SINTERED GLASS

by E. G. DORGELO.

666.189.4

In the manufacture of certain articles in which many metal parts (*e.g.* leading-in wires) must be fused into glass close to each other, it is sometimes impossible, due to the too low fluidity of the molten glass, to force the drop of glass between the metal parts. In such a case glass in powder form can be used, and this can be introduced between the metal parts before fusing. The glass obtained after fusing, which is not completely clear, is called sintered glass, and contains many very small air bubbles. In this article various properties and possibilities of this glass are discussed.

In the manufacture of incandescent lamps, electronic and gas-discharge tubes different special kinds of glass are used, which must satisfy certain requirements for each type of valve or lamp, and therefore may differ very much from each other. A kind of glass whose properties and constructive possibilities are particularly favourable for one type may be quite unsuitable for another type.

For the development of a new lamp or valve, therefore, the possibilities offered by different kinds of glass should be subjected to an extensive investigation. If the existing kinds of glass are unsuitable for the application in view, an attempt is made to find a new and better kind of glass. In the case of different types of valves the progress of their development depended almost exclusively on the manufacture of a suitable new kind of glass.

One example is the sodium lamp covered on the inside with borate glass [1]), which is resistant to sodium vapour; further the high-pressure and super high-pressure mercury lamps, where it was necessary to find types of glass for the covering of the metal leads through the quartz [2]). In the development of transmitting valves for very short waves use was also, successfully made of special new kinds of glass, the electrolysis-free glasses [3]).

While in the examples mentioned here the new glasses are distinguished from the older ones by their chemical composition, in the case of "powder glass" an attempt has been made to create new possibilities by a modification in the physical structure. The stimulus for this attempt lay in a difficulty which occurs in the manufacture of articles in which many metal parts (for example leading-in wires) must be fused into glass close to each other. The liquid glass must then be forced between the metal parts, and the method fails when the spaces between are so small that a drop of molten glass, even under high external pressure, cannot penetrate sufficiently far into them, perhaps because of the fact that cooling takes place too rapidly as a consequence of the heat conduction through the metal parts. This difficulty can be overcome if, before the fusing in, the glass in fine powder form is introduced directly into the space where it belongs, the whole then being heated to a temperature at which the glass melts. The structure of the somewhat turbid sintered glass which is obtained upon fusing the powder is not homogeneous; it contains numerous very small gas or air bubbles which more or less modify the different properties of the glass. In general these changes are not of prime importance. Nevertheless, they may sometimes make possible constructions which are impossible with normal glass. Examples of such cases where preference is given to powder glass will be given later in this paper.

Sintered glass offers great advantages in the manufacture of valves and lamps for experimental purposes, due to the rapid and simple manner in which almost any desired lamp base can be made. Metal leads can be fused in at the same time that the base is made, while the process can be used for every kind of glass, including the kinds which are very difficult to soften.

The employment of sintered glass

The raw material, powdered glass, is obtained by grinding up pieces of glass. This powder is cast in a mould in which the metal parts to be fused in are already present. Care must be taken that the glass powder fills up the spaces well between the metal parts. After covering the mould the whole is heated to a temperature at which the glass is very fluid, so that only slight pressure is enough to fill even the smallest cavities.

The coefficient of expansion of the material of the mould must be adapted in a certain way to that of the glass. The wall of the mould must not clamp the solidified glass article; the coefficient

[1]) Philips techn. Rev. 2, 87, 1937.
[2]) Philips techn. Rev. 3, 119, 1938.
[3]) Philips techn. Rev. 6, 255, 1941.

of expansion of the material of the wall of the mould must therefore be smaller than that of the glass. The bottom of the mould (which is separate from the wall) must, on the other hand, have by preference the same coefficient of expansion as the glass, since otherwise there is a danger that, upon cooling, any leads and the like which are fastened into the bottom will be bent. A mould which possesses the properties mentioned is shown in *fig. 1*.

Besides lead-pins, differently shaped objects can also be fused in, for instance metal strips, tubes, nuts, etc. (*fig. 2*). The number of pins, the distances between them and their grouping is subject to practically no limitations when sintered

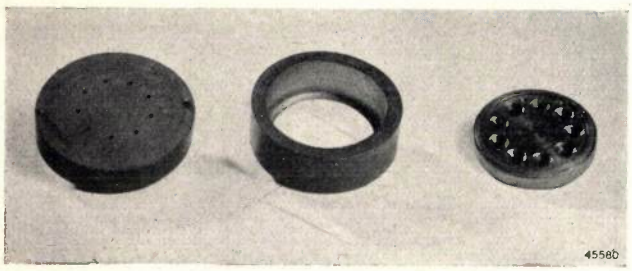


Fig. 1. Mould of metal with separate bottom and cover.

glass is employed. *Fig. 3* illustrates the great variety possible.

Furthermore it may be pointed out that simultaneously with the fusing-in of leads in a lamp or valve base the glass envelope can also be welded on. This envelope is then placed in the mould before the fusing of the powder glass base; the upper part of the mould must then be removed. In this way the separate welding process is eliminated. This process is, of course, subject to the restriction that it can only be applied in those cases where the electrodes inside the envelope are resistant to the heat radiation of the glowing mould. The tubes and bases shown in the photographs *fig. 4* are made in this way. It is obvious that other glass parts, such as an exhaust tube, can be welded in simultaneously with the fusing of the valve or lamp base

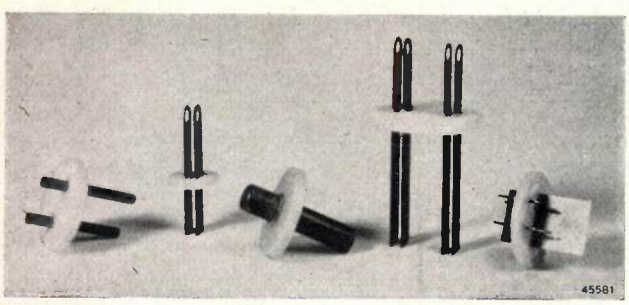


Fig. 2. Fused-in metal tubes and strips in powder-glass bases for valves.

(*fig. 5*). To prevent the exhaust tube from collapsing during the fusion, it is previously sealed at the bottom and filled with fine sand, which is shaken out after the fusion.

