new ceramic materials, such as the magnetic ceramics "Ferroxcube" and "Ferroxdure", the systems TiO_2-Al_2O_3 and $NaCl_1$-$CaCl_2$ which have interesting dielectric properties, and semiconductors such as Fe_2O_3 - TiO_2.

2145: J. Smit and J. Volger: Spontaneous Hall effect in ferromagnetics (Phys. Rev. **92**, 1576-1577, 1953, No. 6).

The coefficients A_H (normal Hall coefficient) and ϱ_{SH} occurring in the equation of the Hall effect:

$$E_y/i_x = A_H . B_z + \varrho_{SH}$$

have been measured for 16 examples of Ni and Ni alloys in fields up to 14 000 gauss (1.4 Wb/m²) at three temperatures (20 °K, 77 °K, 290 °K). The "spontaneous Hall resistivity" ϱ_{SH} proves to be roughly proportional to the ordinary resistivity ϱ and smaller by a factor of the order (10^{-2}) and vanishes for the purest Ni sample at T = 20 °K. It is shown that this last fact is in accordance with the theory.

2146: H. P. J. Wijn: Frequency-dependence of magnetization processes in ferrites and its relation to the distortion caused by ferrite cores (Soft magnetic materials for telecommunications, Pergamon Press, London 1953, pp. 51-63).

Magnetization curves of ferrites have been measured as a function of frequency and it appears that two dispersion mechanisms occur. The relation between the dispersion frequency of the initial permeability μ_i and the value of μ_i at low frequency has been investigated for samples of nickel-zinc ferrite with increasing zinc-content; the greater part of μ_i is caused by a rotation of the spins in the Weiss domains. The dispersion of the permeability corresponding to magnetic fields of the order of magnitude of the coercive force of the ferrite always occurs at a lower frequency, and is attributed to a relaxation of the irreversible domain-wall displacements. As a consequence of the last-mentioned dispersion, the distortion caused by coils with a ferrite core can decrease with frequency, and even vanish. Total losses of the ferrite are given as a function of induction and frequency, and the relation between distortion and hysteresis resistance is discussed.

2147: J. D. Fast and M. B. Verrijp: Diffusion of nitrogen in iron (J. Iron and Steel Inst. **176**, 24-27, 1954, No. 1).

Diffusion coefficients of nitrogen in α-iron at 500° and 600° C are derived from the desorption rates of nitrogen from iron wires in hydrogen, using internal friction measurements to determine concentration ratios. Extrapolation to 21.5° and 9.5° gives values of D that correspond well with those calculated from internal friction measurements at these temperatures and thus strongly support the interpretation by Snoek and Polder of the damping caused by carbon and nitrogen. The combined measurements yield for the diffusivity of nitrogen in α-iron:

$$D = 6.6 \times 10^{-3} \exp(-18\,600/RT) \text{ cm}^2/\text{s}.$$

Determination of the diffusion coefficient of nitrogen in γ-iron at 950 °C shows that diffusion in γ-iron is much slower than in α-iron.

2148*: H. Bruining: Physics and applications of secondary emission (Pergamon Press, London 1954, xii + 178 pp., 129 figs.)

In this book a survey of the physics and applications of secondary electron emission is given. Physical aspects are discussed in the first seven chapters. Ch. 1: Introduction; Ch. 2: Methods and measurements; Ch. 3 & 4: Review of results (metals and compounds); Ch. 5: Influence of externally adsorbed foreign atoms and ions; Ch. 6 and 7: Mechanism of secondary electron emission. A complete theory does not yet exist; a survey is given of various approaches to the problem. The final three chapters deal with applications. Ch. 8: Electron multiplication; Ch. 9. Elimination of disturbing effects caused by secondary emission; Ch. 10: Storage devices (for storage of information in the form of electric charges on an insulating surface, utilizing secondary emission). A list containing about 400 references is included.

2149: K. ter Haar and J. Bazen: The titration of Al with "Complexone III" at *p*H 3.5 (Anal. chim. Acta **10**, 23-28, 1954, No. 1).

A titration procedure for Al, based on the reaction of Al and "Complexone III" (disodium salt of ethylenediamine-tetra-acetic acid, ("versene") at *p*H 3.5-4.3 is described. After adding an excess of "Complexone III", the excess is back-titrated with thorium nitrate, using "Alizarin S" as an indicator. The reaction is nor exactly stoichiometric, but nevertheless quite reproducible; the correction factor is about 1%.

VOL. 16 No. 12, pp. 341-372 JUNE 1955

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

FOREIGN ATOMS IN METALS

by J. D. FAST.

669.017

For many readers, the title of this article will perhaps already evoke the many publications by the same author, both in this journal and elsewhere. The cross-section of his field of study considered here by Prof. Fast — viz. interstitial solutions — formed the principal contents of his inaugural lecture on his appointment as extra-mural professor at Delft on 12th January 1955. The lecture is supplemented here by illustrations and references to the literature.

It is perhaps not always fully realized how important is the role of metals in our lives. Modern civilization would be out of the question without the generation and distribution of electrical energy, without the production and distribution of town-gas and drinking-water, without motor traffic, trains, ships and aircraft, without printing and communications, the telegraph, telephone, radio, television and the film. Our civilization could not exist without surgery, without large industries employing all manner of machines, apparatus and measuring instruments, without bridges, cranes and docks, without electric lighting, vacuum cleaners and typewriters. If there were no metals at our disposal, none of these could have been developed to their present advanced state. Metals played an important role even in primitive communities, especially for making weapons, tools, jewellery and coinage. Indeed, the vital role of certain metals is illustrated by the fact that the history of mankind distinguishes a copper age, a bronze age and an iron age.

Of the hundred or so elements more than two-thirds are metals. In ancient times only seven of these metals were known: copper, silver, gold, mercury, tin, lead and iron. By about 1900 many more were known, but only five of the newcomers have found their way into the ranks of the metals of industrial importance: magnesium, zinc, aluminium, antimony and nickel. Aluminium, the most important of the five, can be produced economically on a large scale only by electrolytic methods; it thus had to wait for the birth of electrical technology in the latter half of the last century.

It remained for the present century to see the application of the mass of the metallic elements. Besides the dozen metals which have been named, the following metals are among those which are now used industrially on a considerable scale: sodium, caesium, beryllium, calcium, barium, cadmium, indium, titanium, zirconium, thorium, germanium, vanadium, niobium, tantalum, chromium, molybdenum, tungsten, uranium, manganese and cobalt. The availability of all these metallic elements has led to the development of a vast number of useful alloys; a survey of the latter is quite impracticable. The metallurgist has developed all these metals in order to meet the demands of an ever-expanding technology: higher melting points, greater mechanical strength, special electrical resistance characteristics, special magnetic properties, improved corrosion resistance, etc.

Hand in hand with and as a necessary condition for this great expansion, metallurgy has developed from purely empirical techniques to an exact science: originally it comprised only extraction metallurgy (the preparation of metals from their ores), the working of metals by casting and mechanical deformation, heat treatments and mechanical testing. Metallography, the microscopic study of the structure of metals, was added in the last century.

Round about the turn of the century, a considerable advance was made when Bakhuis Roozeboom, whose centenary has been recently celebrated, demonstrated how Gibbs' phase rule may be applied to alloys. Van Laar carried this application of thermodynamics to metallurgy a stage further, by showing that the simplest types of binary phase diagrams can be completely derived from the heats of mixing and entropies of mixing for the liquid and solid states and the heats of fusion of the two components. Van Laar's important work, however, remained for ten years almost unnoticed. After 1912 the metallurgist learnt to think in terms of atoms from the pioneer work of Von Laue and the two Braggs. These investigators demonstrated with the aid of X-rays, that all crystals, and hence also the crystals from which metals are built up, are orderly, periodic configurations of atoms; further, they showed that the distances between the atoms in these configurations may be quantitatively determined. Structural investigations by means of X-ray diffraction have been continued in many laboratories, throwing new light on phase diagrams.

Initially it was expected that it would be possible to explain all the properties of a crystalline material from the structure determined from diffraction patterns. It was thought that local deviations from the perfect lattice structure would have only a slight influence on the properties. Gradually, however, it has come to be realised that many of the most interesting properties of crystals owe their origin to just these deviations from the perfect atomic configurations. This applies, for example, to the plastic deformation of metals which, according to modern theory, is due to linear imperfections in the lattice, known as dislocations. Mass transfer (diffusion) in solid metals also seems to be due to the presence of imperfections in the lattice. Point imperfections are responsible for this phenomenon, i.e. extra atoms squeezed into the orderly configuration or absent atoms (vacancies) at points where they are normally present.

Apart from the development of new metals and alloys, the metallurgist's tasks include learning to control these lattice imperfections and to employ them usefully for the improvement of the properties of the metals. A number of phenomena will now be discussed, which depend upon one type of imperfection, viz. the presence of extra atoms (interstitials) in the periodic lattice.

Examples of interstitial solutions

A metal in equilibrium can take up foreign atoms in its interstices only when the foreign atoms are relatively small. This is especially the case where the foreign atoms are certain non-metallic elements of low atomic number: hydrogen, boron, carbon, nitrogen and oxygen. The latter two elements have played a striking part in the history of the metals titanium and zirconium. All ingots made from these metals some thirty years ago were found to be brittle, even those ingots, which had been shown by chemical analysis to be very pure. Consequently it was long assumed that brittleness was an intrinsic property of titanium and zirconium. However, in 1924 a method was discovered in Eindhoven for preparing these metals in a highly ductile form [1]). The ductile metal appeared to have the same crystal structure as the brittle form, so that it had to be assumed that the latter contained one or more impurities which had not been detected by normal chemical analysis. Extensive investigation into this matter showed that titanium and zirconium could contain a considerable atomic percentage of oxygen in solid solution, and that even a few percent were sufficient to render either metal hard and brittle [2]). Nitrogen also appeared to be highly soluble in both metals and to have in the dissolved state an even more unfavourable influence on the mechanical properties than oxygen. Both non-metals are absorbed in atomic form in the largest interstices (octahedral spaces) of the hexagonal metal lattice. In the titanium and zirconium ingots prepared by the early methods it was especially oxygen which was present in an undesirably high amount. Now that this action of oxygen and nitrogen is known, it is possible in principle to add them in controlled quantities to titanium and zirconium to render the latter suitable for applications for which the pure metals are too soft. The amounts of oxygen or nitrogen must either be chosen so small that the ductility of the metal is not completely lost, or the hardening must be confined to the surface layers.

Other metals are likewise rendered hard and brittle by interstitially dissolved atoms. Hardening by means of interstitial atoms has been unknowingly employed for many centuries in the hardening of steel. In this case the hardening element is carbon. In the hardening process this element is dissolved to form a highly supersaturated solution, making use of the fact that its solubility in the iron phase stable above 900 °C (γ iron) is far greater than its solubility in the phase stable below this temperature (α iron).

---

[1]) J. H. de Boer and J. D. Fast, Z. anorg. allg. chem. **153**, 1-8, 1926.

[2]) J. H. de Boer and J. D. Fast, Rec. Trav. chim. Pays-Bas **55**, 459-467, 1936; J. D. Fast, Metallwirtschaft **17**, 641-644, 1938.

Titanium and zirconium have crystallographic transition points at about the same temperature as iron, but the technique used in the hardening of steel may not be employed here since oxygen and nitrogen dissolve far more readily in the Ti and Zr phases stable at low temperatures than in those stable at high temperatures. The same is true for carbon which, however, has a lower solubility in both metals than either oxygen or nitrogen.

This inverse effect of the crystallographic transition on interstitial solubility is illustrated by the binary phase diagrams for Fe-C, Fe-N compared with those of Ti-O, Ti-N, ... etc: the presence of interstitial atoms shifts the crystallographic

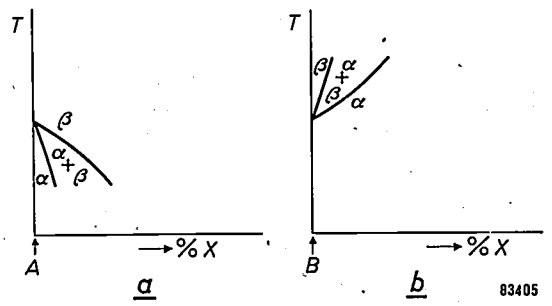


Fig. 1. *a*) The melting point or a crystallographic transition point of a component (*A*) moves to lower temperatures by adding a certain component (*X*), since the solubility in phase α stable at low temperatures is less than the solubility in phase β stable at high temperatures.
b) In the reverse case (*B-X* system) the melting point or or transition point is shifted to higher temperatures.

transition for iron to lower temperatures whilst that for titanium and zirconium moves to higher temperatures. This is schematically represented in *fig. 1*. *Figures 2* and *3* show larger parts of the phase diagrams for two of the binary systems discussed, viz. the Fe-C and Zr-O systems.

Factors determining the interstitial solubility

The facts mentioned above seem to indicate a connection between the size of the available interstices and the solubility of carbon, oxygen and nitrogen. The crystal structure as determined by X-ray diffraction, shows that the γ phase of iron has larger interstices than the α phase, and it can indeed dissolve more carbon and nitrogen; in the case of Ti and Zr however it is the phases stable at lower temperatures which have the larger interstices and which can dissolve the greater quantities of C, N and O. The introduction of a carbon, nitrogen or oxygen atom into an interstice of a metal involves an appreciable local strain in the lattice. The smaller the size of the interstice relative to the interstitial atom, the more will the lattice be deformed and the greater will be the deformation energy, required

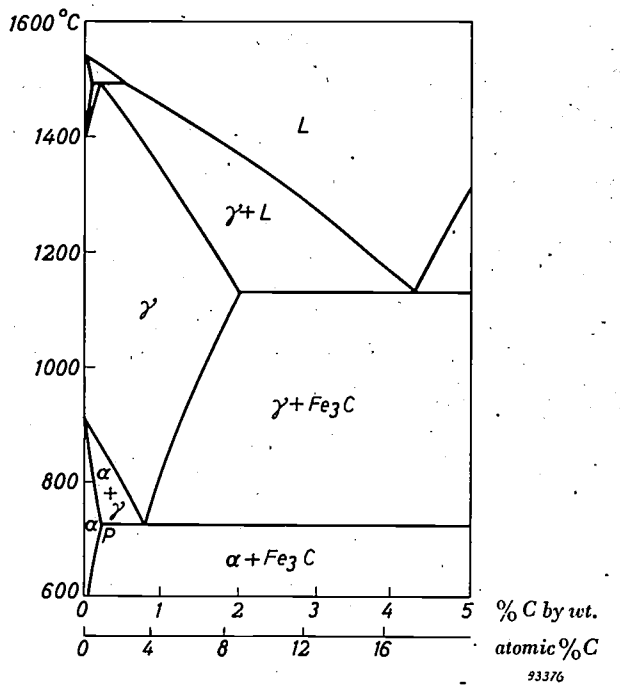


Fig. 2. A part of the iron-carbon phase diagram. For the sake of clarity the distance of P from the % $C = 0$ axis has been exaggerated; in reality the maximum solubility of carbon in α iron corresponding to this point is only 0.02 % by weight. L is the liquid phase. In contrast to fig. 1, the iron phase stable at high temperatures (above 900 °C) is designated γ iron and not β iron; this has arisen historically, since earlier it was incorrectly thought that the transition of iron at about 760 °C from the ferromagnetic to the paramagnetic state corresponded to a transition into another phase which was designated at the time the β phase.

to place it there. This greater deformation energy will mean a smaller solubility of the interstitial atoms.

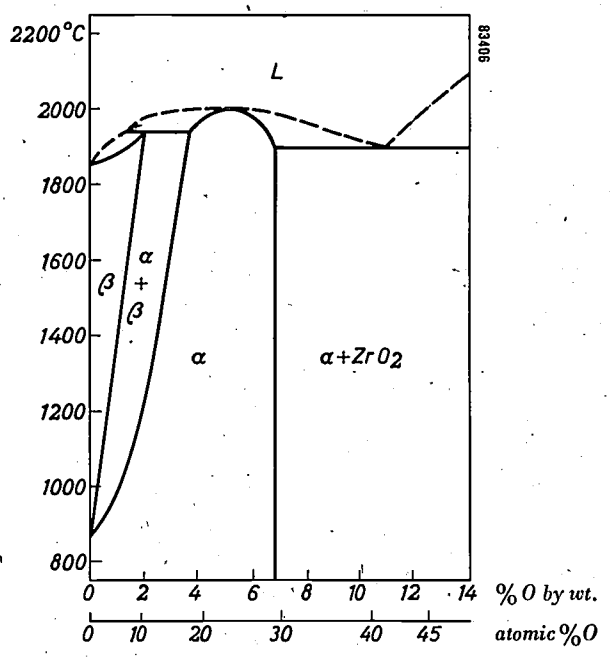


Fig. 3. A part of the phase diagram for zirconium-oxygen (From Domagala and McPherson, J. Metals **6**, 238-246, 1954).

There is a temptation to generalize the foregoing by supposing that there is a general relation between the interstitial solubility of foreign atoms and the ratio of the size of the latter to that of the interstices. Closer consideration shows that the supposition that the solubilities are determined exclusively by geometrical factors does not correspond with the facts. Even if we ignore the fact that entropy changes during solution also play a part in determining the solubility, we must nevertheless take into account the existence of two other factors besides the elastic deformation energy: the electronic interaction energy and the stability of the coexistent phase — the source of the interstitial atoms. Let us now briefly examine these two additional factors.

The significance of the electronic interaction energy is shown, for example, by the fact that helium atoms, which are smaller than either carbon or nitrogen atoms but have a highly stable configuration of two paired electrons, do not dissolve in any metal. That the interstitial atoms must be relatively small is thus a necessary but by no means sufficient condition. The heat of solution can be expressed, to a rough approximation, as the sum of the elastic deformation energy, which is always positive (absorption of heat during isothermal solution) and the electronic interaction energy, which can be either positive or negative. Strictly speaking, this division of the energy of solution into two terms is not justifiable, since the interaction energy depends to a large extent upon the interatomic distances and hence upon the deformation energy. Such a division however, is useful for many considerations. For the sake of brevity we will denote the two terms strain energy and chemical energy. A typical example of those cases in which the chemical energy is dominant is the solubility of oxygen in copper, silver and gold. In silver and gold the interstices are larger than those in copper, yet the solubility of oxygen in these metals at a given pressure decreases in the order Cu, Ag, Au, corresponding to their decreasing affinities for oxygen [3]).

The influence of the stability of the coexistent phase is illustrated by a comparison of the solubilities of carbon, nitrogen, and oxygen in iron. Although the size of the C, N and O atoms decreases in that order, the solubility of oxygen in both modifications of iron, when FeO is the second phase, is appreciably less than that of nitrogen and carbon, when Fe_4N and Fe_3C respectively are the second phases. This is no doubt largely due to the greater stability of the oxides of iron compared to the nitrides and carbides.

[3]) J. L. Meijering, Acta Metallurgica, in the press.