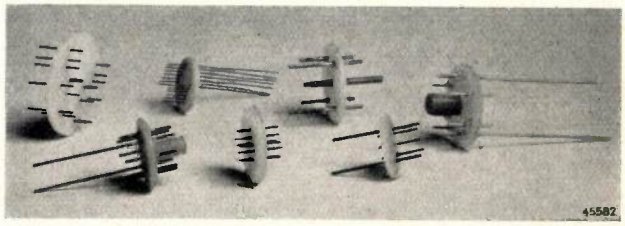
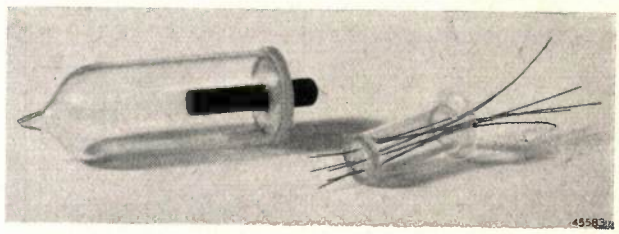


Fig. 3. The leads can be fused-in through sintered glass bases in almost any number and arrangement.

A special manner of fusing in, which is impossible with glass envelopes, can be applied to metal envelopes by strongly heating the edge of the metal envelope and then pressing it into the likewise previously heated sintered glass base. The projecting edge can be removed later. In this way it is possible, for example, to fasten lead-glass discs into an iron can. Because of the fact that the coefficient of expansion of the iron is larger than that of the glass, the iron upon cooling is clamped around



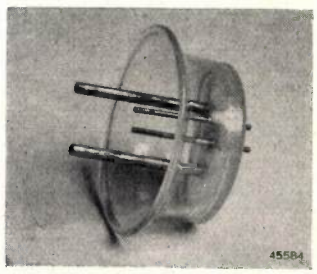


Fig. 4. Envelope and pinches, the powder-glass bases of which are welded to the envelopes at the same time that the powdered glass is fused.

the glass, giving a very reliable connection [4]). Nor is it of importance in this method whether the cross section of the envelope is a true circle.

Properties of powder glass

1) *Specific weight*

The circumstance that there is a very large number of small gas bubbles in sintered glass affects different properties; it is clear that for example the specific weight is smaller than that of

[4]) In principle this method of fusing-in is also possible with discs of clear glass.

the original glass. The decrease depends upon the size and the number of bubbles and is usually of the order of magnitude of 5 to 10 percent; with very fine powder the decrease is greater. The diameter of the bubbles usually lies between 10 and 50 μ. The number of bubbles per mm³ amounts to several thousands.

2) *Electrical properties*

When a block of powder glass is situated in an electric field the field strength in the glass will be much smaller than in the gas or air bubbles, as a result of the very different dielectric constants (glass: approx. 7, air: 1). The field is thus concentrated in the bubbles.

If we consider a piece of clear glass with only a few large air bubbles, the field concentration in these bubbles may be so high that the enclosed gas becomes ionized, which may lead to complete breakdown. Glass for high-voltage apparatus (X-ray tubes, for example) must thus satisfy the requirement of being free of bubbles to a high degree.

The situation becomes quite different when the air in the glass is divided into very many small bubbles. The potential difference between the boundary surface is then uniformly distributed over the numerous intermediate gas bubbles, so that the potential difference per gas bubble is so low that danger of ionization is out of the question. With powder glass indeed values of the breakdown voltage are found which are just as high as those measured on glass which is free of bubbles.

The electrical field in a medium in which there are globular gas bubbles can be calculated [5]). Different electrical constants of powder glass, such as the electrical conductivity λ', the dielectric constant ε' and the angle of loss, determined by tan δ', can be calculated from the corresponding values for the homogeneous glass (λ, ε and tan δ). Thus if

$$p = \frac{\text{volume of all air bubbles}}{\text{total volume}},$$

we find that:

$$\frac{\lambda'}{\lambda} = \frac{2-2p}{2+p} \approx 1 - \frac{3}{2}p + \ldots \quad (1)$$

$$\frac{\varepsilon'}{\varepsilon} = \frac{2\varepsilon+1-2p\,(\varepsilon-1)}{2\varepsilon+1+p\,(\varepsilon-1)},$$

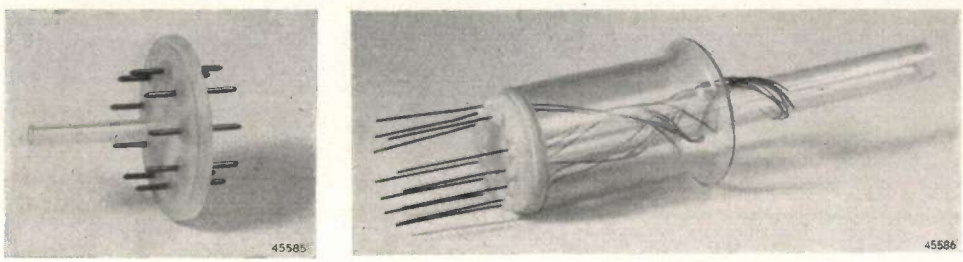


Fig. 5. Exhaust tubes can be welded in simultaneously with the fusing of the lamp or valve base.

which, *e.g.* when $\varepsilon = 7$, gives

$$\frac{\varepsilon'}{\varepsilon} = \frac{5-4p}{5+2p} \approx 1 - \frac{6}{5}p + \ldots \quad (2)$$

$$\frac{\tan\delta'}{\tan\delta} = \frac{2\,(5-3p)}{10-3\,p} \approx 1 - \frac{3}{10}p + \ldots \quad (3)$$

The formulae are valid only when $p \ll 1$. In order to obtain an impression of the influence which the air bubbles have on the constants mentioned we substitute $p = 0.1$. Then $\lambda'/\lambda = 0.85$, $\varepsilon'/\varepsilon = 0.88$ and $\text{tg}\,\delta'/\text{tg}\,\delta = 0.98$.

A further lowering of the values of ε' and λ' can thus be obtained by distributing much air among many small bubbles. For this purpose it is necessary to start with very fine powder and during the fusion the temperature must be raised very rapidly to prevent the escape or flowing together of the air bubbles. By the addition of substances which give off gas the percentage can be very much increased, which, however, involves a lowering of the strength of the glass.

3) *Thermal properties*

The heat conduction in sintered glass shows almost the same variation with p as the electrical conductivity (see formula 1). The heat conductivity of powder glass is thus somewhat less than that of

[5]) K. W. Wagner, Archiv für Elektrotechnik, 2, 382, 1914.

clear glass of the same composition. It is sometimes necessary to take special precautions in fusing because of this fact.

The coefficient of expansion of powder glass is the same as that of normal glass; the coefficient of expansion is not changed by the presence of air bubbles, which is a result of the well-known fact that a hollow body expands as if it were massive.

4) *Tensions*

Objects built up of more than one different material with different coefficients of expansion are not in general free of mechanical tensions at every temperature. This holds also for the welding together of two kinds of glass and for the introduction of a metal lead through glass. The occurrence of tensions in a glass object often leads to breakage and therefore methods have been developed for the checking of this. The tensions make the glass optically anisotropic and give rise to phenomena of double refraction, which can for instance be made visible with a polarization apparatus.

In this investigation of tensions it has now been found to the advantage of sintered glass that in objects manufactured with the help of this glass fewer mechanical tensions occur than in ordinary clear glass. This can be demonstrated by fusing together in pairs discs of different kinds of glass. When the discs fused together consist of powder glass smaller tensions appear after cooling than when the two discs are composed of the corresponding kinds of clear glass.

The explanation of this phenomenon is probably as follows. Since the solidification is accompanied by a decrease in volume, tensions will appear in the glass which are smaller the better the still soft glass is able to accommodate itself.

Now this is more easily possible in sintered glass than in clear glass of the same composition. If there is a large number of gas bubbles in the glass, as long as the surrounding glass is still somewhat soft they can compensate the volume decrease of the solidifying glass by expanding. This conception is thus based upon the fact that some parts of the piece of work become solid before the rest. This phenomenon often plays a part when metal components are fused in.

Completely homogeneous objects can also be made free of tension when normal glass is used by taking care that they are cooled very slowly. One of the advantages of sintered glass is that there is no objection to the cooling taking place more rapidly.

Applications

One of the chief applications of powder glass is in lamps and tubes for experimental purposes. However, it also offers a number of advantages for other applications which in many cases are important. In the following we shall mention several of them.

1) *In tubes with complicated electrode systems*

As already mentioned, the glass can be introduced into the mould in the form of finely divided powder so that it can easily fill all the small cavities. Complicated pieces of work, such as lamp bases with very many leading-in wires or with leads very close to each other can be made of powder glass without much trouble (see fig. 3). The freedom of choice in distance between the leads makes it possible to place them in such positions that the simplest and most logical assembly of the electrode

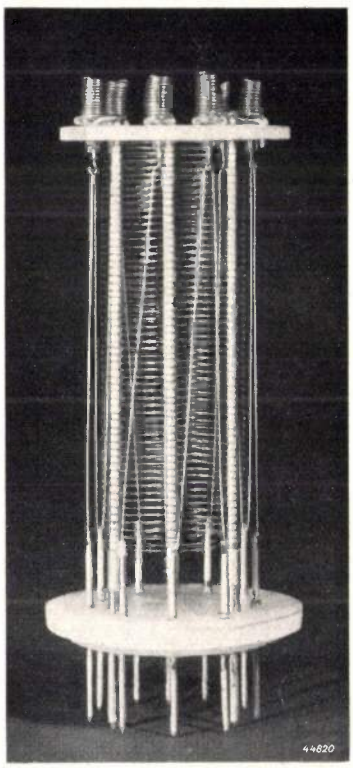


Fig. 6. Electrodes of a transmitting valve for short waves assembled directly on the powder-glass base.

system is achieved. Thus in *fig. 6* may be seen the grid-cathode part of a short-wave transmitting valve, in which the two electrodes are welded directly to the leads. These leads form two concentric circles with diameters corresponding to those of the cathode and grid. Such a construction excels in simplicity and sturdiness, while the slightness

of the self-induction of the lead elements makes possible a satisfactory functioning on short waves. In extreme cases, in order to keep the selfinduction as low as possible, it is possible to lead the electrode itself through the glass. Three examples are shown in *fig.* 7 (see also fig. 2).

In the left-hand valve there are two electrodes bent in a U-shape which pass through the glass base while retaining their cross section. The self-induction of the lead part is very small here. At the same time good cooling is hereby obtained. In the case of the middle valve a shielding plate separating the two parts of the valve is led through the base.

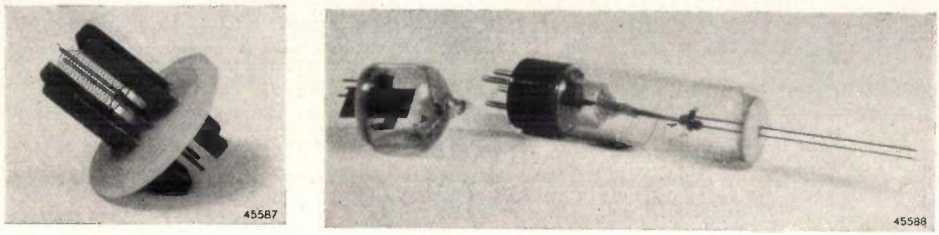


Fig. 7. Several transmitting valves for very short waves in which the electrodes themselves are fused into the base.

The right hand valve of fig. 7 contains two leads in the form of strips which together form a Lecher system. Experience has shown that with clear glass and the ordinary pressing technique such constructions are only made reliable with difficulty. It is difficult to obtain a satisfactory distribution of the drops of glass. At the edges of the metal also the tensions are often too high, so that the valve cracks at that spot.

In the case of powder glass the first difficulty is non-existent, while less hindrance is experienced from the second.

2) *Crowded constructions*

Such constructions often result in high temperatures of the glass wall. It is then necessary to have recourse to glass with a high softening point. Normal boro-silicate glasses with a softening point of 500-600° C are often too soft or are too poor insulators at the high operating temperatures, so that harder glasses must be used. As a rule these glasses cannot be pressed into narrow interstices in the ordinary way. Before the necessary pressure is reached the drop of glass has cooled off too far. The pressing in of metal parts directly is even more difficult. Due to the fact that pressing is unnecessary in powder technique, high requirements need not be made of the mould, and it can be heated to such a high temperature that even a very hard glass still becomes sufficiently fluid. It has hereby become possible to make lamp bases of kinds of glass, such as the so-called electrolysis-free glass, which could not formerly be so used and this again has opened the possibility of constructing new types of valves, especially in the field of transmitter valves for decimetre waves.

3) *Connecting seals*

It is possible to fill the mould with layers of powdered glass having different properties, and then to fuse the whole. By choosing powdered glasses of gradually increasing coefficients of expansion graded seals can be made, *i.e.* tubular parts which at one end can be fused to a glass with a high coefficient of expansion and at the other to a glass with a low coefficient of expansion.

Another possibility offered is that the base of a valve which is to contain the vapour of an alkali metal can be protected by a thin layer of resistant glass. It is well known that most kinds of glass are very severely attacked by such vapours. In the case of glasses containing lead for example such a strong reduction takes place that the glass turns black, due to the liberated lead. Now by first scattering powdered lead glass in the mould and over it a thin layer of borate glass, a protecting layer is formed. Borate glass is not attacked; it cannot, however, be used alone, since the temperature interval in which softening takes place is particularly short. A variant on the foregoing is the use of glasses of different colour, by which means all kinds of indications can be introduced on the object.