The strain energy

A striking picture of the important influence of the strain energy may be obtained from a study of the effects of carbon and nitrogen in iron. From purely geometrical considerations, it is possible to see a connection between many of the phenomena occurring in ordinary mild steel, which at first sight appear to have little or nothing in common. One of these phenomena, viz. the greater solubility of carbon and nitrogen in γ iron, as compared with their solubilities in α iron, has already been mentioned. The influence of the strain energy also manifests itself in each deformation of the lattice, which enlarges some of the interstices, while diminishing others. Such a deformation leads to a change in the distribution of the dissolved C or N atoms, such that the occupation of the contracted interstices decreases and that of the enlarged interstices increases. The fall in the entropy, which accompanies this departure from the random distribution, is overcompensated by a simultaneous fall in the energy, so that a net fall in free energy results from the redistribution.

Deformation by internal stresses

Deformation of the lattice, leading to redistribution, can be caused by internal as well as by external stresses. A familiar source of internal stresses are dislocations, the linear lattice defects already mentioned; these are responsible for the ready plastic deformation of metals. On one side of the slip plane of an edge dislocation, the iron lattice is stretched and on the other side of the slip plane it is compressed (see *fig. 4*). The deformation rapidly decreases as the distance from the dislocation increases. The introduction of a C or N atom into an enlarged interstice of the stretched zone will require

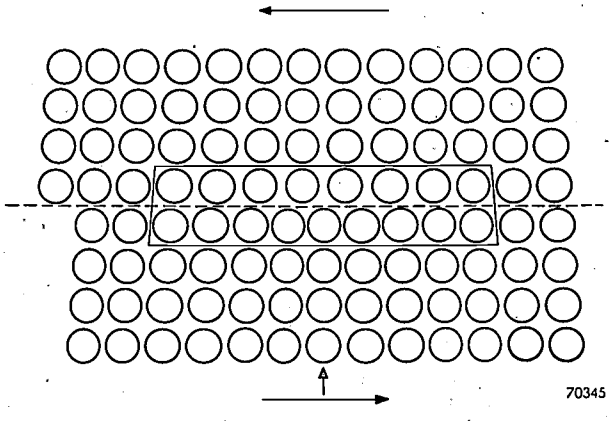


Fig. 4. Schematic respresentation of an edge dislocation in a simple cubic metal lattice. The dislocation is a linear imperfection in the lattice in a direction perpendicular to the plane of the diagram. Its effect on one atomic layer is shown here (see the enclosed region). The dotted line represents the plane in which slip can occur.

less deformation energy than introduction into a normal interstice. Dissolved C or N atoms which arrive at the dilated region by diffusion will be no longer able to leave this zone at low temperatures, e.g. room-temperature. When there are enough carbon or nitrogen atoms present in the lattice, chains of interstitial atoms will be formed along the whole length of each edge dislocation. Formation of these chains or strands greatly influences the plastic properties at ordinary temperatures, since plastic deformation (as, for example, in the tensile test)

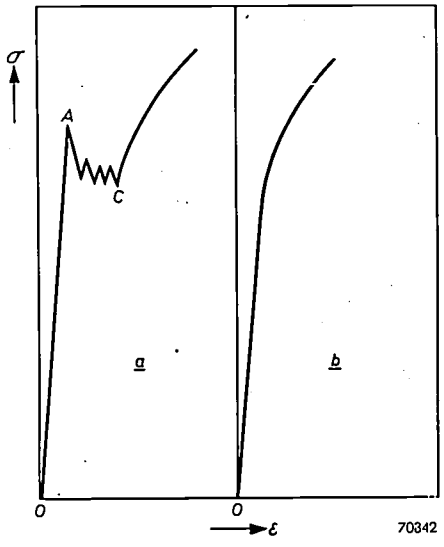


Fig. 5. Stress-strain curve for mild steel. The tensile stress σ is plotted as a function of the elongation ε.
The steel shows a sharp upper yield point at A (fig. 5*a*). If it is plastically stretched to C and the stress removed, then on renewed loading the material shows a stress-strain curve (*b*) without a yield point.

must be preceded by a freeing of the dislocations from the atom strands. This separation brings the metal into a higher energy state, the dissolved atoms being located, after the separation, in normal interstices. An extra stress is therefore initially required to move the dislocations, but after the C or N atom strands have been freed, this stress is no longer needed. This explains the occurrence of an upper and a lower yield point in the stress-strain curve for mild steel (*fig. 5*). Directly after a small plastic deformation the dislocations are still free; a well-defined yield point is lacking on renewed deformation (fig. 5*b*). If the metal is then allowed to stand, the carbon and nitrogen atoms once more diffuse towards the dislocations, as a consequence of which the upper and lower yield points are restored and the metal becomes more difficult to deform (harder). This spontaneous process has been given the name of strain ageing. The theory of the occurrence of an upper and a lower yield point and of strain ageing was originated by the English worker A. H. Cottrell [4]).

Experiments in Eindhoven [5]) have shown that strain ageing can also occur even if the C or N are not dissolved in the iron but present merely in the form of small crystals of Fe_3C or Fe_4N. Under suitable conditions, the C or N atoms move from these crystals into the iron lattice where they may reach the dislocations by diffusion. The experiments show that the C or N atoms are more strongly bound in dislocations than in the Fe_3C or Fe_4N crystals. This is the more remarkable since — as is well known — they are less strongly bound in normal interstices than in Fe_3C or Fe_4N crystals. These facts clearly show the importance of the strain energy. If the carbon or nitrogen is present in the steel only in the form of a very stable carbide or nitride (e.g. TiC or TiN) where it is bound more strongly than in a dislocation, then no strain ageing occurs.

Deformation by applied stresses

Applied stresses can also lead to a redistribution of the interstitial atoms in α-iron. This phenomenon is connected with the asymmetry of the interstices. The dissolved atoms are sited in the octahedral interstices in the iron lattice, whose centres coincide with the middle points of the edges and faces of the unit cell. The six atoms which enclose the octahedral interstice in α-iron form an irregular octahedron (one body diagonal shorter than the other two, see *fig. 6*). A carbon or nitrogen atom in such an

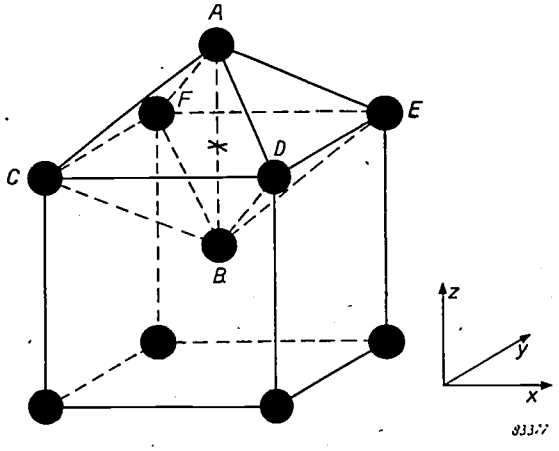


Fig. 6. The cross in the diagram indicates the location of the centre of an octahedral interstice in the unit cell of a body-centred cubic metal. Such interstices are not symmetrical with respect to the x, y and z direction: the distances CE and DF are $\sqrt{2}$ times greater than the distance AB.

[4]) A. H. Cottrell, Progress in Metal Physics **1**, 77-126, 1949.
[5]) J. D. Fast, Revue de Métallurgie **47**, 779-786, 1950; Philips tech. Rev. **14**, 60-67, 1952/53.

interstice therefore exerts a greater force on two of the six iron atoms and causes a unilateral elongation in the direction of the axis defined by these two atoms. Denoting the axes of the unit cell x, y and z, the interstices can be divided into three groups: x, y and z interstices. In an undeformed crystal, the interstitial atoms are distributed equally over the three types of interstices. This does not mean that they occupy fixed positions; the average duration of stay in any one interstice is only about a second at room-temperature. If now the crystal is stretched elastically in the direction of the x axis, there will be a preference for the enlarged x positions and a shift in the distribution equilibrium, so that the occupation of the x positions increases while that of the y and z positions decreases. This redistribution takes a certain time and leads to an extra elongation which occurs after the momentary elastic elongation and gradually reaches a limiting value (*fig. 7*). The rate at which this occurs is determined by the rate of diffusion of the dissolved C and N atoms, i.e. by the above-mentioned frequency of movement between sites — about 1 per second at room-temperature. This phenomenon is known as elastic after-effect. With an alternating load, this leads to a phase shift between load and deformation and thus to a dissipation of elastic energy, i.e. to damping of free vibrations (*fig. 8*).

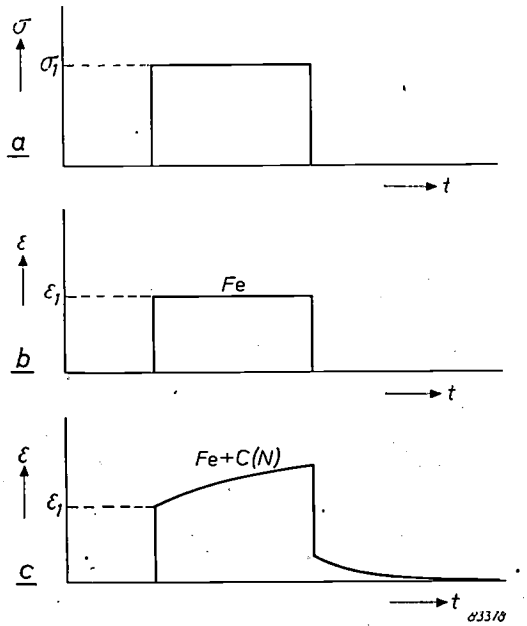


Fig. 7. *a*) A constant tensile stress σ_1 (which is considerably below the yield point) is applied at a certain instant to an iron crystal in the direction of one of the axes of the cube. *b*) The metal responds at once, showing a certain elastic elongation ε_1, which in the case of pure iron remains constant. *c*) If the iron contains dissolved carbon or nitrogen, the sudden elongation ε_1 is followed by a much smaller elongation which tends gradually to a limiting value (elastic after-effect). If at a later moment the tensile stress be suddenly removed, then in the case of pure iron the total strain disappears equally suddenly; with C or N-containing iron the same sudden shortening occurs and is followed by the gradual disappearance of the extra strain.

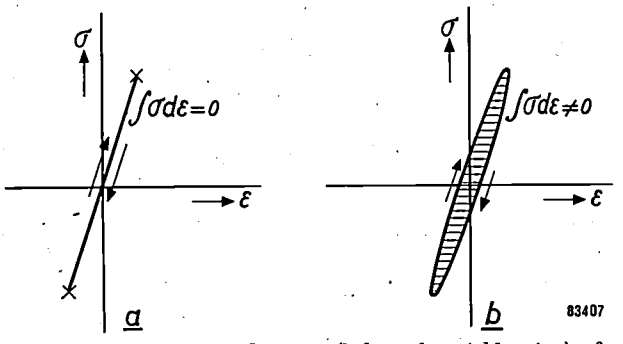


Fig. 8. *a*) For periodic loading (below the yield point) of a body which does not show the after-effect, no dissipation of vibration energy occurs. *b*) If, on the other hand, the material shows the after-effect, then the deformation ε passes its zero value later than the stress σ and energy is dissipated. The area of the shaded envelope gives the energy dissipation per cycle.

The above explanation of the elastic after-effect and damping caused by interstitial atoms was given at Eindhoven by Snoek [6]). Measurements of Snoek-damping has developed within recent years into an important aid to fundamental research on iron and steel. It has made it possible to determine with a high degree of accuracy the diffusion coefficient and solubility of carbon and nitrogen in α-iron at various temperatures. It serves also for the analytical determination of small quantities of C and N in iron and for measuring the rate at which they precipitate from supersaturated solution in α-iron. In connection with the latter it is to be noted that the solubility of carbon and nitrogen in α-iron at the temperatures of the damping measurements (in the neighbourhood of room-temperature) is virtually zero, so that the measurements are performed on supersaturated solutions, obtained by rapid cooling from a higher temperature (say 500 °C). The degree of damping decreases slowly with time at, for example, 20 °C and the rate at which this decrease occurs is a direct measure of the rate of precipitation, i.e. of the rate at which carbide and nitride crystallize out in the mass of the α-iron.

Magnetostrictive deformation

No account has been taken in the foregoing of the magnetostrictive deformation of the iron lattice occurring below the Curie-point. Below the Curie-temperature iron, even in a macroscopically non-magnetic state, is always locally magnetized to saturation. If no external magnetic effect is apparent this merely indicates that the spontaneous magnetization in the various domains, the Weiss domains, is

[6]) J. L. Snoek, Physica **8**, 711-733, 1941 and **9**, 862-863, 1942.

in different directions. The directions always coincide in stress-free iron with one of the six directions of the edges of the unit cell. The iron lattice is slightly elongated in the direction of the spontaneous magnetization (magnetostriction). Hence the assertion made above that the dissolved C and N atoms in stress-free iron are equally distributed among the x, y and z positions, is true only when averaged over a large number of Weiss-domains. It does not hold for the individual domains. If the magnetization vector of a particular Weiss domain is directed, say, in the $\pm x$-direction, there will be a pronounced preference for the x positions.

It is to be expected that the presence of dissolved C and N atoms will have a considerable influence upon the mobility of 90° domain walls (planes between Weiss-domains of mutually perpendicular magnetizations). The displacement of such a 90° domain wall corresponds to a rotation of the magnetization direction through 90° in the region covered and to a contraction of the preferential interstices in which (for the greater part) the foreign atoms are situated (*fig. 9*). This causes a rise in energy. The displacement of a 90° wall in iron containing dissolved C or N, will thus require a greater force than would be necessary for such a displacement in pure iron. It is reasonable to assume that the freedom of movement of 180° domains walls is much less influenced by the presence of dissolved C or N since the magnetostrictive deformation is not altered by the displacement of a 180° wall.

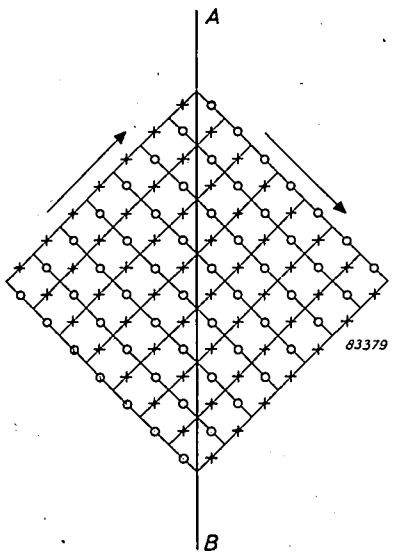


Fig. 9. The line AB represents a 90° domain wall in α-iron. The iron atoms are considered as being located at the corner points of the network. The arrows give the directions of spontaneous magnetization. To the left of AB the locations of the preferred sites for the interstitial atoms are represented by the crosses and to the right of AB by the circles. The difference of location is a consequence of the small magnetostrictive strain in the direction of the arrows. (The finite thickness of a domain wall is neglected here.).

These effects result in a magnetic after-effect which is analogous to the earlier-mentioned elastic after-effect. For small field strengths, figures 7 and 8 may also be applied to the magnetic phenomenon, substituting the field strength H for the stress σ and the induction B for the deformation ε. Suppose, for example, that a piece of iron, containing a few thousandths percent of dissolved C or N, is placed in a weak magnetic field. When the field is removed, then only a part of the induction will disappear virtually instantaneously; the remaining part disappears gradually at a rate which is again determined by the frequency of movement of the dissolved atoms between sites and is therefore dependent upon the temperature. In an alternating field the magnetic after-effect results in a phase-shift between field and induction, involving a dissipation of magnetic energy which must be added to hysteresis and eddy current losses.

Stability of supersaturated solutions

Ordered solutions (hardening of steel)

If we now return to the earlier mentioned hardening of steel in the middle of an argument on magnetic phenomena in iron, it may seem that we are jumping from one subject to another. This, however, is not the case, for an explanation of the *stability* of hardened steel must be sought in the nature of the interstices in α-iron and their division into x, y and z interstices. The local asymmetrical deformation of the iron lattice which accompanies the introduction of a C or N atom, means that the introduction of a second atom into certain neighbouring spaces of the same group (x, y or z) requires a smaller deformation energy. That is to say, the position into which the second atom must enter has already been slightly stretched by the first atom. Thus the occurrence of highly supersaturated, but relatively stable solutions of C or N in α-iron is conceivable, provided that all theses atoms could be introduced into equivalent interstices (say, the x-interstices). This is precisely what does occur automatically in the hardening of steel. The basic material, iron, containing in addition to other elements, a large amount of carbon, about 1 % by weight, is heated to a temperature at which the metal is in the γ state. On account of its high solubility in the γ phase all the carbon goes into homogeneous solution (cf. fig. 2). The metal is then very rapidly cooled, so rapidly that the segregation into α crystals with little dissolved C and iron carbide Fe_3C, as required by the phase dia-

gram, fails to take place. Instead, something else happens in the neighbourhood of 200 °C. The lattice suddenly changes into the α state, the mechanism being such that after the change all the C atoms are located in one kind of interstice, say the x interstices. The lattice constant in the x-direction is in consequence greatly increased, and in the y and z directions decreased (*fig. 10*). This hard tetragonally-deformed variety of α-iron is known as martensite. Jumps of the carbon atoms between unlike interstices cannot occur on account of the

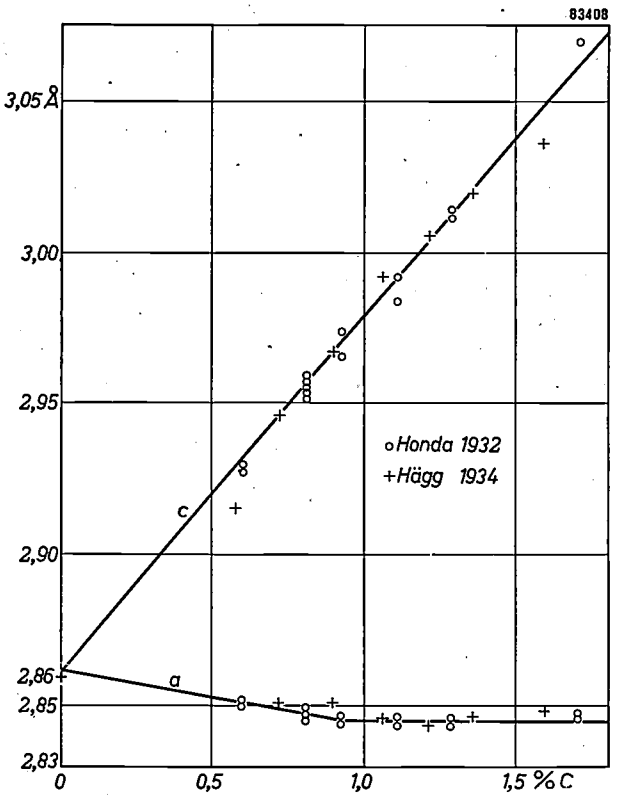


Fig. 10. Lattice constants (a and c) for martensite as a function of the carbon content.

tetragonal deformation and the accompanying co-operative stabilizing action of the carbon atoms. Jumps between like interstices are impossible, because each interstice has unlike neighbours only. Thus a slow precipitation of Fe_3C, such as occurs in disordered supersaturated solutions, in which the carbon atoms make about 1 jump per second at room temperature cannot take place. Old swords still retain their great hardness after many centuries due to this ordered supersaturation with carbon.