4) *Rapid fusing-in and cooling*

Notwithstanding the fact that the conduction of heat in sintered glass is smaller than in the corresponding clear glass, it appears in heating

that with sintered glass without many precautions a more equal heating through the whole object is obtained. Probably the numerous gas bubbles present in sintered glass ensure that during the heating the radiation of heat is strongly dissipated. One result of this is that the so-called pre-heating which precedes the fusing of a lamp base to the envelope can take place quite rapidly without cracking the tube. It is clear that this is of great importance in mass production. The annealing also can in most cases proceed quite rapidly due to the previously mentioned great elasticity of the still not completely solidified powder glass, so that the occurrence of large tensions is combated.

---

50 YEARS X-RAYS

In November 1895 Wilhelm Conrad Röntgen, professor at the university of Würzburg, made his first observations on X-rays; his publication is dated December 27th 1895 and entitled: "On a new Kind of Rays", and it was a communication to the "Würzburger medizinisch-physikalischen Gesellschaft" (1895, p. 137—141).

Now, 50 years afterwards, the whole world is going to commemorate this important contribution to our present day scientific insight — and to our medical and physical instruments.

To look backward on the road achieved during these 50 years is indeed worth while. There the interaction is reflected which took place between the various developing branches of science and technique, and it is possible to find there how an amazing quantity of work has given life to a series of most important practical applications of X-rays. These are too well known to be summed up here, but they engendered, too, a great array of very elaborate apparatus. The Philips' Factories and Laboratories have also had their part in this development. Here it may suffice to refer to the 38 publications which appeared in the first seven years of this periodical on the subject of X-rays and their applications: 9 of these articles dealt with X-ray tubes and X-ray apparatus; 21 on applications and 8 on the methods used in these applications. In this number, too, with which the Philips Technical Review re-enters the world, after the forced interruption during the German occupation, the reader will find a description of an X-ray apparatus which in many respects is representative of the ideas and methods which have developed in the domain of X-rays.

In the 1946 volume of this journal we hope to find from our X-ray Department still more contributions on the development of special tubes, apparatus and methods of research. Thus the great discovery of W. C. Röntgen 50 years ago will be given worthy commemoration.

that with sintered glass without many precautions a more equal heating through the whole object is obtained. Probably the numerous gas bubbles present in sintered glass ensure that during the heating the radiation of heat is strongly dissipated. One result of this is that the so-called pre-heating which precedes the fusing of a lamp base to the envelope can take place quite rapidly without cracking the tube. It is clear that this is of great importance in mass production. The annealing also can in most cases proceed quite rapidly due to the previously mentioned great elasticity of the still not completely solidified powder glass, so that the occurrence of large tensions is combated.

---

50 YEARS X-RAYS

In November 1895 Wilhelm Conrad Röntgen, professor at the university of Würzburg, made his first observations on X-rays; his publication is dated December 27th 1895 and entitled: "On a new Kind of Rays", and it was a communication to the "Würzburger medizinisch-physikalischen Gesellschaft" (1895, p. 137—141).

Now, 50 years afterwards, the whole world is going to commemorate this important contribution to our present day scientific insight — and to our medical and physical instruments.

To look backward on the road achieved during these 50 years is indeed worth while. There the interaction is reflected which took place between the various developing branches of science and technique, and it is possible to find there how an amazing quantity of work has given life to a series of most important practical applications of X-rays. These are too well known to be summed up here, but they engendered, too, a great array of very elaborate apparatus. The Philips' Factories and Laboratories have also had their part in this development. Here it may suffice to refer to the 38 publications which appeared in the first seven years of this periodical on the subject of X-rays and their applications: 9 of these articles dealt with X-ray tubes and X-ray apparatus; 21 on applications and 8 on the methods used in these applications. In this number, too, with which the Philips Technical Review re-enters the world, after the forced interruption during the German occupation, the reader will find a description of an X-ray apparatus which in many respects is representative of the ideas and methods which have developed in the domain of X-rays.

In the 1946 volume of this journal we hope to find from our X-ray Department still more contributions on the development of special tubes, apparatus and methods of research. Thus the great discovery of W. C. Röntgen 50 years ago will be given worthy commemoration.

AN X-RAY APPARATUS FOR CONTACT THERAPY

by H. A. G. HAZEU, J. M. LEDEBOER and J. H. v. d. TUUK. 261.386.1 : 615.849

In the X-ray treatment of tumours on the surface of the skin it is desirable, in order to spare the underlying healthy tissue, that the radiation intensity should decrease rapidly with increasing depth below the skin. In order to realize this it is necessary that the distance between the source of X-rays and the skin should not be too large (often there exists immediate contact between X-ray tube and the skin, hence the name contact therapy) and the X-ray tube should possess only a slight "own filter". In order to satisfy these and other requirements connected with medical practice, the tube of the Philips contact therapy apparatus is so constructed that the X-radiation leaves the tube through an opening in the earthed cathode. This construction is described in detail in the following paper; it permits an irradiation from a distance of 18 mm of the focus with a filter equivalent to only 0.2 mm of aluminium. The tube is fed with 50 kv DC voltage at a current of 2 mA and possesses forced air cooling. The X-ray intensity on the skin is so high that an irradiation time of a few minutes is usually sufficient. This type of therapy is thereby made accessible to a much wider circle of patients than is possible with radium treatment.

Depth therapy and surface therapy

The treatment of tumours with radium or X-rays depends upon the fact that the affected tissue is attacked by these rays and if a sufficient dose is administered this tissue will die off. However, the healthy tissue around the tumour is, also more or less exposed to the X-rays and attacked by them. If this impairment is too severe, the healing of the diseased tissue can be retarded or even prevented. The aim of the doctor, therefore, must be to make the ratio of the radiation dosage on the healthy tissue to that on the diseased tissue as small as possible. The measures to be taken for this purpose are quite different according as the diseased tissue lies deep below the surface of the skin or is situated only slightly below or on the surface of the skin.

Let us first consider the first case of "depth therapy". Since the intensity of the X-rays decreases with the square of the distance from their source, and, moreover, since the rays suffer an attenuation in the tissue, the intensity will be smaller at some depth than on the skin. The dosage has a certain depth gradient, see *fig. 1a*. In depth therapy, therefore, the diseased tissue always receives a smaller dose than the healthy one at the surface. The fact that a healing effect can nevertheless be obtained is due to the fact that the diseased tissue is sometimes more severely attacked than the healthy tissue by the same dose. In fig. 1a the relative positions are indicated of the dose required for killing the diseased tissue and the dose permissible with the object of sparing the healthy tissue. It may be seen that with these relative positions therapy is possible, but only when the radiation does not have too great a depth gradient. In depth therapy, therefore, the smallest possible depth gradient of the radiation is desirable.