A process analogous to that with carbon is likewise possible with nitrogen; the product is then referred to as nitrogen martensite.

Disordered solutions

The precipitation of carbide or nitride which may occur in disordered supersaturated solutions of carbon and nitrogen in α-iron, likewise leads to an increased hardness which remains however far below that attained by the ordered hardening discussed above. One of the reasons for this is that the concentration of carbon or nitrogen is much lower. Precipitation is particularly undesirable if the metal is to be magnetically soft. On account of the decrease in the solubility with decreasing temperature, precipitation can even occur in steel with a very small C and N content. In moving towards a state of equilibrium, there first forms in the supersaturated solution a very fine precipitate, which gradually coarsens, since in this process the total grain boundary surface energy must fall. The precipitate impedes the movement of the domain walls, attaining a maximum effect when the particles of the precipitate are of the same order of size as the thickness of a domain wall (about 10^{-5} cm). The unfavourable influence on the magnetic properties of soft magnetic materials is then appreciably greater than when the same amounts of C or N are present in solution. This gradual deterioration in the magnetic properties is called magnetic ageing. That slowly-cooled mild steel also shows this phenomenon appears to result from the joint presence of manganese and nitrogen [7]). The carbon present precipitates completely during the slow cooling, but the manganese greatly retards the precipitation of the nitrogen The coercive force of mild steel containing a mere 0.005% of N, increases on ageing the steel for a few hundred hours at 100 °C often to twice its former value. The hysteresis losses will therefore also increase. The small amount of nitrogen may be redissolved and the original coercive force restored by heating the metal for a short time at only 250 °C. Ageing will set in anew, however, when the temperature is lowered again.

This undesirable effect can be largely eliminated by adding metals with a strong affinity for nitrogen to the steel. Examples of such metals are aluminium, titanium, zirconium and vanadium. These form stable nitrides AlN, TiN, ZrN and VN. The small magnetic ageing of silicon iron, which is used in large quantities for the construction of transformers, motors and generators seems to be due to the same effect: according to our investigations, the nitrogen present in this metal is fixed in the form of a stable nitride, viz. silicon nitride, by a suitable heat treatment.

Ageing of the type discussed above, based on rapid cooling is called quench ageing to distinguish it from strain ageing discussed earlier. Quench ageing

[7]) J. D. Fast and L. J. Dijkstra, Philips tech. Rev. **13**, 172-179, 1951/52.

likewise manifests itself in the mechanical properties, since movement of the dislocations is also impeded by a precipitate. The maximum hardening effect occurs at a much lower particle size than in the case of the magnetic properties, since the range of action of a dislocation is considerably less than the thickness of a domain wall. This difference is demonstrated in *fig. 11*, which gives a rough indication of the variation of hardness and coercive force of mild steel containing 0.07% of C, which was quenched from 680 °C and subsequently heated for 1 hour at temperatures of 50 °C, 100 °C, 150 °C, 200 °C etc. The data are derived from the measurements of

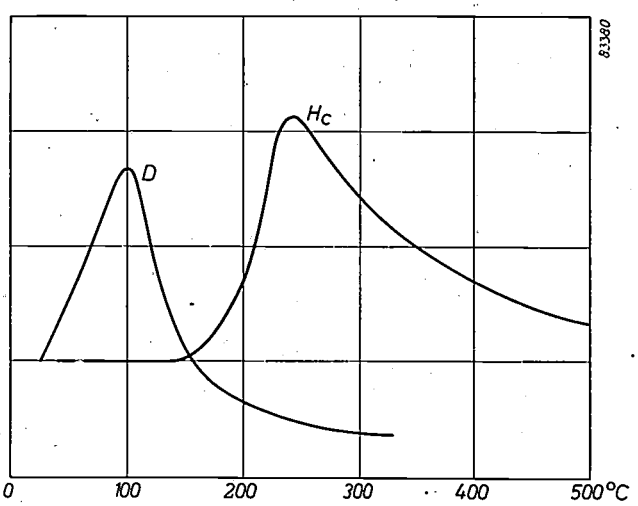


Fig. 11. Rough curves of the mechanical hardness (D) and the coercive force (H_c) for a mild steel containing 0.07% of carbon, after quenching from 680 °C and heating for one hour at the temperatures indicated on the abscissa (from data of Köster [8]).

W. Köster [8]). As may be seen, the maximum mechanical hardness occurs on heating at 100 °C. On heating at higher temperatures, the precipitate is too coarse to have very much effect on the mechanical properties. On the other hand, the maximum influence on the magnetic hardness (coercive force) occurs after heating at 250 °C, when the particles are on the verge of microscopic visibility and their influence on mechanical hardness is negligible.

Transport phenomena in interstitial solutions

In the foregoing examination of the effects of interstitial atoms in metals, attention has been paid particularly to the influence of the strain energy. That the accompanying electronic interaction cannot be ignored even in the simple case of carbon or nitrogen in unalloyed iron, may be seen from experiments on transport phenomena in electric fields. Seith and co-workers [9]) established that carbon in solid iron at 1000 °C moves under the influence of an electric field in the direction of the negative pole, while nitrogen on the other hand moves in the direction of the positive pole. Transport phenomena are shown particularly clearly by oxygen-containing zirconium [10]). In a zirconium wire heated by means of a direct current, the oxygen moves towards the positive pole. At the end adjacent to the negative pole practically pure zirconium is formed. This is indeed the sole method which has been found up to now for purifying solid zirconium of oxygen. Purification by chemical methods is impossible. If the oxygen-containing metal is brought into contact with molten calcium, for example, only a part of the oxygen is removed, since the affinity of zirconium for oxygen increases as its oxygen content decreases. Heating the metal at high temperatures in vacuo to drive off the oxygen is likewise of no avail. Unfortunately the electrolytic method of purification is of no practical significance, since the mobility and the transport number of the oxygen particles are smaller by many orders of magnitude than the mobility and transport number of the conduction electrons. The term particles and not ions is used here intentionally, since the oxygen should not be thought of as clearly distinguishable O^{--} ions. It is merely that the electron gas in the metal is distributed in such a way that the oxygen atoms, on the average have a slight excess of negative electric charge. Experiments so far give the impression that this effective charge on the oxygen atoms in zirconium lies far below one elementary charge.

Also of great interest are the phenomena resulting from the presence of hydrogen in metals. The above discussed strain energy seems to play hardly any part in solutions of hydrogen in metals. For example, the solubility of hydrogen in the body-centred cubic modification of Zr with its small interstices is greater than the solubility in the hexagonal close packed modification which has larger interstices [11]). It is natural to seek an explanation of this in the extremely small size of the hydrogen ion: the radius of this ion (the proton) is but a hundred thousandth of the radius of an atom. Again, one should not think of the hydrogen as present in metals in the form of clearly distinguishable ions; averaged over a sufficiently long time, however, a dissolved hydrogen

[8]) W. Köster, Arch. Eisenhüttenwesen **2**, 503-522, 1928/29.
[9]) W. Seith. Diffusion in Metallen, Springer, Berlin, 1939.
[10]) J. H. de Boer and J. D. Fast, Rec. Trav. chim. Pays-Bas **59**, 161-167, 1940.
[11]) J. H. de Boer and J. D. Fast, Rec. Trav. chim. Pays-Bas **55**, 350-356, 1936.
[12]) A. Coehn and W. Specht, Z. Physik **62**, 1-31, 1930; A. Coehn and H. Jürgens, Z. Physik **71**, 179-204, 1931; A. Coehn and K. Sperling, Z. Physik **83**, 291-312, 1933.

atom will have a deficiency of negative charge. This is in agreement with transport experiments by Coehn and co-workers [12]) who showed that hydrogen in palladium moves towards the negative pole. If the dissolved hydrogen proceeds through the metal lattice in the form of protons, much greater diffusion coefficients are to be expected for this element than for carbon, nitrogen and oxygen. This expectation is borne out by the facts. Thus the diffusion coefficient of hydrogen in α-iron at 20 °C is more than 10^{12} times greater than the diffusion coefficients of carbon and nitrogen [13]). The diffusion coefficient of hydrogen is indeed so large that it might be expected that iron would be permeable to it even at room temperature. The fact that it is possible to store this gas under high pressure in iron cylinders without any perceptible reduction in pressure as a result of diffusion through the walls seems, however, to be in conflict with the above. A similar apparent discrepancy may be seen in the previously mentioned fact that oxygen cannot be driven out of zirconium by heating, even at temperatures at which its diffusion coefficient in this metal apparently has a high value. The explanation of these conflicting observations lies in the fact that the permeability of a metal wall to a gas is often determined not by diffusion, but by a surface reaction. As we have noted on more than one occasion, this explanation is not generally known (publication took place at a rather inopportune moment [14]); it will therefore be briefly discussed below.

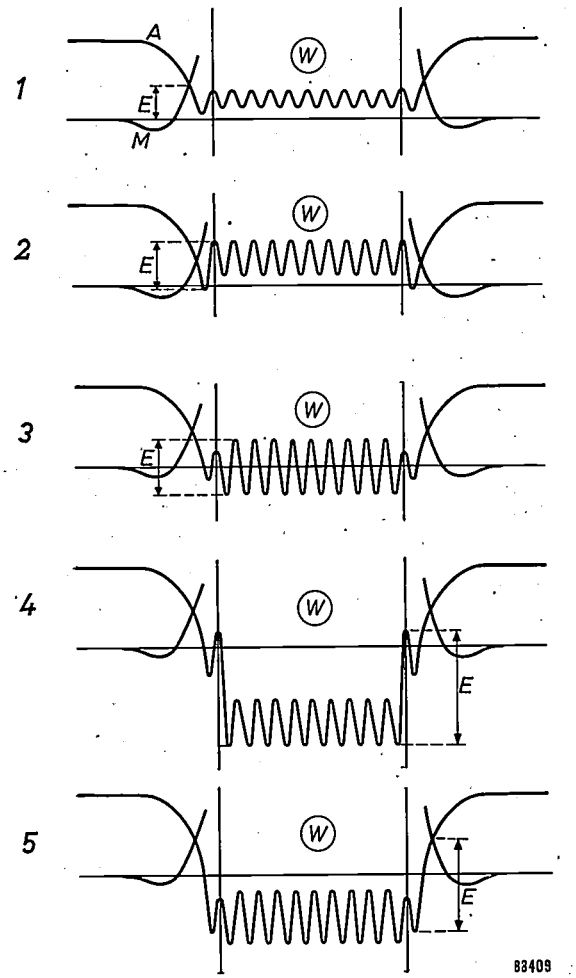


Fig. 12. Schematic representation of the passage of a gas through a metal wall (W). The "potential curves" M and A give the energy of a gas molecule and of its two separate atoms, respectively, as a function of position. The greatest potential jump occurring (E) determines the rate of the whole process. It may occur at any one of five different locations, corresponding to cases *1-5* summarized in the text.

The passage of gases through metals

The passage of a diatomic gas through a metal wall requires five successive processes: 1) the splitting of the molecules into atoms at the surface of entry, 2) the transition of the atoms from the adsorbed state into the interior of the metal, 3) diffusion within the metal, 4) transition of the atoms from the dissolved state into a state of adsorption on the surface of exit, 5) the recombination of the adsorbed atoms into molecules on the surface of exit (*fig. 12*). The slowest of the five processes will determine the permeability. In the passage of hydrogen through an iron wall the first of the five processes, the splitting of H_2 into 2H, is the slowest and is thus the rate determining process. This may be immediately demonstrated by an experiment in which atomic hydrogen is fed to one side of an iron wall. Even at room temperature a ready penetration occurs. If the surface of entry of the iron is made very rough, then the second process, the penetration of the adsorbed hydrogen atoms into the interior, becomes the slowest process. In this case, splitting of the hydrogen molecules into atoms at sharp points and edges takes place spontaneously even at room temperature. The atoms formed are now so much more strongly attached to the surface than is the case for a smooth surface, that the transition to the interior of the metal is rate determining.

In experiments with other gas-metal combinations, instances where the third, fourth of fifth process is rate-determining, have been noted. Thus, for example, diffusion is rate-determining in the passage of hydrogen through a copper wall. Above some hundreds of degrees centigrade, diffusion also determines the rate at which hydrogen passes through an iron wall, in contrast to the situation

[13]) H. D. Fast and M. B. Verrijp, J. Iron and Steel Inst. **176** (I), 24-27, 1954.
[14]) J. D. Fast, Philips tech. Rev. **6**, 365-371, 1941; **7**, 74-82, 1942.

at room temperature. The processes occurring at the surface of exit determine the permeability of a wall of zirconium to oxygen and of a wall of palladium to hydrogen.

To close this necessarily incomplete review on the influence of interstitial atoms in metals let us just return once more to those cases where their presence causes hardness and brittleness in metals. As we have seen, the great hardness of martensite is based on the formation of a new lattice with much smaller interstices, in which the carbon or nitrogen are present under constraint. It is conceivable that hardening could result not only from the formation of new interstices but also from the formation of new atoms. This is actually the case in nuclear reactors using uranium as fuel, and constitutes one of the most urgent metal problems there. The energy production of such a reactor is based on the splitting of nuclei of the isotope U^{235} with the formation of two new atoms. Under the most favourable conditions only one of these atoms can be accomodated in a lattice position, and the other must be taken up interstitially. The ever rising concentration of interstitial atoms eventually causes the uranium rods to break up. One of the tasks of the metallurgist is to prepare uranium in such a state that it can take up as many extra atoms as possible.

THE "DUPLO" CAR HEADLAMP BULB WITH AN ASYMMETRIC DIPPED BEAM

628.971.85:629.113.06

Motorists driving in the dark should dip their headlamps for each oncoming car in order not to dazzle its driver. Dipping is usually done by switching over from the main headlamp filament to an auxiliary filament which supplies the dipped beam. The problem of getting sufficient visibility for safe driving with this limited beam has been tackled from different angles on the continent of Europe and in America.

On the continent the main requirement has been laid on the least possible dazzle. The continental lamps based on this idea give a beam which is symmetrical about a vertical plane through the axis of the lamp and has a sharp horizontal cut-off obtained by means of a small metal cap mounted under the auxiliary filament. In America the emphasis has been laid on the illumination of possible obstacles on the road, the requirement of minimum dazzle taking second place. The asymmetric beam of the so-called "sealed-beam" lamp is based on these premises. This gives more light on the near-side of the road than on the off-side (the off-side is lit by the lamps of the oncoming car). This arrangement gives little consideration for dazzle on bends or the dazzle of cyclists or pedestrians on the near side of the road. This difference between American and continental practice has been described earlier in this Review [1]).

For the motorist, the higher light intensity on the nearside kerb [2]) is an advantage of the American dipped beam. The reduced dazzle of the oncoming motorist and other road users and the sharp light-dark cut-off which helps in the aiming of the beam (which, of course, is very important) are advantages of the continental system.

It is possible to combine the advantages of both systems to a considerable extent by cutting off a part of the metal cap under the auxiliary filament of the continental type of bulb (see *fig. 1*). The dipped deam then obtained retains the sharp light-

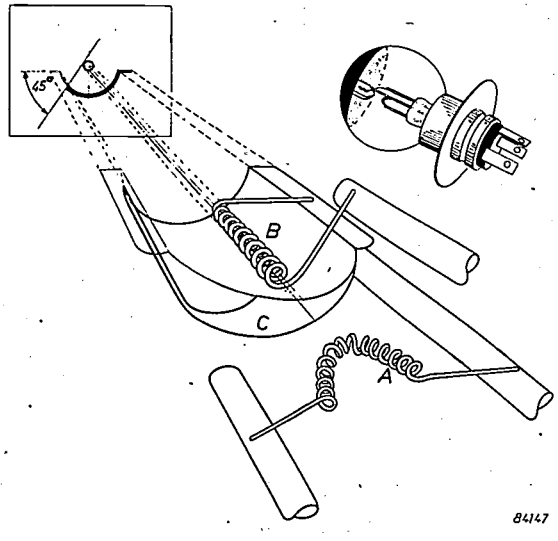


Fig. 1. Main filament *A*, auxiliary filament *B* and metal cap *C*, showing their positioning in a headlamp bulb of the new type. (See right top). The part of the metal cap drawn in thin lines is cut away in order to give the asymmetric beam. This is further clarified in the projection of the metal cap shown in the top left of the figure.

[1]) J. B. de Boer and D. Vermeulen, Motorcar headlights, Philips techn. Rev. 12, 305-317, 1950/51.

[2]) Note that figs. 1 and 2 relate to lamps developed for those countries where the traffic drives on the right-hand side of the road. A lamp for left-hand traffic can of course be designed on the same principles.

at room temperature. The processes occurring at the surface of exit determine the permeability of a wall of zirconium to oxygen and of a wall of palladium to hydrogen.

To close this necessarily incomplete review on the influence of interstitial atoms in metals let us just return once more to those cases where their presence causes hardness and brittleness in metals. As we have seen, the great hardness of martensite is based on the formation of a new lattice with much smaller interstices, in which the carbon or nitrogen are present under constraint. It is conceivable that hardening could result not only from the formation of new interstices but also from the formation of new atoms. This is actually the case in nuclear reactors using uranium as fuel, and constitutes one of the most urgent metal problems there. The energy production of such a reactor is based on the splitting of nuclei of the isotope U^{235} with the formation of two new atoms. Under the most favourable conditions only one of these atoms can be accomodated in a lattice position, and the other must be taken up interstitially. The ever rising concentration of interstitial atoms eventually causes the uranium rods to break up. One of the tasks of the metallurgist is to prepare uranium in such a state that it can take up as many extra atoms as possible.

THE "DUPLO" CAR HEADLAMP BULB WITH AN ASYMMETRIC DIPPED BEAM

628.971.85:629.113.06

Motorists driving in the dark should dip their headlamps for each oncoming car in order not to dazzle its driver. Dipping is usually done by switching over from the main headlamp filament to an auxiliary filament which supplies the dipped beam. The problem of getting sufficient visibility for safe driving with this limited beam has been tackled from different angles on the continent of Europe and in America.

On the continent the main requirement has been laid on the least possible dazzle. The continental lamps based on this idea give a beam which is symmetrical about a vertical plane through the axis of the lamp and has a sharp horizontal cut-off obtained by means of a small metal cap mounted under the auxiliary filament. In America the emphasis has been laid on the illumination of possible obstacles on the road, the requirement of minimum dazzle taking second place. The asymmetric beam of the so-called "sealed-beam" lamp is based on these premises. This gives more light on the near-side of the road than on the off-side (the off-side is lit by the lamps of the oncoming car). This arrangement gives little consideration for dazzle on bends or the dazzle of cyclists or pedestrians on the near side of the road. This difference between American and continental practice has been described earlier in this Review [1]).