The situation is quite different in the case of tumours on the surface of the skin, *i.e.* in "surface therapy". In this case the healthy tissue, which will also be exposed upon irradiation, lies for the most part under the diseased tissue, the positions being thus reversed. This case is naturally more favourable for radiation therapy, since here the diseased tissue always receives a larger dose than the healthy

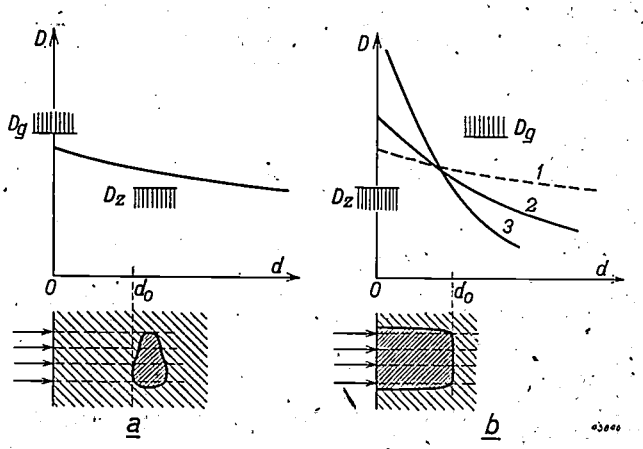


Fig. 1*a*) In depth therapy the dose of X-rays administered on the surface of the skin where the tissue is healthy may not exceed a value D_g, while at a depth d_o under the skin, where the diseased tissue is situated, it must attain at least a value D_z. The smallest possible depth gradient of X-ray intensity is required.

b) In surface therapy a minimum value D_z is required on the surface and a maximum value D_g at the depth d_o. Here a steep depth gradient is favourable, for instance one according to curves *2* or *3*, which are much steeper than the broken line curve *1* representing the curve of fig. 1*a*.

tissue. If we now indicate the required minimum dose on the diseased and the maximum permissible dose on the healthy tissue (*fig. 1b*) it is clear that in principle any depth gradient of the dose can be used. In order, however, to spare the healthy tissue as much as possible, it is clear that a fairly steep depth gradient is preferable, while one

also has the possibility of administering a larger dose on the diseased tissue.

How is it now possible to realize the desired slight or steep depth gradient, as the case may be? The gradient is characterized by comparing the intensity of radiation at the surface of the skin I_h with the intensity I_d at a depth d under the skin. If a is the distance from the source of X-rays (focus of the tube) to the skin and μ the coefficient of attenuation of the tissue for the rays, the square law gives

$$\frac{I_d}{I_h} = \frac{a^2}{(a+d)^2} e^{-\mu d} \quad \ldots\ldots \quad (1)$$

This "depth quotient", which is a measure of the depth gradient, becomes large (*i.e.* approaches the value unity) when a is large compared with d and if μ is small. In depth therapy, therefore, the X-ray tube will be set up at some distance from the patient (30 to 100 cm) and hard radiation (short wave length) will be used, as it is only slightly attenuated in the tissue. This means high tube voltages [1]) — practically usually 200 kV, sometimes up to 1000 kV — and the employment of a heavy metal filter to suppress the soft parts of the mixture of radiation emitted from the focus. In order to obtain the desired dose in the tumour without too long an exposure time in spite of the great distance and the loss of radiation in the heavy filter, a high tube power (1-4 kV) is needed.

In the case of surface therapy, where the object is a low depth quotient I_d/I_h, exactly the opposite measures mus tbe taken: the distance a between focus and skin will be made as small as possible and radiation will be used which is subject to great attenuation in the tissue, thus soft radiation (low tube voltage, approximately 50 kV) without more filtering than is inevitable due to the passage of the radiation through the wall of the tube (so-called own filter of the tube).

Apart from the fundamentally more favourable situation in surface therapy compared with depth therapy, there is also the extra advantage that, thanks to the small distance from focus to object and the weak filtering, only a low tube power is necessary for the required dosage. The X-ray tube may therefore be small and easily adjustable, the whole apparatus may, also because of the ralative low tube voltage, be light, even portable, while in addition very short exposure times are sufficient.

These favourable aspects of surface therapy, or "contact therapy" as it is often called, the X-ray tube being in immediate contact with the patient's skin, have contributed much to the adoption of the X-ray treatment of skin diseases and surface tumours.

In the following description of the Philips CT apparatus, which has been specially developed for this therapy (CT), we shall enter into more detail about some of the aspects of this therapy [4]).

Construction of the X-ray tube

From the above it follows that an X-ray tube for contact therapy must fulfil the following requirements:

1) The distance from the focus to the window must be very short in order to make it possible to place the focus close to the skin.
2) The filter of the tube must be small, which means that the window of the tube at the spot where the rays pass through the tube wall must be very thin and made of a light material (low absorption).

In addition to these there are several practical requirements which emerge from the desire to be able to apply contact therapy to tumours and diseases of the mucous membrane in cavities of the body, such as the mouth, throat, etc. For this purpose the focus should be close to the end of the tube and this end should be small enough to be introduced into such cavities. At the same time it is usually desirable that the radiation should be emitted in a forward direction (not lateral, as is customary).

The requirement of a thin tube with the focus at the end also holds in the employment of X-rays for the testing of material, when tubular casts and the like are examined. For this purpose X-ray tubes have been constructed with a hollow, earthed anode projecting from the tube and shaped like a funnel [2]), see *fig. 2*. For our purpose, however, this construction has the disadvantage that the second requirement mentioned, small own filter, cannot easily be satisfied. The rays, which are emitted in the direction of the length of the tube, must pass through the anode plate upon which the focus (lozenge) is situated. For the sake of high efficiency the lozenge must consist of a heavy metal, usually tungsten; this metal, however, also has a high absorption, especially for the soft rays. Moreover,

[1]) See: Philips techn. Rev. 4, 161, 1939. It should also be noted that until now it has not been determined whether the healing effect for tumours is fundamentally different for X-rays of different wavelengths corresponding to tube voltages between 50 and 1000 kV. The choice of wave length can therefore, apart from technical considerations, be determined primarily by the desired depth gradient.

[2]) *Cf:* Philips techn. Rev. 5, 69, 1940 (fig. 6). For a contact therapy tube of this kind see H. W. Ernst, K. Frik and P. Ott, Strahlentherapie 52, 369, 1935.

the anode plate must not be very thin, since it must be resistant to the strong heat transfer by the current of electrons and must also be vacuum tight. In this way a heavy own filter is obtained and in consequence a much harder and less intense radiation than is desirable.

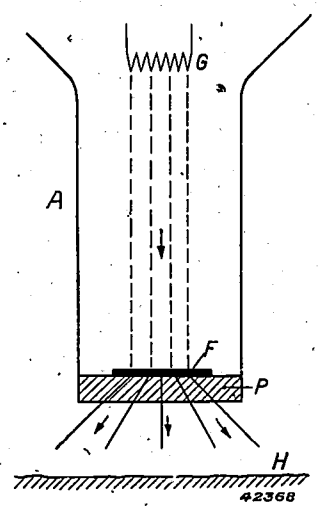


Fig. 2. Construction of an X-ray tube with earthed anode projection. The electrons emitted from the filament G move through the earthed hollow anode A and impinge upon the copper plate P. The upper layer of this plate, the lozenge, upon which the focus F is formed, consists of tungsten. The X-rays emitted by P must pass through the anode plate P to reach the surface to be irradiated H.

Of course these two objections can be partially compensated by making the distance between focus and skin very small. When the distance is small enough (several mm) it is possible to work with both hard and soft radiations, as is indeed the case with radium irradiation, since then, due to the dominance of the factor $a^2/(a+d)^2$ in equation (1), a very steep depth gradient is nevertheless obtained. Working with such small distances has in turn, however, other objections. In the first place very small variations in the distance then have immediately a great influence on the intensity on the skin, so that dosage is made difficult. Especially, however, at such small distances it is necessary to use very large angles of divergence of the beam of the X-rays. The peripheral rays of the beam then pass through the anode plate obliquely, as may be seen in fig. 2, and are thus more attenuated than the rays along the axis; and, moreover, they must cover a greater distance to reach the skin. The result is a very non-uniform distribution of the intensity on the field irradiated, a phenomenon which meets strong objections from the doctor. The doctor, on the contrary, for the sake of easy and reliable dosage, requires the most uniform distribution possible.

In this laboratory an entirely different construction has been worked out, in which the own filter of the tube could be very much restricted [13]). The cathode is here earthed and the X-rays emitted by the anode pass through the opening in the ring-shaped anode.

In *fig. 3*, which shows a diagram of the end of the tube, this construction may be seen. The filament G, led through with one pole connected and the other insulated, is fastened in the earthed metal cathode can K. By means of the ring D the electrons emitted from the filament are focussed on the massive tungsten anode, which has a positive voltage of 50 kV with respect to the cathode. The X-rays excited on the anode pass out of the tube through a glass window behind the filament and fused into the cathode can. It is of importance here that practically cathode potential prevails over the whole space occupied by the window. There is thus no danger that secondary electrons freed on the anode will bombard the window, which therefore may be large and thin.

Surrounding the tube is an earthed metal jacket O, which is closed at the spot where the rays emerge by a thin cap of "Philite" in order to provide mechanical protection of the thin window. This cap can be placed directly against the skin surface to be treated, so that due to the very short distance between focus and skin a very steep depth gradient is obtained. The minimum distance between focus and skin amounts to 18 mm, while the X-rays need pass through no other filter than the glass window and the "Philite" cap; the combined filter

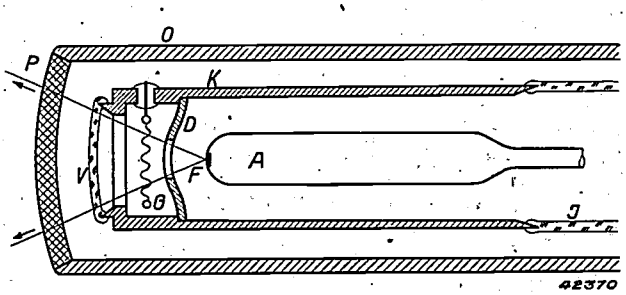


Fig. 3. Construction of the X-ray tube of the Philips *CT* apparatus for contact therapy. K earthed cathode can, G filament, D focussing ring, A anode, F focus, V window, I insulated connection between cathode can and anode holder, O jacket, P "Philite" cap.

effect of these is equivalent to 0.2 mm of aluminium. For the sake of comparison it may be mentioned that the own filter of other tubes for the same voltage is at least 1 mm of aluminium.

[3]) *Cf.* J. H. van der Tuuk and G. J. van der Plaats, Ned. T. Geneesk. **79**, 4025, 1935 and J. H. van der Tuuk, Ned. T. Natuurk. **3**, 129, 1936.

In *fig. 4* the variation of the intensity as a function of the depth beneath the skin is shown as measured [4]) upon irradiation with the tube described. It may be seen that, using the smallest possible distance (curve *1* for a distance of 2 cm), a very steep depth gradient is obtained: at a depth of 1 cm under the skin the intensity has already fallen to 22 percent of that on the surface.

It must be remarked that there is still some difference of opinion in medical circles as to what is practically the most suitable value for the depth gradient [5]). It will indeed depend more or less on the individual case. In some cases, therefore, a less steep gradient will be preferred. This can be obtained in a simple way by slightly increasing the distance from focus to skin or by introducing an extra filter for the rays. Curves *2* to *6* in fig. 4 show the different forms of intensity gradient under the skin which can be realized by these means. In the case of curve *6* (4 cm distance and filter of 2.7 mm of aluminium) the intensity at 1 cm depth still amounts to 49 percent of that on the surface.

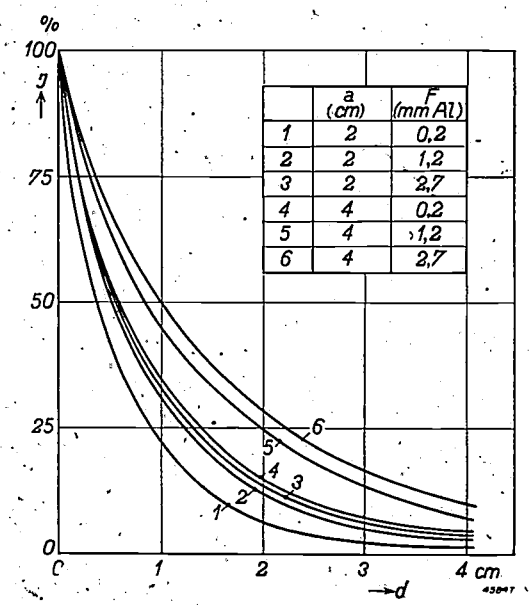


Fig. 4. Measured depth gradient of the tube of fig. 3 with 50 kV voltage. Intensity I in percent as a function of the depth d in cm.

The field distribution is very homogeneous with this tube, as is shown in *fig. 5*, where so-called isodose curves are drawn. These are lines connecting all points receiving the same dose. The field irradiated is here limited by a metal cap to a circle of 10 mm diameter. The dose over the whole skin surface exposed is practically uniform. At the same time the rapid decrease in the dose with depth and the sharp bounding at the sides may be seen from the figure. This bounding, also an important requirement of the doctor, is obtained by using a small focus. With a large focus only the field on the surface of the skin would be sharply limited by the metal cap, but the edge of the cap would cast a half-shadow, owing to which at some depth the lateral limitation of the field would no longer be sharp.