For the motorist, the higher light intensity on the nearside kerb [2]) is an advantage of the American dipped beam. The reduced dazzle of the oncoming motorist and other road users and the sharp light-dark cut-off which helps in the aiming of the beam (which, of course, is very important) are advantages of the continental system.

It is possible to combine the advantages of both systems to a considerable extent by cutting off a part of the metal cap under the auxiliary filament of the continental type of bulb (see *fig. 1*). The dipped deam then obtained retains the sharp light-

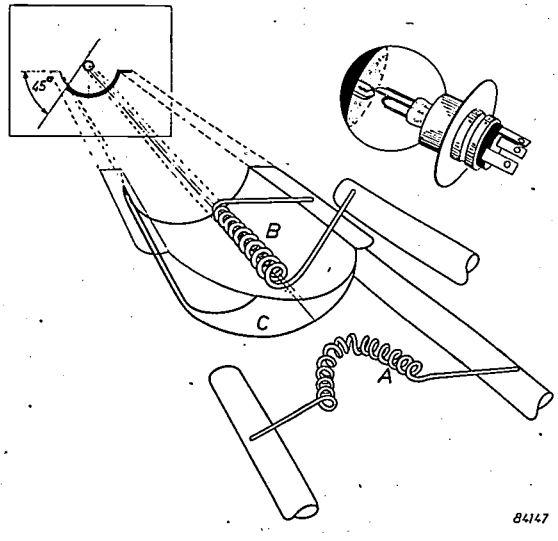


Fig. 1. Main filament *A*, auxiliary filament *B* and metal cap *C*, showing their positioning in a headlamp bulb of the new type. (See right top). The part of the metal cap drawn in thin lines is cut away in order to give the asymmetric beam. This is further clarified in the projection of the metal cap shown in the top left of the figure.

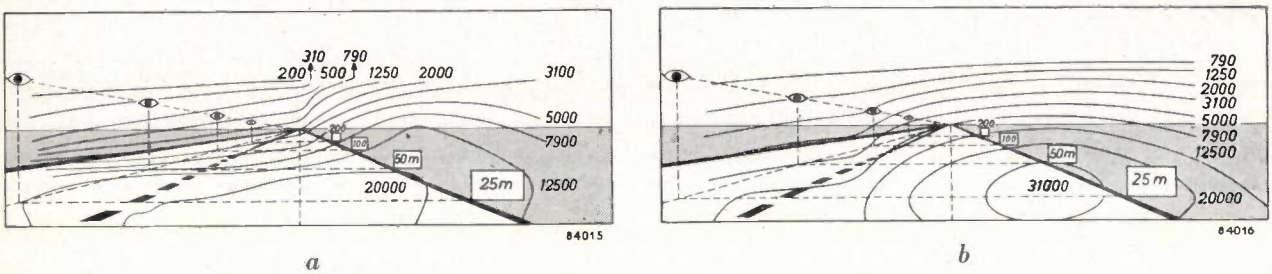
[1]) J. B. de Boer and D. Vermeulen, Motorcar headlights, Philips techn. Rev. 12, 305-317, 1950/51.

[2]) Note that figs. 1 and 2 relate to lamps developed for those countries where the traffic drives on the right-hand side of the road. A lamp for left-hand traffic can of course be designed on the same principles.

dark cut-off of the original lamp on the off-side but on the near-side more light is radiated so that the beam is here more like that of the American lamps.

Road tests according to internationally approved procedures have shown that the new asymmetric beam lamp gives at least the same visibility as the American sealed beam lamps [1]) (since publication of the article [1]), these have been somewhat modified), and hardly more dazzle than that of the existing continental lamps. Because of the latter there can be no objections from a technical point of view to the use of the new asymmetric lamp in countries which approve the existing continental dipped beam.

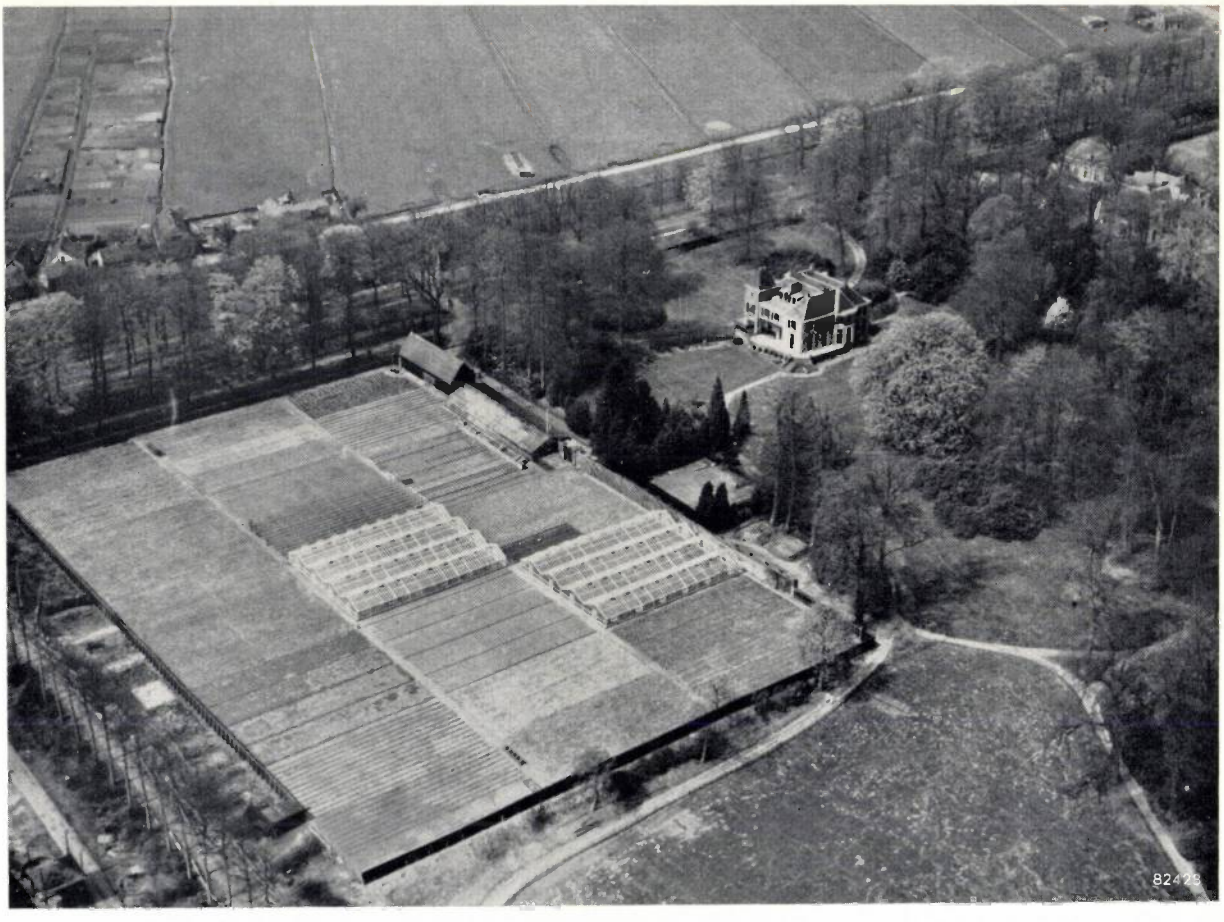
The light distribution obtained with the new lamps and that with the sealed-beam lamps are compared in *fig. 2a* and *b*. In the direction of points on the right side of the road [2]) in the region 50-200 yds in front of the car, the light intensity from the new asymmetric bulb is about 2/3 that of the new type of sealed beam lamp. In the direction of the eyes of the oncoming motorist, the light intensity of the assymmetric bulb is only 1/3 that given by the American headlamp.



a *b*

Fig. 2. Light distribution [2]) *a*) of the new asymmetric beam European headlamp bulb, *b*) of the latest type of American sealed beam lamp.
The light distributions are here given by iso-candela lines. The six-yard-wide road is viewed from a point midway between the headlamps of a car in the middle of the right half of the road and projected on an imaginary screen in front of the car. In the perspective picture so obtained, each iso-candela line joins the points corresponding to directions in which the light intensity due to both headlamps together has the value marked by the line. Note that the distances are actually given in metres.

J. de BOER.



"BOEKESTEYN"
THE AGROBIOLOGICAL LABORATORY OF N.V. PHILIPS-ROXANE

by R. van der VEEN. 632.931.33:615.777/.779

A number of articles are due to appear in this Review on the work of the Boekesteyn agrobiological laboratory. As the fields of disease and pest control in agriculture and horticulture will be unfamiliar to most of our readers, we have thought it desirable to begin this series with the introductory survey given below.

Protection of cultivated plants

When N.V. Philips-Roxane entered the field of disease and pest control in agriculture and horticulture, it became necessary to have available a laboratory for experimental work in these fields. Such a laboratory has now been established at the old country estate of "Boekesteyn" at 's-Graveland, near Hilversum. The Boekesteyn estate consists of a manor house (*fig. 1*), a fine park, a farm, a kitchen garden with several hothouses, and an orchard of young trees. The park, which is open to the public, is quite separate from the laboratory. The farm is also separate and is let. The house, however, is completely equipped as a laboratory, and the kitchen garden and orchard are likewise turned over to experimental purposes.

It is the aim of the research at Boekesteyn to develop efficient and economic means to combat diseases and pests in all types of cultivated plants. Although in an agricultural country such as the Netherlands intensive measures of disease and pest control have been in progress for a considerable time, at least 10% of the possible harvest is lost as a result of commonly occurring diseases. In most other countries this percentage is considerably higher.

It is sometimes asked why there is so much spraying and dusting nowadays. "There was none of that in the old days, but the fruit still grew on the

trees and there was still corn in the fields", it is argued. That may be so, but to-day much more exacting demands are made as regards the yields par acre and there are many more acres under cultivation. This increased cultivation usually involves a wider spread of diseases and pests.

The plant is a living organism; therefore, by careful selection, a certain resistance can be built up to specific diseases. It would be ideal if exclusively resistant varieties could be raised. Plant improvement laboratories and institutes are indeed busy with the breeding of such races and much has already been achieved in this field. The enemies of the plant, for the most part fungi and insects, are likewise living organisms, however, and show great adaptability. As a consequence it is often found that a few years after a plant, resistant to a certain fungus, has been cultivated on a large scale, the fungus has also produced a new variety which is adapted to the improved plant. Therefore, for the time being at least, the employment of chemical protective measures remains necessary.

It goes without saying, however, that injudicious use of these chemical agents may well lead to results just the reverse of those expected. A good example of this was the large-scale spraying of citrus plantations with DDT which was carried out at a time when it was still considered that DDT was effective against all insects. There was, however, a scale which was unaffected by DDT. Its enemies were exterminated by the spraying, so it was able to multiply unhindered. The damage wrought by this single species was greater than that which would have been caused by all the other pests together had there been no spraying!

Other examples might be cited which show how the balance of nature may be disturbed by spraying. But here it should be remembered that agriculture itself is a disturbance of the balance of nature. It begins whenever the original vegetation is rooted out and a new form of vegetation is artificially cultivated in its stead. Only by hard work can the farmer maintain this disturbance.

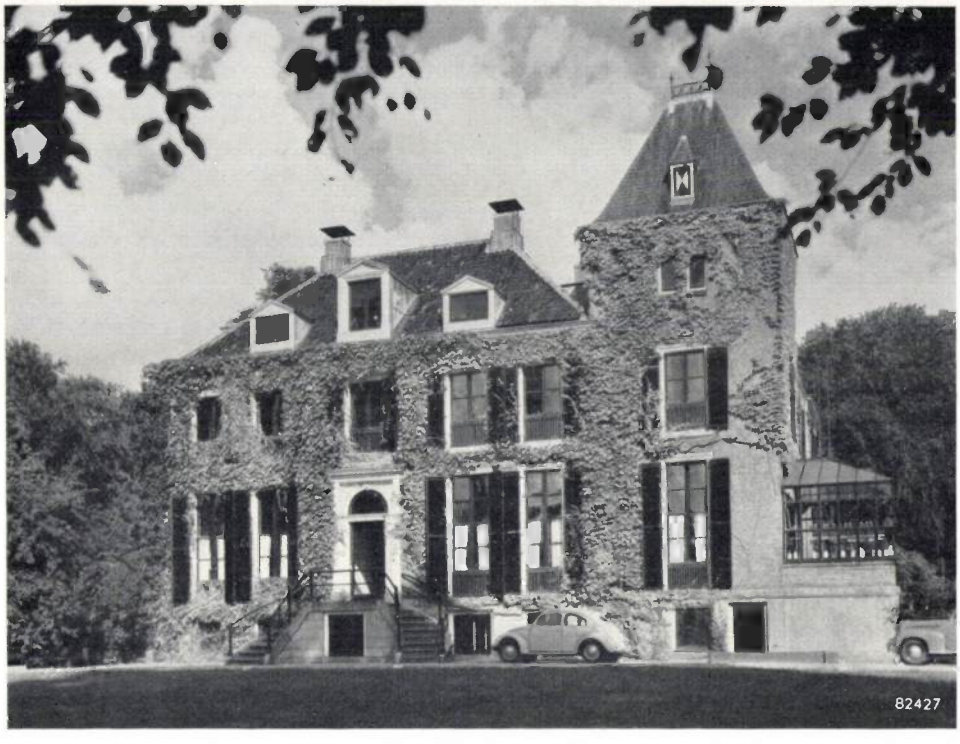


Fig. 1. Boekesteyn House, now fitted out as an agrobiological laboratory for Philips-Roxane.

Better manuring brings new weed problems, more productive varieties frequently have new susceptibilities, extension of the area under cultivation brings with it more virulent diseases and pests. Chemical measures against diseases have their drawbacks, but they are nevertheless essential in a densely populated world.

One of these drawbacks has already been named: not only the parasites, but also the parasites' enemies may be exterminated by chemical agents and the chance of a rapid spread of a new plague is thereby increased. Another danger is that substances toxic to insects are often harmful to man also. The

further development of agricultural chemicals must therefore be directed towards the discovery of substances which are not toxic to man and are selectively toxic to the pests, leaving all other insects undisturbed.

This means that the substance must have a highly selective action. The ideal solution would be to have available a series of substances, each active against one particular type of harmful insect without being poisonous to other creatures. Something of a similar nature would be desirable for fungous diseases; here, however, a substance will suffice which is harmless towards green plants and animals, but active against all fungi.

Before going further into this ideal objective, it is necessary to give a brief review of the development of chemical disease and pest control over the course of the years.

Chemical agents for combating plant diseases and pests

Plant diseases are caused by fungi, by bacteria or by viruses. Moreover, plants are exposed to insect pests and the encroachment of weeds.

A whole arsenal of chemical agents is available for combating all these evils except the virus diseases. For opposing the latter we are still wholly dependent upon the selection of resistant varieties and on the rigorous weeding-out of infected plants and other measures for reducing the danger of contagion.

We will here limit ourselves to the methods by which fungous and bacterial diseases and insect and weed pests may be combated.

Fungous diseases

In general, fungi propagate themselves by means of spores. Therefore, the spread of the disease can be controlled by an agent which will kill the spores or prevent them from germinating.

Various substances (fungicides) are known which operate in this way. The oldest agent and one which is still very widely employed is Bordeaux mixture, a mixture of copper sulphate and lime. To-day there are various other copper compounds on the market which are equally effective, for example, copper oxychloride and the so-called "colloidal copper". Mercury compounds have also made progress, especially as seed desinfectants; they are less suitable for field use, being frequently injurious to plants. There are, however, organic mercury compounds which can be used for spraying fruit trees, although a certain degree of caution must be exercised.

In the last 15 years important applications have been found for organic compounds containing no toxic metals like copper and mercury. The dithiocarbaminates and the quinones in particular have furnished a number of excellent fungicides.

Dithiocarbaminates. Unlike bacteria, most fungi appear to be sensitive to certain dithiocarbaminates, for example tetramethylthiuram disulphide [1] [1]) — known as TMTD — and zineb [2].

Even when diluted to $1 : 10^6$ in water, these compounds completely prevent the germination of spores. A plant which has been sprayed or dusted with them is thus very well protected against most fungous diseases. These compounds are gradually replacing the copper and mercury compounds.

Quinones. Among the quinones, compounds are found [3], [4] whose fungicidal action is of about the same intensity as that of the compounds of the previous group. Their price, however, is somewhat higher and as a consequence they are gaining less ground than the dithiocarbaminates.

Latterly, yet another organic fungicide, known as captan [5], has come to the fore. In general, it is perhaps of a somewhat lesser fungicidal value than the dithiocarbaminates, but it has the favourable property of appreciably improving the colour of both leaves and fruit when used on fruit treess. For the fruit this means a higher market value and for the leaves a longer period of assimilation extending into the autumn, and hence a more vigorous growth of the tree. Philips-Roxane have developed a very useful fungicide which likewise enjoys this property, possibly to an even higher degree.

Frequently the resistance of a plant to fungal diseases rests upon the fact that the plant contains some substance toxic to the fungus. If growers could introduce such substances into their plants artificially, it would be possible to achieve resistance in this way. The substance must be such that the plant may assimilate it and distribute it throughout its whole system without suffering any resultant damage. Such substances are known as "systemic" fungicides.

A few substances are known which have a clearly systemic action, but not one of these however, has been developed far enough to find practical application. Nevertheless, an intensive search for systemic fungicides is in progress at Boekesteyn. A great advantage of such a fungicide would be that, on being sprayed solely on the leaves, it would be distributed throughout the whole plant and would thus also protect the roots against ground fungi.

---

[1]) The numbers in square brackets placed after the names of the compounds refer to the corresponding structural formulae given in the appendix.

Bacterical diseases

Bacterial diseases are usually more difficult to combat than those caused by fungi, since the bacteria spread throughout the soil and infect the plants via the roots. Disinfection of the soil is the best treatment available, but is expensive.

In recent years antibiotics such as streptomycine have been employed as "systemic" agents against bacterial diseases, sometimes with success. They render the plants less susceptible to bacteria, but the price and often the toxicity to the plant form obstacles to their general application.

Insect pests

The increasingly intensive cultivation of the land in the present century has led to a marked increase in the number of insect pests. Insects, it appears, quickly adapt themselves, and on being offered many acres of food, will propagate themselves with extraordinary rapidity.

Like fungi, insects were formerly combated with inorganic poisons, such as lead arsenate and copper arsenite. In the last ten or twenty years, changes have occurred also in this field.

It was first found that some plants contain substances which are extremely poisonous to insects: nicotine from tobacco, rotenone from Derris, pyrethrine from Pyrethrum. Tobacco, Derris and Pyrethrum were specially cultivated for the production of these substances.

Since 1940 these substances are no longer used on such a wide scale, owing to the discovery of extremely efficient insecticides which can be produced synthetically at low cost.

Dichloro-diphenyl trichloroethane [6] — widely known as DDT — was the first, and rapidly gained general acceptance when it was observed that many insects were susceptible to it and that warm-blooded animals could tolerate large doses of it. DDT has given good service in combating malaria and typhus.