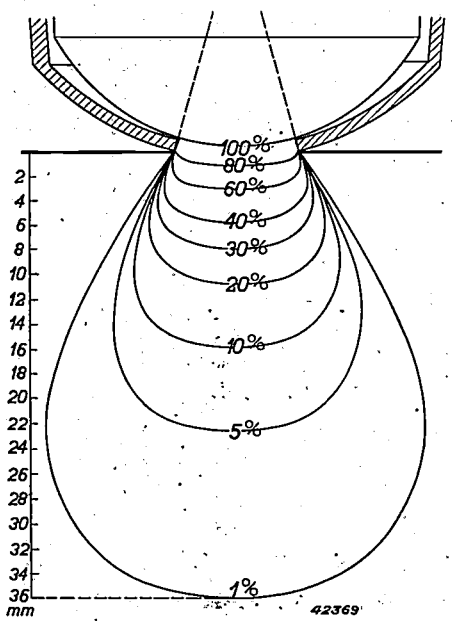


Fig. 5. Isodose curves obtained upon irradiation with the tube of fig. 3. The field irradiated on the skin is limited to a diameter of 10 mm by means of a metal cone with an opening.

The jacket of the tube

In *fig. 6* a sketch is shown of a simplified cross section through the complete tube with jacket. (As may clearly be seen from the photographs of figs. *7*, *8* and *9* the tube is relatively still much thinner). The anode A on a long stem is supported by an anode holder H, which is insulated from the cathode K by a glass joint I. Between the tube and the earthed metal jacket O, which completely protects doctor and patient from the high voltage, a conical insulating tube of ebonite E is introduced. This made it possible to limit the diameter of the whole to about 50 mm (in the middle). The external diameter 2 r_2 of the ebonite tube near the anode holder is 48 mm, whilst the anode holder itself has a

[4]) *Cf.* for extensive results: G. J. van der Plaats, diss. Utrecht 1938. Compare too L. F. Lamerton, Brit. J. Rad. **13**, 136, 1940.

[5]) *Cf.* for example: W. Schäfer and E. Witte, Strahlentherapie **33**, 578, 1929; W. Chaoul and A. Adam, Strahlentherapie **48**, 31, 1933; G. J. van der Plaats, Theses, Utrecht, 1938; cf. also footnote 6); D. den Hoed, Acta Rad. **19**, 239, 1938 J. M. Woodburn-Morison, Medical World **12**, 231, 1933, P. A. Flood and D. W. Smithers, Brit. J. Rad. **12**, 462, 1939. Sven Hultberg, Acta Rad. **24**, 328, 1943.

diameter $2r_1 = 22$ mm and is at a voltage V of 50 kV. The greatest field strength in the ebonite (namely on the inner wall) is then [6])

$$E = \frac{V}{2.3\, r_1 \cdot \lg r_2/r_1} = 58 \text{ kV/cm}, \ldots . \quad (2)$$

which is still far below the maximum permissible field strength in ebonite (150 kV/cm). Without the ebonite the diameter of the jacket would, as

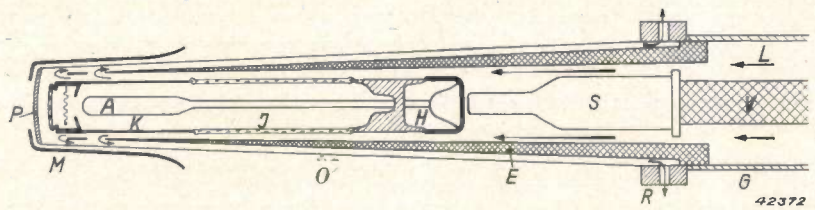


Fig. 6. Cross-section of the X-ray tube with jacket. From the photographs of figs. 7, 8 and 9 it may clearly be seen that the tube is actually very much thinner. K cathode can, A anode, H anode holder, I glass joint, O jacket with "Philite" cap P, E insulating tube of ebonite, V high-voltage cable, S plug, G rubber tube, L air inlet, R ring with holes for escape of air, M metal cone for accurate placing of tube on the skin.

can be calculated with formula (2), have to amount to at least $2r_2 = 112$ mm, in order to limit the field strength on the anode holder to the maximum permissible value for air of 28 kV/cm.

During use a power of about 100 W is dissipated on the heated anode (tube voltage = 50 kV DC voltage, tube current = 2 mA). The anode gives off the heat developed by radiation to the cathode can. If this can had in turn to get rid of the heat by radiation towards the outside, it, and with it the whole extremity of the tube, would become much too hot, so that it would not be possible to place the tube directly against the surface of the skin or in a cavity of the body. It was therefore necessary to provide an intensive cooling. For the sake of simplicity in construction and ease in manipulation of the tube, air cooling was chosen. In the cabinet for the high-voltage generator a fan is mounted. A rubber tube surrounds the cable, which supplies the high voltage to the tube, and the fan blows air through the space between cable and rubber tube. As may be seen in fig. 6, the air flows between the tube and the ebonite insulation can, along the cathode can and back along the outside of the insulator can, to pass through holes in the ring to the outside. In this way the patient experiences no inconvenience from the current of air. The amount of air blown through the tube amounts to 2.4 liters per second. Since air has a heat capacity of 0.285 cal. per litre and per degree centigrade and the air passing through when the tube is in continuous use takes up 100 W sec = 24 cal, the temperature of the air rises $24/(2.4 \times 0.285) = 35°$ above room temperature. The extremity of the tube does not therefore, become uncomfortably hot even after long use.

The method of cooling just described makes an air gap necessary between the anode holder and the ebonite insulator. Due to the fact that ebonite has a dielectric constant of 3, the field strength in this air gap is three times as high as in the adjacent ebonite [7]), thus a field strength which, according to the above mentioned values, is certainly many times greater than the breakdown strength in air. It may therefore appear remarkable that in this case no account is taken of the customary requirement that the breakdown field strength may not be exceeded at any point. Thanks to the presence of the ebonite no complete breakdown can occur; moreover, the ionized air is continually replaced by the cooling air current.

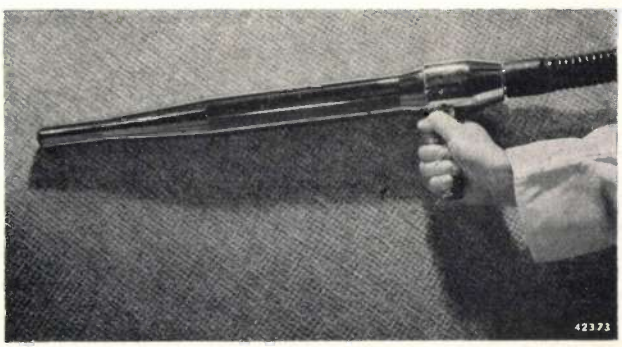


Fig. 7. The tube in its jacket. The whole is 50 cm long and weighs 2.5 kg; it can easily be held in one hand by the handle.