Hexachloro-cyclohexane [7] (known as HCH or BHC) followed shortly afterwards and appeared to be more toxic to many insects which are insensitive to DDT. One disadvantage of HCH, however, is its unpleasant smell. Further research showed that HCH consists of a mixture of isomers and that it is the gamma-isomer almost exclusively which is responsible for the insecticidal properties; this isomer is odourless. The gamma-isomer is produced by Philips-Roxane under the name of lindane (called after Van der Linden, who was the first to isolate the gamma-isomer in the pure state).

More recently a new group of insecticides is gaining ground (aldrin [8], dieldrin [9]).

Substances extremely toxic to insects were found during war-time research on poison gases. Several of them (for the greater part phosphorus compounds) have rapidly gained ground in agriculture. The best known of these is parathion [10]. A disadvantage of such materials is that they are toxic not only to insects but also to all animals and thus also to man. However, insecticides have more recently been discovered in this group which are less toxic to warm-blooded animals, for example, malathion [11].

Just as substances are being sought which render the plant itself poisonous to fungi or bacteria, so also a search is in progress for "systemic" insecticides. Several compounds from the above-named group of phosphorus compounds show this effect, that is to say they are assimilated by the plant without causing it any damage and thereby render the plant toxic to insects which feed upon it; an example of this is schradan [12]. Since these substances are broken down only slowly in the plant, the treatment of plants intended for human or animal consumption must take place a long time before they are due to be eaten, since the systemic insecticides are also very poisonous to man.

Spider mites (Acaridae) are also very harmful to plants. Although these creatures are not insects, but members of the spider family, the agents used to combat them (acaricides) are usually considered as insecticides. Normal insecticides as a rule have little effect on mites; DDT appears indeed to exert a stimulating influence upon them. Parathion, which is fatal to all animals, does, however, exterminate mites; as do also mineral oils. Philips-Roxane has developed a further acaricide which will come into production in 1955. Its toxicity to insects is so low that bees, for example, may be fed upon it without suffering any harm.

Weed pests

It is curious that substances which must be classified as belonging to the group of "growth-hormones" (further details regarding which will be published in a later article) may also be employed to effect the antithesis of growth, namely to eradicate certain plants. Thus the widely-known compound, 2,4-D [13], will kill most dicotyledons in certain concentrations but will leave monocotyledons unharmed. With this substance, therefore, meadows and cornfields may be rendered free from weeds.

Two other weed-killers (herbicides) based upon growth-hormones, namely 2,4,5-T [14] and 2-methyl-4-chloro-phenoxy-acetic acid [15] — known

as MCPA — have an analogous action. However, their effects differ: brambles and nettles may be combated with 2,4,5-T, while 2,4-D and MCPA have little effect upon these weeds.

Weed-killers are also sought which will selectively kill grasses while leaving dicotyledons unharmed. Up till now, however, this search can hardly be said to have been successful. There is a preparation which is known as a "weed killer for carrots", which will kill all plants, save only the umbellifers and the majority of conifers, so that it can be used on carrots and celery and in coniferous nurseries.

Finally we must mention a group of compounds which exterminate all plants and are used, for example, in keeping paths free from overgrowth. For this purpose, certain types of chlorates or arsenic compounds are employed and sometimes also pentachlorophenol [16], known as PCP. Various new compounds for this type of application, e.g. trichloroacetic acid (TCA) [17] and chlorophenyl dimethyl urea (CMU) [18] are still in the experimental stage.

Agrobiological research at Philips-Roxane

Agrobiological research work at Philips-Roxane is pursued along the following broad lines:

1) By investigating the natural chemical defensive agents in plants, an attempt is being made to gain an insight into the mode of operation and the character of those substances which nature, as it were, has provided to make plants resistant to fungi and insects.
2) By a search for "systemic" chemicals, such as are harmlessly assimilated by plants, attempts are being made to increase the resistance of plants to certain diseases or to render them toxic to certain insects.
3) By testing numerous substances for their effect on fungi, insects and the higher plants, it is hoped to find compounds with as little toxicity as possible to man and the domestic animals; such substances must be as selective as possible so as to kill a very definite class of insects but be non-injurious to other insects.

Research at Boekesteyn is conducted by five departments which are in close contact with the Philips-Roxane chemical laboratory at Weesp. In the latter laboratory new chemical compounds are synthesized and are later tested at Boekesteyn.

The departments at Boekesteyn are as follows:

a) The entomological department, where compounds are tested for their insecticidal properties.
b) The mycological department, where the same process is carried out with respect to fungicidal properties.
c) The systemic department, where systemic fungicides and the naturally occurring defensive agents of plants are studied.
d) The herbicide department, which investigates the concentrations of compounds injurious to green plants.
e) The field service, which conducts field tests on substances which have given promising results in the laboratory.

A substance received for investigation usually undergoes examination roughly as follows. It is first ascertained in the entomological and mycological departments and sometimes also in the systemic department, whether small concentrations of the substance are lethal to insects or fungi. Various species of insects and fungi are employed for this purpose; this means that an extensive collection of these species has to be maintained in culture (*fig. 2*). If the compound appears to be active when very highly diluted, then it is sprayed onto plants for the purpose of ascertaining whether the latter suffer any damage from the compound — this is all too often the case. If, however, the plant is much less susceptible than the fungi or insects (the susceptibility

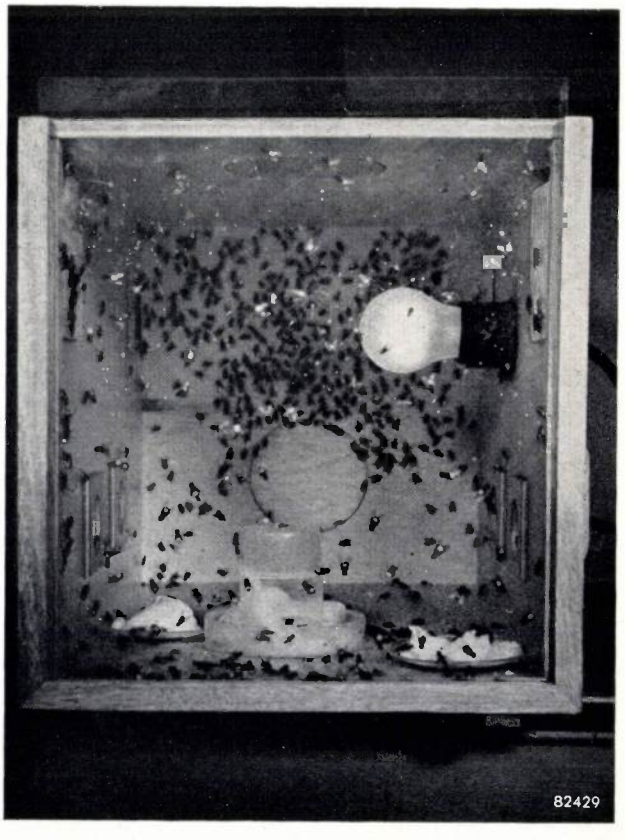


Fig. 2. Breeding of flies, which are used for initial tests on new disease and pest control substances.

must be at least a hundred times less), then there is here a case for further research. This consists in the first place in accurately determining the concentration necessary to give 50% mortality first for fungi or insects and then for the plants. Here the investigation extends over a greater range of fungous and insect species than in the previous test.

If after all this, the compound still appears to show promise, an investigation follows into its toxicity to warm-blooded animals and a few experiments are conducted with it in the field.

If the possibilities still appear favourable, the feasibility of industrial-scale production is examined. This involves the following: choosing the simplest method of production, calculating the cost price, carrying out "formulation" tests (i.e. tests to ascertain the form in which the compound can be best employed: as spraying powder, as dusting powder, as emulsion, etc). and an exhaustive investigation by means of field tests to ascertain those diseases or pests against which the compound is effective; finally permission to market the compound for specific uses must be sought from the competent authorities of the various countries (such applications are directed in the Netherlands to the Director General of Agriculture via the Plant Disease Service).

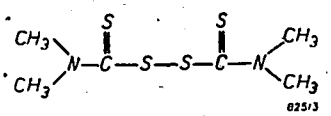
As may well be imagined, only a very small number of the hundreds of compounds which are sent each year from Weesp to Boekesteyn satisfactorily pass through all stages. The overwhelming majority get no further than the first test for killing properties, and of the few which pass through this stage, many fall by the wayside during the later investigations. The very few exceptions which do come to fruition must atone for the many disappointments.

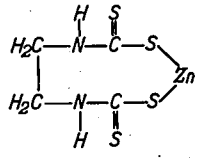
APPENDIX: STRUCTURAL FORMULAE

Below are given the structural formulae of the compounds mentioned in the text.

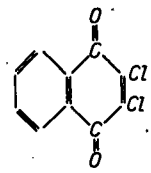
Fungicides

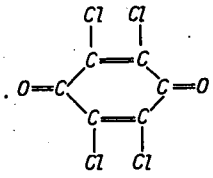
Dithiocarbaminates

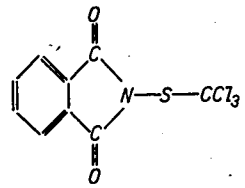
[1] TMTD . . . . . 

[2] Zineb . . . . . 

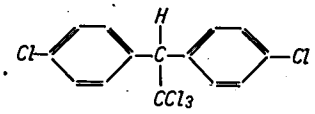
Quinones

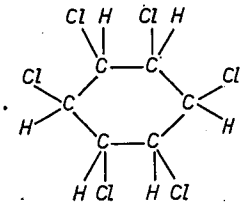
[3] Dichloronaphthaquinone . . . . . 

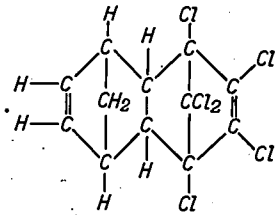
[4] Tetrachlorobenzoquinone . . . . 

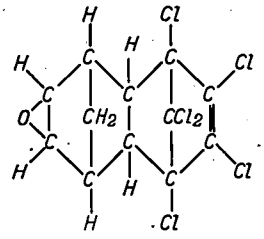
[5] Captan . . . . . 

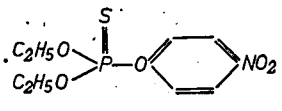
Insecticides

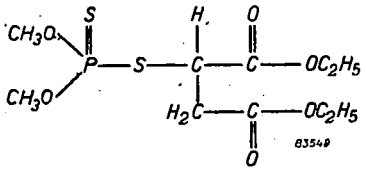
[6] DDT . . . . . 

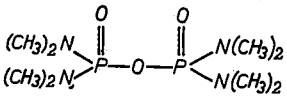
[7] HCH. BHC, Lindane . . . . . 

[8] Aldrin . . . . . 

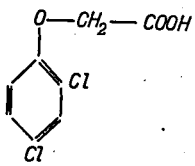
[9] Dieldrin . . . . . 

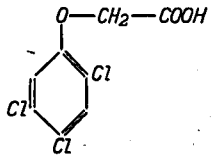
[10] Parathion . . . . . 

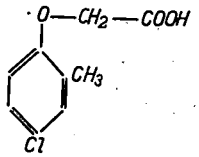
[11] Malathion . . . . 

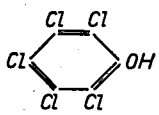
[12] Schradan (systemic insecticide) . . . . . 

Herbicides

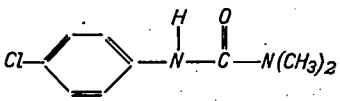
[13] 2,4-D . . . . . 

[14] 2,4,5-T . . . . . 

[15] MCPA . . . . . 

[16] PCP . . . . . 

[17] TCA . . . . . $CCl_3—COOH$

[18] CMU . . . . . 

Summary. This article forms the introduction to a series of articles on the work of the "Boekesteyn" agrobiological laboratory of N.V. Philips-Roxane. This work consists in investigating compounds produced in the company's chemical laboratory at Weesp (Holland).

It is pointed out that the intensification of both agriculture and horticulture in the present century has made the control of diseases and pests more necessary than ever. Chemical agents have a large part to play in this control. Means of control against fungous and bacterial diseases (no chemical agents are yet known against the virus diseases) and against insect and weed pests are reviewed. The organization and departments of Boekesteyn are then outlined (entomological, mycological and "systemic" departments, together with a field service for practical evaluation). A description is given of the procedure followed in investigating the various substances. An appendix is given containing the structural formulae of many of the substances mentioned.

FAST COUNTER CIRCUITS WITH DECADE SCALER TUBES

by E. J. van BARNEVELD.

621.385.832:621.318.57

*The E1T scaler tube *) is a decade scaler within very small compass — a tube no larger than an ordinary radio valve. It has the further advantages of direct reading (the number counted appears as an illuminated spot opposite the digit on the tube) and a high counting rate when used with a suitable input circuit. Suitable circuits, together with the counting rates actually attained by them during laboratory tests, are discussed in the present article. In particular, the counting of random pulses is considered.*

The need for fast scalers, in particular as a means of measuring radioactive radiation and other phenomena associated with nuclear physics, has increased considerably in recent years. Radiation intensities can be measured by means of a Geiger-Müller counter tube [1]), in which every incident particle (e.g. β particle) causes a discharge and so produces an electric pulse, or with a scintillation counter, containing a fluorescent substance which scintillates under radioactive radiation, the scintillations being converted into electric pulses by a photo-multiplier [2]).

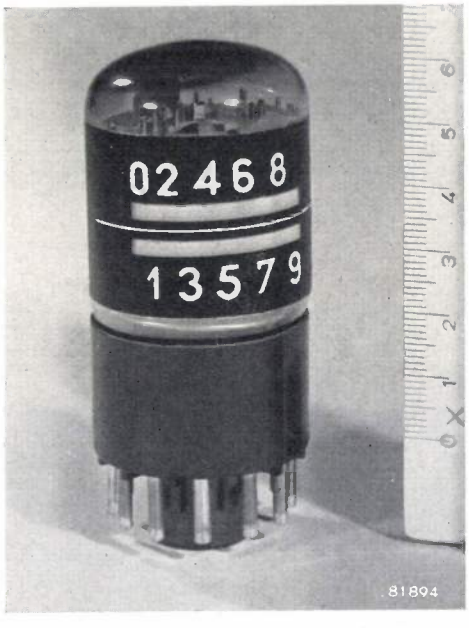


Fig. 1. The E1T decade counter tube. A blue-green fluorescent spot indicates the digit corresponding to the position of the beam in the scaler tube, that is, to the number of pulses counted.

Provided that the repetition rate of these pulses is not too high (i.e. does not exceed 100 per second), they can be counted with a simple mechanical counter. Where the pulses follow one another more rapidly, however, it is necessary to employ electronic methods. These methods enable very fast scalers to be constructed. However, if such a scaler be equipped entirely with standard electronic tubes, it will be rather expensive and will require quite a large number of tubes, particularly if it is to be used for decade counting.

Here, the decade scaler tube *) E1T (*fig. 1*), already described in this Review [3]) and in other publications [4]), offers a better solution. Although the functioning of this tube is fully described in article [3]), referred to in the following as I, a brief description of it will now be given.

Operation of decade scaler tube E1T

The E1T scaler tube, shown in cross-section in *fig. 2a* and diagrammatically in fig. 2*b*, is in essence a cathode-ray tube, whose electron gun produces a ribbon-shaped beam (the advantages of which are described in articles [3]) and [4])). This beam can be deflected to and fro by varying the potential of the right-hand deflector plate (D'). Beyond the deflector plates, the beam impinges upon the so-called slotted electrode (g_4), which has ten vertical slots. As the beam traverses this electrode, a certain number of the beam-electrons (depending on the position of the beam) pass through a slot and strike the anode of the tube, which is behind the slotted electrode. The slots are so designed as to produce a variation of the anode current (i_{a_2}) depending on the potential of deflector plate D' roughly in accordance with curve *I* in *fig. 3*. Since the anode current is virtually independent of the anode potential, curve *I* is likewise valid when the anode is connected direct to D'. Now, if a_2 and D' are connected to the 300 V line via a common resistor, the potential of a_2 and D' will be:

$$v_{D',a_2} = V_B - i_{a_2} R_{a_2} \quad \ldots\ldots\ldots \quad (1)$$

*) The E1T, previously referred to as a "counter tube", is now termed a "scaler tube" to preclude confusion with counter tubes such as Geiger-Müller tubes etc.

[1]) See e.g. N. Warmoltz, Philips tech. Rev. **13**, 282-292, 1951/52.

[2]) See e.g. H. Dormont and E. Morilleau, Le Vide **8**, 1344-1352, 1953 (No. 45); R. Champeix, H. Dormont and E. Morilleau, Philips tech. Rev. **16**, 250-257, 1954/55

[3]) A. J. W. M. van Overbeek, J. L. H. Jonker and K. Rodenhuis, A decade counter tube for high counting rates, Philips tech. Rev. **14**, 313-326, 1952/53.

[4]) J. L. H. Jonker ,Valves with a ribbon-shaped electron beam; contact valve, switch valve, selector valve, counting valve, Philips Res. Rep. **5**, 6-22, 1950. J. L. H. Jonker, A. J. W. M. Overbeek and P. H. de Beurs, A decade counter valve for high counting rates, Philips Res. Rep. **7**, 81-111, 1952,

This relationship is represented by line *II* in fig. 3, viz. the load line. The points of intersection of lines *I* and *II* represent equilibrium states. It will be seen, however, that only the points in the diagram indicated by dots correspond to *stable* positions of the beam.

In each of the ten stable positions, a certain number of the beam electrons pass through an aperture in the anode and strike a layer of fluorescent material (*l*, fig. 2*a*) on the wall of the bulb, thus producing a luminous spot to indicate which of the ten stable positions, marked 0 ... 9, is occupied by the beam.

The actual counting process may be described in the following manner. With the beam in a stable position, each count pulse gives rise to a positive voltage step of 14 V which is applied to the left-hand deflector plate (*D*). Owing to the presence of stray capacitances, the potential of the right-hand deflector plate cannot change so very quickly, and may therefore be considered constant, at least for the time being. Hence the beam shifts abruptly to the left, that is, to the next stable position. If the potential of the left-hand deflector plate be returned gradually to its original level, the stray capacitances will be charged in the meantime, thus enabling the beam to stabilize itself in the new position.

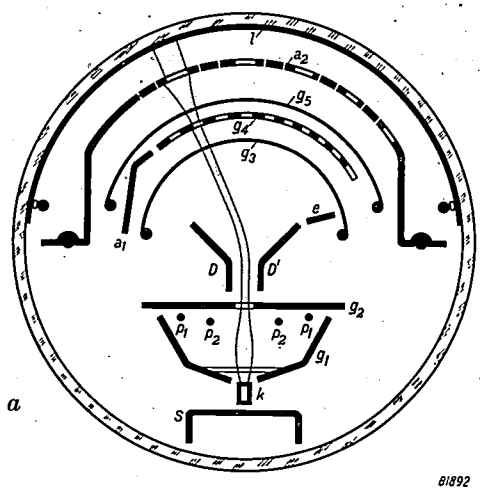


Fig. 2. *a*) Cross-section and *b*) Schematic representation of the decade scaler tube. The cathode (*k*) (with heater *f*), the control grid (g_1), the four internally connected focusing electrodes (p_1 and p_2) and the accelerating electrode (g_2) constitute the electron gun, which produces a ribbon-shaped beam (the width of the ribbon is normal to the plane of the drawing). *D*, *D'* deflector plates; g_3, g_5 suppressor grids; a_1 reset anode; g_4 electrode with ten slots; a_2 anode; *l* fluorescent layer; *s* screen (connected internally to *k*) preventing primary electrons from striking the tube envelope. The auxiliary anode *e* (connected internally to g_2) captures secondary electrons.

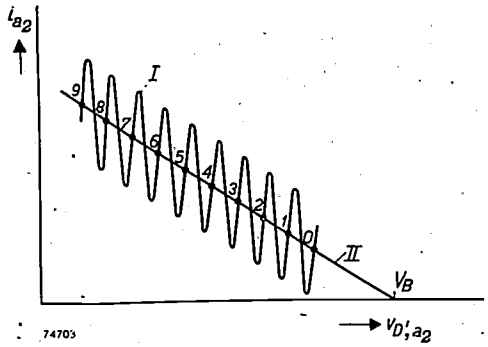


Fig. 3. Undulating line *I*: characteristic of the decade counter tube (anode current I_{a_2} versus potential V_{D',a_2} of the right-hand deflector plate and the anode).
Line *II* is the diagrammatical equivalent of equation (1). Only those points of intersection of *I* and *II* that are numbered from *0* to *9* represent stable beam positions; all others corresponding to unstable positions.

Starting from the extreme right-hand position (0), the beam after nine similar movements, reaches the extreme left-hand position (9). With the tenth pulse it must return to position 0, at the same time causing the beam of another counter tube, counting the tens, to move from 0 to 1. The re-set from 9 to 0 is initiated by the so-called reset anode (a_1, fig. 2*a*): one of the methods of resetting the tube is described in article I, page 323.

The counting of random pulses

A circuit to count 30 000 pulses per second is described in article I: it was there assumed implicitly that the pulses were periodic. However, it is often necessary to count completely random pulses, as in the case of a Geiger-Müller or scintillation counter. Almost all practical cases are covered by these two extremes.

Pulses belonging to the latter group can be counted mechanically or electrically, but in either case a fraction of the pulses produced will escape detection. In the decade scaler tube, for example, the beam requires a certain time to pass from one position to the next; during this time the tube is inactive and any pulses that happen to arrive will not be counted. Similar losses occur in all electronic scalers and the losses are even more serious in mechanical counters. This counting loss depends on the time required by the scaler to record one digit, that is, on the so-called dead-time. Hence it is usual to specify this characteristic property of a scaler.

However, in trying to do so in the case of the tube E1T we encounter a difficulty, i.e. that this tube has two dead-times, namely what may be described as the "step-time" (τ_s), or time required for stepping the beam from one position to the next (0-1, 1-2, ... 8-9), and the "reset time" (τ_r), required to reset the beam from 9 to 0. In general, τ_r exceeds τ_s. In the case of periodic pulses, the counting rate is limited only by the longer of the two dead-times, that is τ_r; if the interval between successive pulses is shorter than τ_r the tube will not operate. However, a more complex situation arises in the case of random pulses.

To understand this, consider pulses distributed entirely at random, and let n represent the average number of pulses recorded per second during a particular count. On the average, then, n will also represent the number of times per second that the counter is inactive. In nine out of every ten cases, the cause of this inactivity is a step, i.e. 0-1, 1-2, ..., 8-9, with its associated dead-time τ_s, and in only one of the ten cases is it the reset 9-0, with dead-time τ_r. It will be seen, then, that on an average, the counter is inactive for a time $\tau = (0.9\, n\tau_s + 0.1\, n\tau_r)$ second per second, i.e. it is able to count during only $(1-\tau)$ second in each second.

Accordingly, the number of pulses actually arriving per second is not n, but:

$$N = \frac{1}{1-\tau} n. \quad \ldots\ldots \quad (2)$$

Here, then, we have what may be described as an average dead-time τ_m:

$$\tau_m = \frac{\tau}{n} = 0.9\, \tau_s + 0.1\, \tau_r. \quad \ldots \quad (3)$$

Now, τ_s and τ_r are of the same order of magnitude: $\tau_s = 24$ μsec, $\tau_r = 27$ μsec [5]); therefore τ_s contributes considerably more than τ_r. We find that:

$$\tau_m = 0.9 \times 24 + 0.1 \times 27 = 21.6 + 2.7 = 24.3 \text{ μsec.}$$

which turns out to be smaller than τ_r. However, this by no means implies that it would be possible to count, on an average, more than 30 000 random pulses per second, as will be seen from the following. From equations (2) and (3): $n = N/(1 + N\tau_m)$. Substituting $\tau_m = 24.3$ μsec and $N = 30\,000$ pulses/sec, we find that $n = 17\,300$ pulses/sec; hence the loss is no less than 42 %.

Such a loss cannot be accepted. Ideally, in fact, the loss should be negligible, e.g. 1%, but the number of pulses that can be counted per second is then relatively small. To illustrate this point we have, from formula (2):

$$n = \frac{1}{\tau_m} \cdot \frac{N-n}{N} \quad \ldots\ldots \quad (4)$$

where $(N-n)/N$ represents the relative loss. It is seen, then, that n is proportional to the loss. If the loss be limited to 1% and if τ_m be 24.3 μsec, we find that $n = 411$ pulses/sec and $N = 1.01\, n \approx n$.

From formula (4), we see that if the number of pulses per second be increased, the loss will also increase, thus necessitating the use of a correction factor to calculate the real quantity (N) from the quantity recorded (n). It can be derived from (2) that this correction factor is $1/(1-\tau)$, which, in conjunction with formula (3), gives:

$$N = \frac{1}{1 - n\,\tau_m} n. \quad \ldots\ldots \quad (5)$$

The count n can be corrected, employing the correction factor to an extent consistent with the accuracy of τ_m, since the error in the final result arising from any inaccuracy in τ_m is all the larger, the larger the correction factor $1/(1-n\tau_m)$. In fact, if Δ be the accuracy of τ_m, and δ the required accuracy of N (both in %), it is necessary to satisfy the following condition:

$$n\,\tau_m \leqq \frac{\delta}{\Delta + \delta}.$$

For example, given that the required $\delta = 1\%$ and $\Delta = 10\%$, then $n\tau_m$ must not exceed 1/11. Accordingly, n may be increased to $1/(11\,\tau_m) = 3740$ pulses per sec. It will be seen from the above that even a moderate degree of accuracy in the determination of τ_m is sufficient to procure an increase of a factor of ten in the count rate.

However, if the loss is to be limited to 1% to enable the correcting factor to be dispensed with, but a count rate of about 400 pulses/sec is considered inadequate, it will be necessary to find a means of shortening the average dead-time τ_m. It will be evident that, at all events as far as random pulses are concerned, we should consider primarily the possibility of shortening the beam step-time (0 to 9) rather than the reset time, which contributes only 2.7 to the total of 24.3 μsec.

Now, the "step-time" τ_s may be divided into two parts, i.e. one governed by the scaler tube itself (in a given circuit), and one governed by the pulse shaper (that is, the circuit preceding the scaler tube, whose function is to convert each applied pulse into one of constant shape and

[5]) In article I, page 324, an average value of 27.2 μsec is specified, with due regard to the different safety margins. This value is here rounded off to 27 μsec.

amplitude suitable for counting). Let us now consider to what extent it is possible to shorten τ_s by employing an improved pulse shaper.

Improved pulse shaper

Fig. 4 is the circuit diagram of a pulse shaper able to shape pulses arriving at a higher rate than the corresponding unit described in article I (fig. 20). The circuit itself is fully described in another publication [6]).

It will be seen that fig. 4 is divided into two parts, i.e. the "squarer" (A), to convert the input signal (pulsating, sinusoidal, or square) into square pulses, and the actual pulse shaper (B), to convert the squared pulses into triangular ones.

The square pulse thus produced at P is differentiated in unit B by a combination of a capacitor (C_4) and a resistor (R_{13}). A diode (d_1) prevents the occurrence of any negative peak (corresponding to the jump from 173 V to 130 V), and another diode (d_2) passes the positive peak (resulting from the jump from 130 V to 173 V), which then produces an increase in the cathode voltage of d_2 (point Q) from 156 V to about 170 V.

Given a suitable choice of the time constant C_4R_{13}, it is possible to make the anode voltage of d_2 decrease faster than the cathode voltage, thus driving this valve beyond cut-off. From then on, the potential at point Q will decrease exponentially to an asymptotic value of 90 V, by reason of the

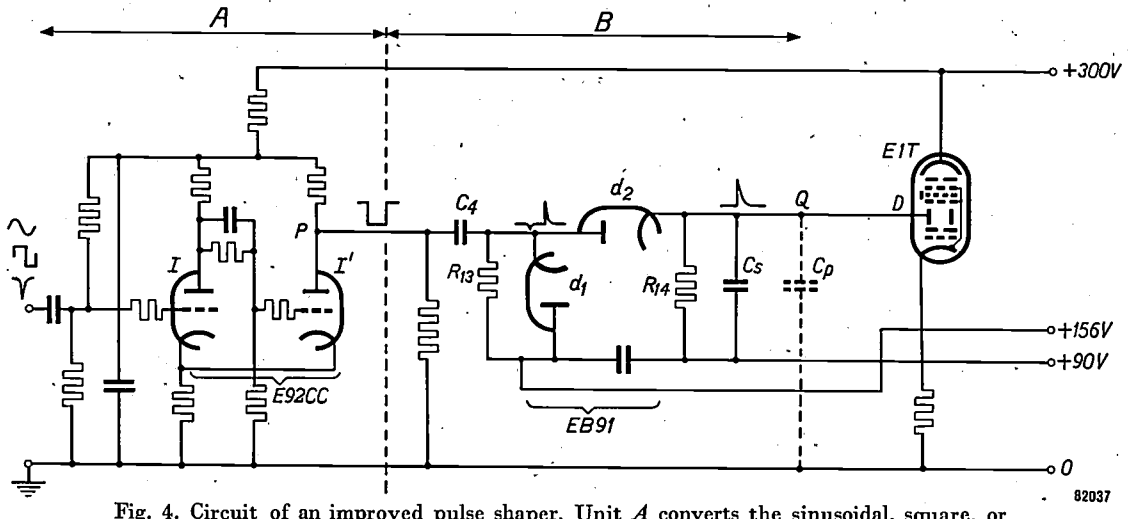


Fig. 4. Circuit of an improved pulse shaper. Unit A converts the sinusoidal, square, or pulsating input signal into square pulses at point P. Unit B differentiates these square pulses which are applied via point Q to the left-hand deflector plate (D) of scaler tube E1T. Section I of double triode E 92 CC is conductive, and section I' non-conductive, in the steady condition. A drop in the grid potential of section I reverses this situation, and a subsequent rise in potential restores it. C_4-R_{13} is a differentiating network. Section d_1 of double diode EB 91 cuts off the negative pulses, and section d_2 passes the peaks of the positive ones. $(C_5 + C_p)R_{14}$ is the time constant of the trailing edges of the pulses occurring at Q (C_p stray capacitance).

The principal component of unit A is a double triode E 92 CC, of which one section (I) is conductive and the other (I') non-conductive, in the quiescent condition; the anode voltage of section I' is then 173 V.

A negative pulse arriving on the grid of section I will drive this section beyond cut-off, and make section I' conductive for the duration of the pulse. During this pulse, then, the anode voltage of section I' will decrease to 130 V, and after it the original condition of the tube will be restored. Accordingly, the potential of point P in the circuit (see fig. 4) drops first from 173 V to 130 V, and then returns to 173 V.

fact that capacitor C_5 (together with the stray capacitances in parallel with it) discharges across resistor R_{14}. However, as soon as the decreasing potential reaches 156 V (that is, the anode voltage of d_1), diodes d_1 and d_2 become conductive again, thus maintaining the potential of point Q constant at or near this value. Since it is small, the voltage drop from 170 V to 156 V is virtually linear, as will be seen from *fig. 5*. Given certain values of C_5 and R_{14}, the base-width of the triangular pulse, with an amplitude of 14 V, thus produced is 13 μsec; such a pulse is suitable for the E1T scaler tube. This enables us to connect point Q to the left-hand deflector plate D (fig. 4).

The dead-time to be taken into account in the counting of random pulses is:

$$\tau_m = 0.9\,\tau_s + 0.1\,\tau_r = 11.7 + 2.7 = 14.4\ \mu\text{sec}.$$

[6]) R. van Houten, A decade counter stage with a counting rate of 100 000 pulses per second, Electronic Appl. Bull. 15, 34-43, 1954, fig. 3.

Assuming that a 1% loss is acceptable, it follows from (4) that $n = 10^{-2}/14.4 \times 10^{-6} = 695$ pulses/sec, that is, a counting rate much higher than that calculated above.

Again, the question arises whether even an average counting rate of 695 pulses/sec is high enough. However, it will be seen from the following argument that this average is at all events of the correct order of magnitude for many cases.

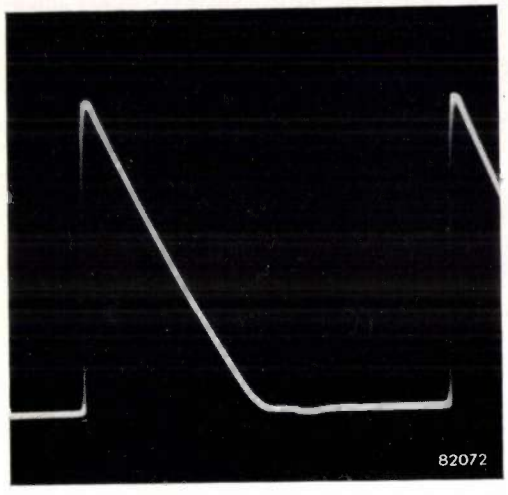


Fig. 5. Oscillogram of a pulse produced at point Q (fig. 4) by a 50 000 c/s A.C. voltage applied to the input of the pulse shaper. Amplitude 14 V, duration 10 μsec.

If N is to be measured accurately to within 1%, the total number of pulses counted will have to be large enough to ensure that the natural fluctuation (which is completely independent of the dead-time of the scaler) does not produce any error greater than 1%; the minimum total pulse count required to satisfy this condition is 10 000. Moreover, it will then be necessary to measure the actual counting time accurately to within 1%; the shortest counting time that can be measured with a stopwatch with this accuracy is 20 seconds. Now, although the fact that at least 10 000 pulses must be counted in at least 20 seconds does not necessarily imply that 10 000/20 = 500 is either the precise, the average, or the maximum number of pulses per second that the scaler must be able to record. It does imply, however, (assuming a count accurate to within 1%) that this is the correct order of magnitude. In general, there is little to be gained by extending the count beyond a total of 10 000 pulses, since the extension required to produce any essential increase in counting accuracy would be enormous.

In the case of radioactive radiation so intense as to produce more than 10 000 pulses in 20 sec, it is usually possible to compromise either by increasing the distance between the sample and the counter, or by employing a diaphragm. However, such measures are sometimes undesirable, as in the case of coincidence circuits. In these circuits two or more scintillation counters are connected such that a count is recorded only when both counters are struck by the same elementary particle. The number of such coincidences may be expressed as a fraction of the total number of particles striking one of the counters in the same period. This fraction may be very small. Where it is necessary, owing to the low counting rate of the scaler, that the number of pulses per second be kept relatively small, coincidences will be so rare that adequate accuracy can be ensured only by counting over a very long period. Accordingly, such measurements require scalers capable of an average counting rate much higher than some hundreds of pulses per second (assuming a loss of 1%). The same holds good for the counting of a wide variety of effects whose rate of repetition is not controllable.

Hence it is well worth while to examine the possibility of shortening the step-time τ_s still further.

Reducing the contribution of the scaler tube to the step-time

The electron beam in the scaler tube is advanced one step at a time by a count-pulse, that is, a sudden rise in the potential of the left-hand deflector plate (D, fig. 2). Owing to the stray capacitance of the interconnected anode (a_2) and right-hand deflector plate (D'), the change in the potential of these electrodes is relatively slow. After each step of the beam in the left-hand direction, the trailing edge of the pulse takes effect, i.e. the potential of the left-hand deflector plate gradually returns to its original level, so gradually, in fact, as to enable the beam to remain in the new position.

At the same time, the potential of the anode and the right-hand deflector plate drops by an amount $V \approx 14$ V. Then, however, the charge on these two electrodes (whose stray capacitance will be denoted by C_p) also changes, a negative charge $Q = C_p V$ being added to it. This charge is produced as follows.

Having been transferred one step to the left by a positive voltage step of the left-hand deflector plate, the beam tries to return to its former position, that is, to move to the right, during the trailing edge of the count-pulse. As will be seen from fig. 3, the anode current then tends to increase, thus providing extra current to charge the stray capacitance. The anode current increases all the faster, and the stray capacitance is therefore charged all the sooner,

the more rapidly the potential of the left-hand deflector plate is reduced.

The speed of these changes is limited owing to the fact that the above-mentioned extra current itself cannot exceed a certain limit. Some of the peaks of the $i_{a2} = f(V_{D',a_2})$ characteristic (see fig. 3) lie only about 100 μA above the stable points, and in fact, to ensure a thoroughly reliable result, we do not employ more than 20 μA of this margin (controlled by giving a suitable slope to the trailing edge of the count-pulse). The extra current (I) must reduce the potential difference across the stray capacitance (15 pF) by 14 V. The time required for this reduction is:

$$\vartheta_2 = \frac{C_p V}{I} = \frac{15\times10^{-12}\times 14}{20\times10^{-6}} \cdot 10^6 = 10.5 \text{ μsec.} \quad (6)$$

If the build-up time (ϑ_1) of the pulse, and a certain safety margin, be added to ϑ_2, the duration of the pulse will be at least 13 μsec. Accordingly, ϑ_2 must be shortened. To see how this may be accomplished, let us consider equation (6). It will be clear that it is impossible to reduce either the stray capacitance C_p, or the potential difference V between two stable points. This leaves only the possibility of increasing the charging current I. It cannot be increased through the agency of the scaler tube itself; however, an extra current can be supplied from an outside source. Now, the direction of this current must be such that it produces a drop in the potential of the anode a_2 and the right-hand deflector plate D'; in other words, the current must supply electrons to these two electrodes. A current in the required direction can be obtained by connecting the anode of a suitable valve, e.g. a triode, to the right-hand deflector plate and the anode of the scaler tube (*fig. 6*). Such a "charging valve" must be cut off during the major part of the time, that is, it must pass current only for a brief period during the counting of a pulse.

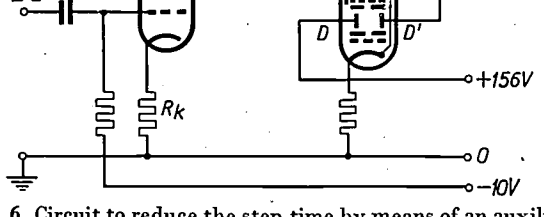


Fig. 6. Circuit to reduce the step-time by means of an auxiliary triode (T_1, "charging valve") which becomes momentarily conductive when gated by an incoming pulse and then supplies a negative charge to the right-hand deflector plate (D') of the scaler tube. The potential of the left-hand deflector plate (D) is constant (156 V). R_k cathode resistor.

Here, a difficulty arises, namely that the charging pulses applied must not be of uniform size, but should be smaller, the lower the voltage v_{D',a_2} of the right-hand deflector plate and the anode. Given a decrease in v_{D',a_2} (owing to the increase in anode current resulting from each step of the beam, i.e. 0-1, 1-2, and so on), the average potential in the space between the deflector plates will also decrease, by reason of the fact that the deflection is asymmetrical; hence the deflection-sensitivity will increase, that is, the stable points of intersection of the resistance line and the $i_{a_2} = f(v_{D',a_2})$ characteristic will lie closer together at a low, than at a high v_{D',a_2}. Now, the points *0,1, 2, ... 9*, as shown in fig. 3 are evenly spaced, but in fact the horizontal distance between *0* and *1* corresponds to 17.5 V (v_{D',a_2} being 245 V for position 0) and that between *8* and *9* to 12.8 V (v_{D',a_2} being 109 V for position 9)

Hence the ratio of the two extreme positions of the beam is $17.5 : 12.8 \approx 4 : 3$.

Accordingly, the circuit should be so designed that the charge supplied by the charging valve decreases with the anode voltage in such a way that the ratio of the charges consistent with the highest and lowest values of the anode voltage is 4 : 3. This condition is satisfied by providing the triode employed as a charging valve (fig. 6) with a suitable cathode resistor R_k, whose value can be established by the method illustrated in *fig. 7*. This diagram shows two characteristics of the triode

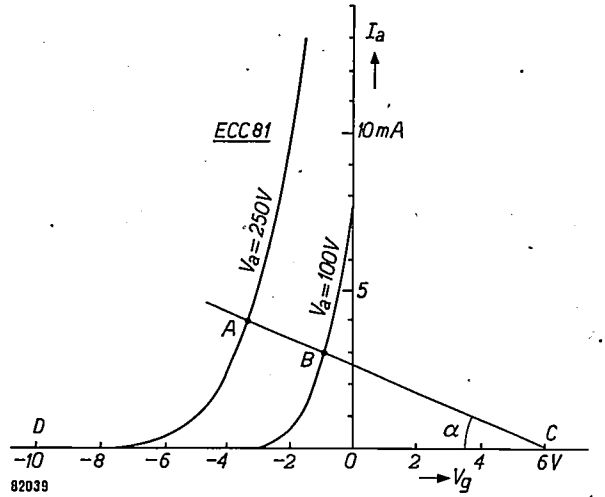


Fig. 7. $I_a = f(V_g)$ characteristics for $V_a = 100$ V and $V_a = 250$ V of one half of double triode ECC 81. Note the straight line drawn through points A (at 4 mA on the characteristic for $V_a = 250$ V) and B (at 3 mA on the characteristic for $V_a = 100$ V). It will be seen that this line makes an angle α with the V_g axis and cuts it at a point C. R_k (fig. 6) is equal to cot α. The required pulse amplitude is $CD = 16$ V.

employed (one half of a double triode ECC 81), viz. the anode current I_a plotted as a function of the grid voltage V_g for two values of the anode voltage V_a (100 V and 250 V). Now, a certain point A, corresponding to a current I_a defined by $I_a = Q/\Delta t = CV/\Delta t$ (where Δt is the duration of the pulse and V is 17.5 V), say $I_a = 4$ mA, is chosen on the characteristic for $V_a = 250$ V (practically equal to the highest anode voltage, 245 V), and another point B, whose ordinate is $^3/_4$ $I_a = 3$ mA, on the characteristic for $V_a = 100$ V (almost the lowest anode voltage, 109 V). The angle α of line AB then governs the size of R_k (= cot α). Line AB cuts the V_g co-ordinate at a point C, corresponding to $V_a = 6$ V. Since V_g must be about -10V to drive the valve far enough beyond cut-off at $V_a = 250$ V, the amplitude (DC) of the pulse must be about 16 V.

It is also necessary to establish the time Δt, since the precise value of R_k depends on it. Again, it is advisable to employ a pulse shaper capable of converting variable pulses (as produced, say, by a scintillation counter) into pulses of suitable size and constant duration required for counting. A suitable shaper is described in the Appendix (the pulse shaper shown in fig. 4 cannot be used here, since it does not produce pulses of the correct shape and duration.).

A charging current of 3 mA, that is, 150 times the original current (20 μA), is readily obtained by the above method. On the other hand, the charging valve almost doubles the stray capacitance of the scaler; despite this, however, an overall improvement by a factor of 75 is obtained, and time ϑ_2 is thus shortened to $10.5/75 = 0.14$ μsec.

The average dead-time τ_m may be calculated with the aid of formula (3), assuming firstly that the counts 0-1, 1-2, etc. involve no dead-time other than the 0.14 μsec already derived (hence $\tau_s = 0.14$ μsec), and secondly that the re-set time τ_r is still 27 μsec; we then have:

$$\tau_m = 0.9 \times 0.14 + 0.1 \times 27 = 0.13 + 2.70 = 2.83 \text{ μsec.}$$

Here, then, it would be possible to count, on the average, $10^{-3}/(2.83 \times 10^{-6}) = 3540$ random pulses per second (at 1% loss).

In general, however, the apparatus preceding the scaler tube (scintillation counter, photomultiplier, amplifier, limiter, pulse shaper) is also affected by a certain dead-time (τ_c), usually exceeding 0.1 μsec, and in many cases even longer than 0.5 μsec. It will be evident that to make the dead-time of the scaler shorter than τ_c would serve virtually no useful purpose, since the scaler would then be ready to count some time before the preceding circuit could be ready to supply pulses.

In fact, we are really concerned only with the overall dead-time of the apparatus. This, also, can be calculated with the aid of equation (3), provided that τ_c be substituted for τ_s if it happens to exceed this, and likewise for τ_r if it be the greater of the two. Given $\tau_s = 0.14$ μsec, $\tau_r = 27$ μsec and $\tau_c = 1$ μsec, then, we have:

$$\tau_m = 0.9\,\tau_c + 0.1\,\tau_r = 0.90 + 2.70 = 3.60 \text{ μsec,}$$

which corresponds to: $n = 2800$ pulses/sec (at 1% loss).

Reducing the reset time

According to our original argument, the reset time is less important than the step-time as a means of increasing the counting rate for pulses occurring at irregular intervals. However, it will be seen from the last numerical example that the measures described so far enable the step-time to be so shortened that the major part of the average dead-time τ_m is in fact now the reset time τ_r. Although this in itself is reason enough to examine the possibility of reducing τ_r, there is a further reason, that is, that τ_r (and it alone) governs the maximum counting rate for periodic pulses.

Acceleration of the reset by means of a flip-flop circuit

The obvious course is to try to shorten the reset time in very much the same way as the step-time, that is, by employing an ancillary valve to reduce the retarding effect of the stray capacitance either by supplying, or by taking away, a charge.

During the reset, the potential of the right-hand deflector plate rises. To accelerate this rise — and so the reset — therefore, it is necessary to take away electrons from the right-hand deflector plate. Here, then, the cathode, not the anode, of the ancillary valve must be connected to the right-hand deflector plate (and the anode) of the counter tube.

Fig. 8 shows, in principle, how this may be done;

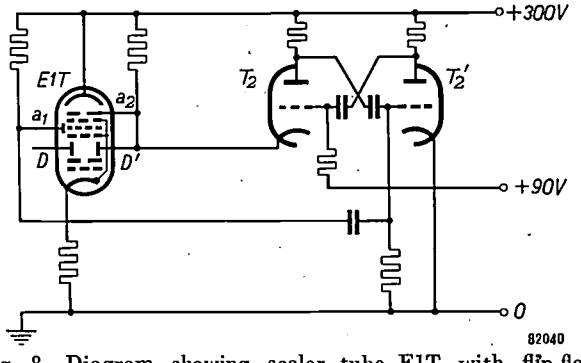


Fig. 8. Diagram showing scaler tube E1T with flip-flop circuit (T_2-T_2') to shorten the reset time. In the steady state T_2 is cut off and T_2' conductive. This negative pulse produced on the reset anode (a_1) when the beam returns from 9 to 0 reverses this situation.

a flip-flop circuit is employed. Ordinarily, the right-hand half T_2' of the valve in this circuit is conductive and the left-hand half (T_2) is cut off. However, a slight increase in the negative grid bias of T_2' is enough to reverse the situation; T_2' is then driven beyond cut-off and T_2 becomes conductive, thus drawing electrons from the right-hand deflector plate (D'), whose potential therefore rises very rapidly. This goes on until the current of T_2 decreases owing to the decrease in the potential difference between anode and cathode; T_2' then rises above cut-off and the circuit is restored to its original condition.

The flip-flop is triggered by the reset anode (a_1) of the scaler tube, that is, by a drop in the potential of this anode produced by the beam striking it each tenth pulse.

Circuits based on this principle lead to a very short reset time, viz. 4 μsec (ordinarily 27 μsec). Reduction of this time still further would involve the disadvantage that the auxiliary valve, whose anode current is inversely proportional to the reset time, would then draw a rather heavy current. Furthermore, these circuits impose a severe strain on the insulation between cathode and heater.

These disadvantages will now be explained. Consider the anode current first; during the reset, the cathode potential of the auxiliary valve T_2 rises from + 90 to about 245 V. To keep the valve about cut-off despite this rise, the potential difference between grid and cathode must remain virtually constant; in other words, the circuit parameters must be so chosen that the grid potential will also increase by about 245 — 90 = 155 V during the reset.

Unless the anode resistance be low, however, stray capacitance will make this increase too slow, and to obtain the necessary voltage drop across a low anode resistance it is necessary to employ a rather heavy current.

With regard to the heater insulation, if the heater be supplied from a transformer winding earthed, as usual, at the centre, the voltage between cathode and heater will rise well above the limit imposed on most valves. Hence it is necessary to employ a separate heater-current winding and to maintain it at a potential relative to earth roughly midway between 90 and 245 Volts (e.g. at 156 V, this being required, in any case, as bias for the left-hand deflector plate).

Despite this separate winding, however, even a very small leakage current may affect the count, since it will augment the anode current of the scaler tube and may thus eliminate one or more of the stable beam-positions. Accordingly, the auxilliary valve employed must have very good cathode-heater insulation.

However, in the circuit that will now be described, these disadvantages are avoided by employing a cathode with secondary, instead of thermionic, emission.

Acceleration of the reset by means of a secondary emission tube

Fig. 9 is the schematic representation of a secondary emission tube, type EFP 60. In this tube, primary electrons produced by a thermionic cathode (k_1) pass three grids (g_1, g_2 and g_3) and then strike the dynode (k_2). The dynode is an electrode coated with a substance with a high secondary emission

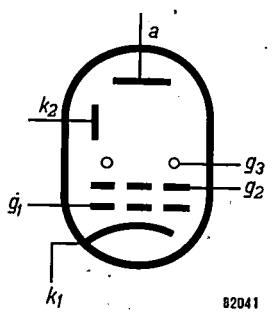


Fig. 9. Schematic representation of a secondary emission tube. k_1 thermionic cathode, g_1 control grid, g_2 screen grid, g_3 suppressor grid, k_2 dynode (secondary emission cathode), a anode.

factor, that is, a substance from which each incident primary electron releases several secondary electrons. Since the anode (a) is positive with respect to the dynode, the secondary electrons thus released proceed to the anode. By virtue of this secondary emission, the anode current of the tube exceeds its cathode current, the difference being supplied by the dynode.

Now, if the dynode be connected to the anode and to the right-hand deflector plate of the scaler tube (*fig. 10*), and the secondary emission tube be gated

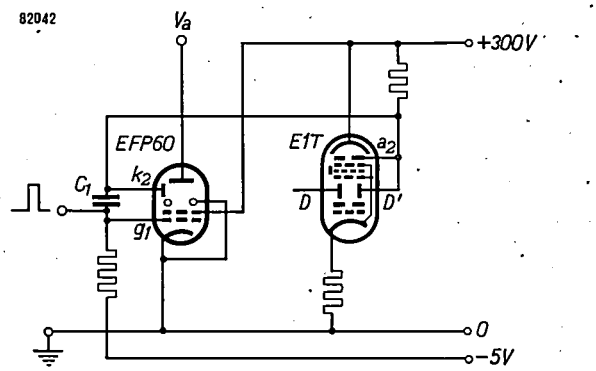


Fig. 10. Diagram showing scaler tube E1T with secondary emission tube EFP 60 to shorten the reset time. Tube EFP 60 is triggered by a positive pulse on g_1 and maintained conductive for the required period by coupling capacitor C_1 between k_2 and g_1.

at the correct moment, the desired effect will be obtained, that is, the dynode will draw electrons from the right-hand deflector plate, producing a sharp rise in the potential of this plate and thus accelerating the reset.

Here, then, we are presented with the problem of triggering the secondary emission tube at the correct moment.

To accomplish this, it is necessary to procure a positive voltage step on the control grid (g_1) at the precise moment when the reset is to take place. However, the only voltage step available at that moment is a negative one (on the reset anode, a_1, of the scaler tube). We therefore convert this negative step into the required positive one by means of a single stage amplifier between the reset anode and the control grid g_1 (*fig. 11*). In the present circuit, a double triode ECC 81 is employed, the one half (T_1') operating as an amplifier, and the other (T_1) as a charging valve to reduce the step-time (see fig. 6).

The secondary emission tube must remain conducting long enough to ensure the complete reset of the scaler tube. To ensure that it will do so, the control grid of the former is coupled to the dynode across a capacitor (C_1, fig. 10 and fig. 11); during the reset the potential of the dynode, and therefore

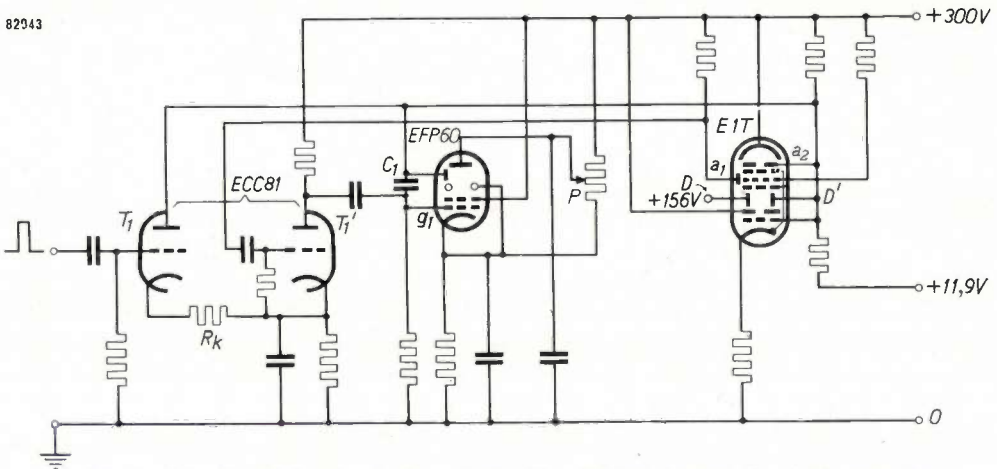


Fig. 11. Combination of the circuit shown in fig. 10, with that shown in fig. 6. The negative step produced on a_1 at the start of the reset is converted by amplifier T_1' into a positive step to bias the control grid (g_1) of secondary emission valve EFP 60. P potentiometer to control the anode voltage of valve EFP 60 (that is, to compensate for variations between individual scaler tubes). T_1 is a charging valve to shorten the step-time (see fig. 6). T_1 and T_1' are the two sections of a double triode ECC 81.

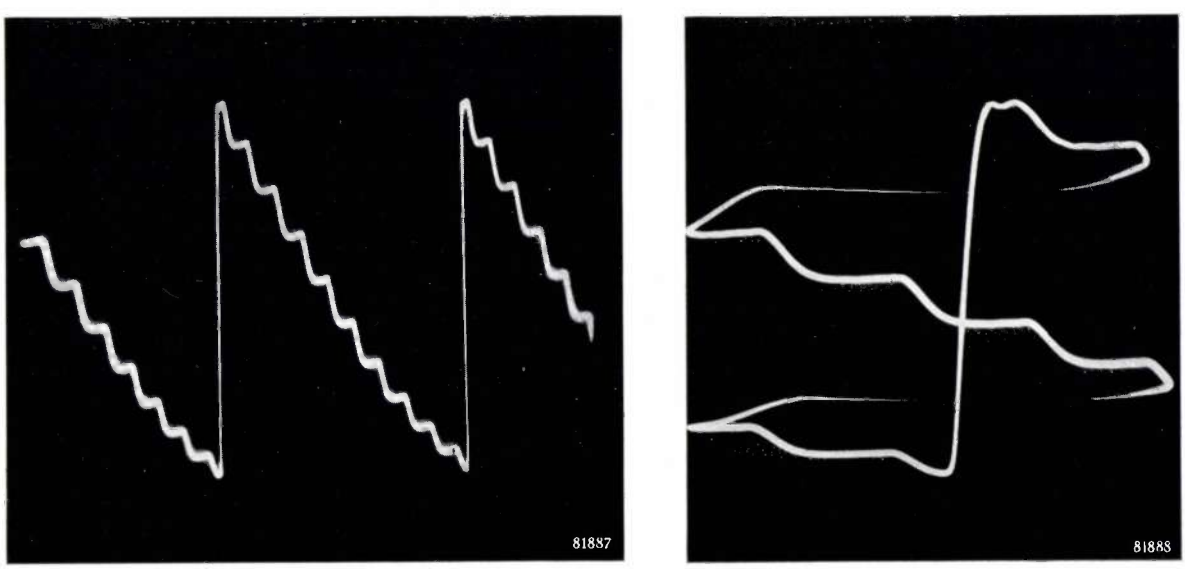
that of the control grid, rises, thus maintaining the valve in operation.

The dynode potential continues to rise until it nearly equals the anode potential; then, however, secondary electrons return to the dynode, the rise in the potential of this electrode reaches its limit, and the beam reaches the end of the reset. Here, then, we have a simple means of compensating for variations as between individual E1T scaler tubes, that is, for differences in the particular voltage V_{D',a_2} associated with position 0; without such compensation, the beams in some of these tubes might well reset only as far as position 1, instead of to position 0, owing to the above-mentioned "spread". The method of compensation consists in making the anode voltage of the secondary emission valve variable (potentiometer P in fig. 11).

After the reset, the positive charge leaks away from the control grid of the secondary emission tube, thus cutting off the tube and restoring the original situation.

The circuit shown in fig. 11 enables the reset time to be reduced to 0.45 μsec, thus raising the maximum counting rate for *regular* pulses to 2.2 million per second (*fig. 12*).

As far as *random* pulses are concerned, it should



a *b*

Fig. 12. Oscillograms of V_{D',a_2} in the circuit shown in fig. 11, taken during a count of regular pulses at the rate of 2 million per second; the peak count rate attained by the scaler was 2.2 million. Note that the speed of horizontal deflection was made a factor of four greater in (*b*) than in (*a*); this was done for more accurate measurement, particularly of the reset time.

be borne in mind that a scintillation counter with a photomultiplier has a dead-time of about 0.20 μsec (see the articles referred to in note [2])). This exceeds the step-time of 0.14 μsec obtained in the manner described; at the same time, the contribution of the reset to the average dead-time is only $0.1 \times 0.45 = 0.045$ μsec. The average dead-time is therefore:

$$\tau_m = 0.9 \times 0.20 + 0.045 = 0.225 \text{ μsec.}$$

Accepting a loss of 1%, this value of τ_m corresponds to a counting rate for random pulses of 44 500 per second; this rate increases directly proportionally to the loss which is acceptable.

Finally, it should be noted that in as far as this article refers to laboratory work, the results quoted in this article take no account of tolerances on valves and other circuit elements, whereas such tolerances are included in the results given in article I.

Appendix: A pulse shaper to drive the charging triode

To operate efficiently, the charging triode, which shortens the step-time (T_1 in fig. 6 and fig. 11), should be driven by pulses of constant height and width. However, the pulses produced by particle counters, particularly scintillation counters, do not satisfy this requirement. Hence a pulse shaper is employed to convert them into pulses of uniform width and of the required amplitude. To give an idea of the manner in which such a pulse shaper operates, one or two of the possible designs will now be briefly described.

Every elementary particle striking the particle counter ultimately produces an avalanche of electrons, that is, in the tube itself if it be a Geiger-Müller tube, and in the associated photomultiplier if a scintillation counter is employed. The electrons so produced strike an electrode, thus charging its stray capacitance which subsequently discharges through a resistor (*fig. 13*). In this way, saw-tooth voltage pulses having a

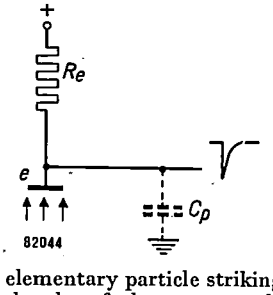


Fig. 13. Each elementary particle striking a particle counter produces an avalanche of electrons on electrode. *e* This, in turn, produces at this electrode a negative voltage step having a steep front and an exponential decay. The time constant of this tail is R_eC_p, where R_e is the charging resistance of electrode *e* and C_p the stray capacitance.

steep front and an exponential decay are produced. Owing to the appreciable variation between the individual charges supplied, the pulses differ considerably in amplitude, the amplitude variation between suitable pulses being as 1 : 20; the smallest pulses are those which just emerge above the noise-level. Since the time constant of the exponential decay is the same for all the pulses, the differences in amplitude correspond to differences in duration.

The usual procedure is to cut down the peak pulses to some extent by limiting, the signal thus obtained being employed — if necessary after amplification — to control a multivibrator, e.g. as shown in *fig. 14*. The multivibrator reverses every time the input signal passes a certain level. The output signal then consists of square-waves equal in amplitude but different in duration. Now, pulses of uniform duration can be obtained by differentiating the square-wave signal in an *RC* network; this produces identical positive and negative pulses. If desired, the negative pulses can be cut off by means of a diode.

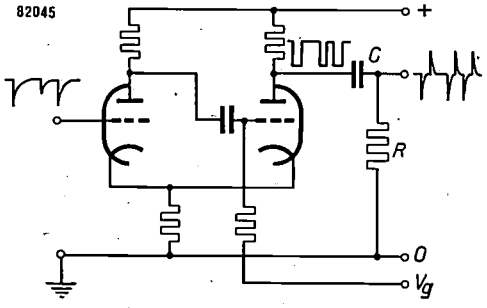


Fig. 14. Diagram of a multivibrator to convert incoming, dissimilar pulses into square pulses of uniform amplitude but different widths. Positive and negative pulses identical in shape and size are obtained by differentiating these square pulses (capacitor *C*, resistor *R*). The input-voltage level to trigger the multivibrator can be varied by varying the grid bias V_g.

The input-signal level to trigger the multivibrator can be varied by varying the grid bias V_g (fig. 14), this bias then being so adjusted that the smallest pulses that can be regarded as noise, have no effect. However, this involves the disadvantage that two pulses arriving in quick succession are likely to be counted as one, since if the second pulse happens to arrive before the first has dropped below trigger level (*fig. 15*), it cannot trigger the multivibrator and will therefore be missed in the count.

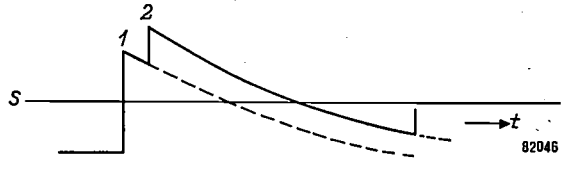


Fig. 15. If the second of two successive pulses (*2*) arrives before the first (*1*) has dropped below the trigger-level (*S*) of the multivibrator (fig. 14), pulse *2* will not be counted.

A considerable improvement in this respect is obtained by employing a delay line in the pulse shaper (*fig. 16a*). The electron avalanche on electrode *e* produces a negative voltage step at *e*. This step is propagated along the line (that is, from right to left in the diagram). Now, the end of the line is short-circuited, so that reflection with change of sign takes place, producing a positive step which travels from left to right. The input end of the line is terminated with a resistor (viz. resistor R_e, through which electrode *e* is supplied): this resistor matches the characteristic impedance of the line and therefore prevents reflection at the input. The effect of this arrangement is, then, that the negative step is followed, after an interval equal to twice the delay-time of the line, by a positive step, the two steps finally combining to form one square pulse (fig. 16*b*). All that is then necessary to make such a square pulse suitable to drive the charging triode is to change its sign and give it the required height.

Since a real transmission line providing the necessary delay would be many metres long, an artificial transmission line is employed; this may be a filter of coils and capacitors, or a specially designed transmission line, of high self-inductance and capacitance per unit length so that the required length is kept within reasonable bounds (say 10 or 20 cm).

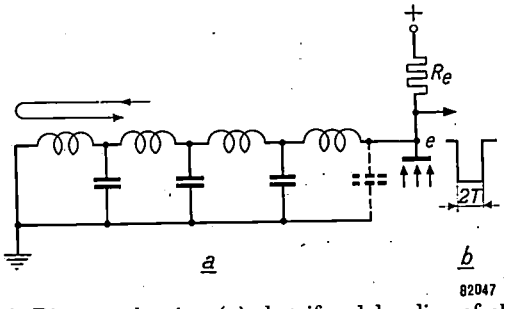


Fig. 16. Diagram showing (*a*) that if a delay line of characteristic impedance R_e, short-circuited at the free end, be connected to anode *e* (fig. 13), the electron avalanche will produce on this anode square pulses (*b*), whose duration is twice the delay (T) in the line.

In practice, it may be necessary to employ one decade scaler in conjuction with several particle counters whose pulse shapers produce pulses differing in width. Here, it would be necessary to vary the bias resistance R_k of the charging triode for each incoming pulse, which would not only be difficult in itself, but would also be a possible source of error in the count; thence it is better to introduce another pulse shaper (*C*, *fig. 17*) to convert the different pulses into identical

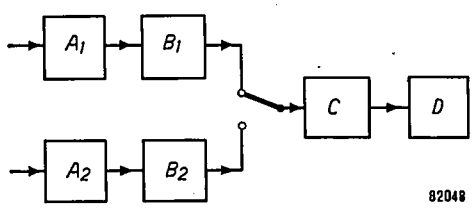


Fig. 17. A_1, A_2 particle counters, and B_1, B_2 their associated pulse shapers. If the pulses produced by B_1 and B_2 differ in duration, another pulse shaper (*C*) should be added to convert these pulses into pulses of uniform duration. *D* scaler.

ones. By a fortunate chance, the charging triode is able to operate with pulses shorter than those produced by most particle counters. When once the dissimilar pulses have been rendered uniform in amplitude by a limiter, pulses of identical (triangular) shape can be derived from them (*fig. 18*); these derived pulses are very short, but nevertheless usable. If the amplitude of such triangular pulses is too small, they can be re-converted into square pulses by means of a multivibrator. Any sensitivity on the part of the counter to the width of the original pulses is thus eliminated.

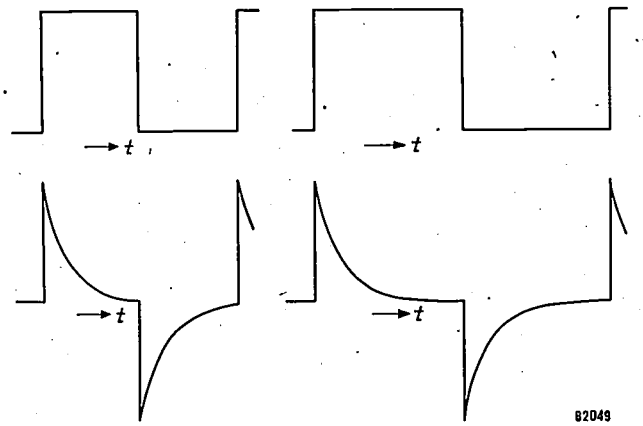


Fig. 18. The identical triangular pulses to gate the charging triode (T_1 in fig. 6 and fig. 11) are obtained by differentiating square pulses uniform in amplitude but differing in width.

Summary. This article describes circuits developed to count as many pulses per second as possible, particularly random pulses by means of a decade scaler tube type E1T. Tube E1T has two "dead-times", viz. the step-time (τ_s) required for each step (0-1, 1-2,..., 8-9), and the re-et time (τ_r) to return the beam from 9 to 0. The maximum counting rate for regular pulses is governed entirely by τ_r (since $\tau_r > \tau_s$). In the case of random pulses, an "average dead-time" (τ_m), depending on τ_s as well as τ_r, governs the counting rate; here, the average number of pulses recorded per second (n) is invariably smaller than the number of pulses arriving per second (N). Provided that N is not taken unduly large, the difference between it and n will be negligible, e.g. less than 1%. Where N is relatively large, its value can be calculated from n with the help of a correction factor. The effect of a possible inaccuracy in τ_m is more apparent in the correcting factor, the greater the difference between N and n.

An improved pulse shaper, a method of shortening the step-time, and two methods of shortening the reset time are described. The highest count rates attained are 44 500 random pulses per second with a loss of 1%, and 2.2 million regular pulses per second. Particulars of the pulse shaper used in the circuit for shortening the step-time are given in an appendix.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Administration of the Philips Research Laboratory, Eindhoven, Netherlands.

2150: K. ter Haar and J. Bazen: The titration of bismuth with "Complexone III" at *p*H 2.0-2.8 (Anal. chim. Acta **10**, 108-112, 1954, No. 2).

A titration procedure for Bi is described; by adding an excess of "Complexone III" (Versene) and back-titrating at *p*H 2.0-2.8 with standard thorium nitrate solution using "Alizarin-S" as an indicator, it is possible to determine Bi with a precision of about 0.3%.

By choice of the proper *p*H it is possible to eliminate most of the interfering elements, but at *p*H 2.0, Fe and Ni and Cu (slightly) still interfere. Attention is drawn to the fact that it may be possible to determine Bi at *p*H 2.0 by direct titration with "Complexone III" in the presence of excess tartrate, using thiourea as an indicator. This method might be of value in the presence of Sn and Sb.

2151: E. W. Gorter: Ionic distribution deduced from the g-factor of a ferrimagnetic spinel: Ti^{4+} in fourfold coordination (Nature, London **173**, 123, 1954, Jan. 16).

Where, in oxides with spinel structure, the distribution of metallic ions over tetrahedal (A) and octahedral (B) sites cannot be found with sufficient accuracy by X-ray diffraction, magnetic data may provide the necessary information, particularly the saturation moment and the factor $g = 2M_{tot}/M_{spin}$, where M_{tot} is the magnetic moment due to orbits plus spins. Because of the negative exchange coupling, the M's are equal to the difference $M_A - M_B$ for both kinds of ions. It is found that for a ferrite $Ni^{2+}_{1.5}\ Fe^{3+}\ Ti^{4+}_{0.5}\ O^{2-}_4$, the distribution corresponds to the formula $[Fe_{0.7}\ Ti_{0.3}]_A\ [Ni_{1.5}\ Fe_{0.3}Ti_{0.2}]_B\ O_4$.

2152: L. Heijne, P. Schagen and H. Bruining: Television pick-up tube for both light and X-ray pictures (Nature, London **173**, 220, 1954, Jan. 30).

Short description of a television camera tube provided with a screen consisting of lead oxide, evaporated on a "Pyrex" glass window. The layer is scanned by a low-velocity electron beam. The light sensitivity (peak in blue part of spectrum) amounts to 100-200 μA/lm ($T_C = 2600$ °K). The relatively high sensitivity to X-rays is the result of the high absorption coefficient of lead (5% absorption in a 5 micron layer for 70 kV D.C; filter 5 mm Al). Photographs are given, showing the same objects illuminated with visible light and with X-rays (shadow picture.)

2153: R. van der Veen and J. Daams: Experiments on the growth of helianthus seedlings (Proc. Kon. Ned. Akad. Wetensch. Amsterdam, Serie C**57**, 81-91, 1954, No. 1).

The growth of isolated parts of the stems of helianthus seedlings in light and in darkness was investigated. The experiments indicate that the growth of green stem sections is controlled by a photochemical reaction. This reaction results in a much quicker generation of organic phosphates, which in turn accelerates growth if auxin is present. The formation of phosphates is not due to enhancement of respiration by illumination, as is evident from the fact that stem parts containing no chlorophyll showed no enhanced respirations. Therefore photosynthesis is more probable. This is corroborated by an investigation on spectral sensitivity to be dealt with in a forthcoming paper.

2154: A. Claassen, L. Bastings and J. Visser: A highly sensitive procedure for the spectrophotometric determination of aluminium with 8-hydroxyquinoline and its application to the determination of aluminium in iron and steel (Anal. chim. Acta **10**, 373-385, 1954, No. 14).

Aluminium hydroxyquinolate can be quantitatively extracted by chloroform from an ammoniacal solution containing hydroxyquinoline, complexone and cyanide.

Titanium, vanadium, tantalum, niobium, uranium, zirconium, gallium, antimony, bismuth, indium and traces of beryllium are similarly extracted. Aluminium can be separated from the first five elements by an extraction in ammonical solution containing hydrogen peroxide. Zirconium, gallium, bismuth and antimony can be eliminated by a cupferron extraction and indium by extraction with diethyldithiocarbamate. Beryllium is eliminated by performing an extraction with hydroxyquinoline at *p*H 5. The proposed method enables a practically specific photometrical determination of aluminium. Applications are given of the determination of trace and

higher amounts of aluminium in steels, non-ferrous alloys and in glass.

2155: W. J. Oosterkamp: General consideration regarding the dosimetry of roentgen and gamma radiation. Addendum. (Appl. sci. Res. B3, 477-478, 1954).

Addition to the original paper under the same heading (Appl. sci. Res. **B3**, 100-118, 1953) to adapt this to the new recommendations of the International Commission on Radiological Units (Copenhagen, July 1953). The definitions of the new quantity *absorbed dose* and the unit in which this quantity is measured, the rad (= 100 erg/gram) are given.

2156: H. P. J. Wijn: Ferromagnetische resonantie en relaxatieverschijnselen (Ned. T. Natuurk. **20**, 45-53, 1954, No. 3). (Ferromagnetic resonance and relaxation phenomena; in Dutch.)

See these abstracts, No. **2092** and **2128**.

2157: J. Volger: Further experimental investigations on some ferromagnetic oxidic compounds of manganese with perovskite structure (Physics **20**, 49-66, 1954, No. 1).

Polycrystalline substances of the type $La_{1-\delta}Sr_{\delta}MnO_3$ have been investigated. These are ferromagnetic semiconductors which exhibit some remarkable second order effects related to the electrical conductivity. The specific heat has been found to be in agreement with Weiss' theory. New data on the direct current resistivity including that at liquid hydrogen temperature have been obtained. The direct current magneto-resistance appears to be independent of the mutual orientation of magnetic field and current; most of the effect seems to be due to domain rotations. The alternating current resistivity of some samples depends strongly upon frequency and high values of the dielectric constant have been observed at low frequencies. Samples exhibiting these properties also show a magneto-resistance dependent on the frequency of the alternating current, and a variation of both resistivity and magneto-resistance with applied voltage. These effects strongly support the hypothesis of barrier-layer resistance in ceramic semiconductors. A phenomenological analysis of these effects is given and may be of some importance for the general problem of magneto-resistance in ferromagnetics. The Seebeck effect of various samples is roughly in agreement with certain basic ideas on conductivity in oxidic semiconductors. From Hall effect measurements no conclusions could be drawn except that the apparent electron mobility is extremely small.

2158: E. J. W. Verwey: Oxyd-Systeme mit interessanten elektrischen und magnetischen Eigenschaften (Angew. Chemie **66**, 189-192, 1954, No. 7). (Oxidic systems with interesting electric and magnetic properties: in German.

Survey of properties of oxidic compound systems containing elements of the first transition period of the periodic system. Sintered ceramic materials with the NaCl, haematite and spinel lattice are described, which are of interest as semiconductors for the production of resistance materials and further provide magnetic materials of industrial importance.

2159: F. A. Kröger and H. J. Vink: The origin of the fluorescence in self-activated ZnS, CdS, and ZnO (J. Chem. Phys. **22**, 250-252, 1954).

It is proposed that the luminescent centre in "self-activated" ZnS consists of a cation vacancy whose nearest surroundings have lost one electron. Such a centre is consistent with the fact that at low firing temperatures, the appearance of the blue fluorescence of self-activated ZnS depends upon the presence of "promoter ions" (monovalent anions or trivalent cations) whereas, if the firing temperature be sufficiently high, some blue fluorescence is obtained without the presence of such promoter ions. The luminescence of reduced ZnS, CdS, and ZnO is also discussed, and is attributed to anion vacancies that have trapped one electron.

2160: S. Duinker: Magnetische versterkers (T. Ned. Radiogenootsch. **19**, 91-114, 1954, No. 2). (Magnetic amplifiers; in Dutch.)

A simplified graphical analysis of the principle of operation of magnetic amplifiers, based on an idealized *B-H* characteristic of the core material, under no-load conditions. Two fundamental types, the series-connected and the parallel-connected magnetic amplifier are considered. The factors determining the power-amplification, feedback, supply frequency, core-construction and bias are discussed.

A survey is given of the properties of magnetic amplifiers and the applications in various fields, which can be broadly separated into measurement, control, switching and a.c. amplification.