Fig. 7 is a photograph of the tube in its jacket. The tube can easily be held with one hand, so that the treatment can also be by hand (*fig. 8*). This method has the advantage that during the

[6]) See for example: Philips techn. Rev. **6**, 270, 1941. We assume that we are here concerned with the case of two concentric cylinders with a homogeneous dielectric, and we shall later deal with more complex dielectrics.

[7]) See for instance the article cited in 6).

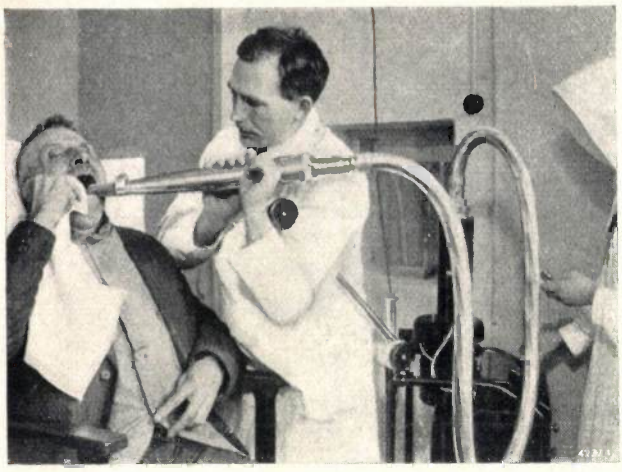


Fig. 8. The apparatus for contact therapy during treatment.

irradiation the tube follows the slight movements of the patient. The placing and adjusting of the tube is usually as follows; A metal cone (*M* in fig. 6) with an opening which is determined by the part to be irradiated is placed upon the spot to be treated. When the opening is exactly over the desired spot the doctor inserts the tube into the cone, which is held in position by the pressure of the tube, and the irradiation can be begun.

Instead of holding the tube in the hand it can also be mounted in a universally moveable arm fastened to the high-voltage generator, see *fig. 9*.

The tube supply

The tube is fed with direct current voltage, according to a scheme the principle of which is shown in *fig. 10*. It is desirable to use DC voltage of opposing the high-voltage in the negative phase is now, as it were, passed on from the X-ray tube to the rectifier valve. The construction of the jacket with the glow discharges in the air gap might also be dangerous with AC voltage [9]).

For the avoidance of flashover a pulsating DC voltage would, in principle, also be sufficient, such as is obtained with the well known and often used Villard connections, see *fig. 11a*. As may be seen upon comparison with *fig. 11b*, in the Villard connections transformer, condenser and valve are subjected to only half as high voltages as in the case of a connection for slightly rippled DC voltage with the same peak voltage. The latter,

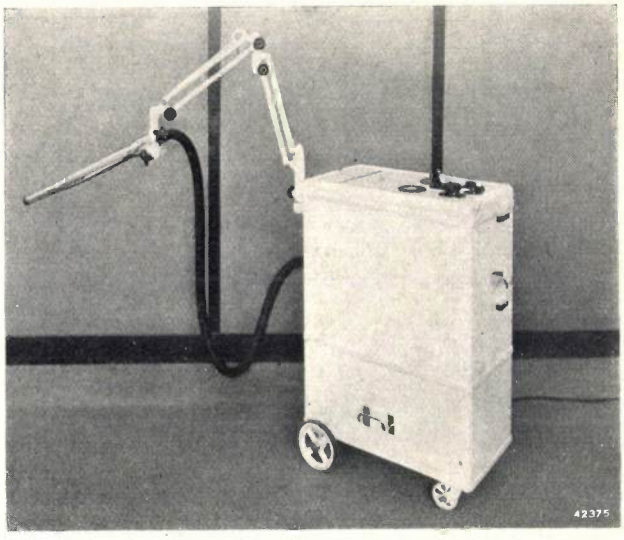


Fig. 9. The cabinet, containing the high-voltage generator and, above it, all the control and regulation elements, bears a universally moveable arm into which the X-ray tube can be fastened.

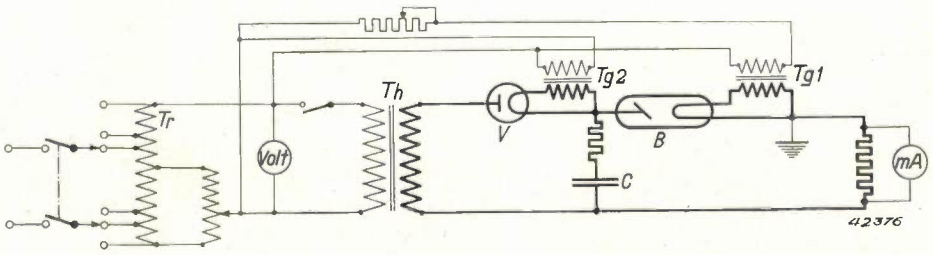


Fig. 10. Diagram showing the principle of the feeding of the tube. The high-voltage part is indicated by heavy lines. *B* X-ray tube, T_h high-voltage transformer, *V* rectifier valve, *C* tondenser, T_{g1} and T_{g2} filament current transformers, *Tr*. regulator transformer for correccion of mains voltage.

especially because of the danger of flashover as a result of electron emission caused by the heating of the hot anode when the voltage across the tube is reversed [8]). By the rectification the function however, offers the advantage that all the electrons attain the maximum velocity. With equal power, thus at a given capacity of the cooling, this is

[8]) See: Philips techn. Rev. **6**, 309, 1941.

[9]) Compare the analogous situation in the case of gas-filled cavities in the dielectric of paper condensers: Philips techn. Rev. **4**, 254, 1939.

manifested directly in a shortening of the necessary times of irradiation. In addition there is also the fact that, thanks to the low power required and the relatively low voltage, the economy possible with the Villard connections is not of much importance here. The whole generator, which is housed in the cabinet shown in fig. 9, weighs only 20 kg.

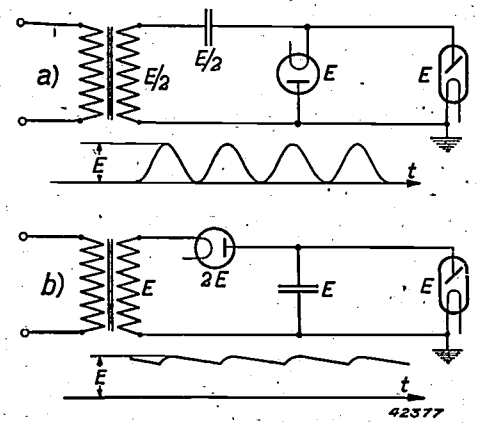


Fig. 11*a*) Villard connections.
b) Connections for obtaining much less strongly pulsating *DC* voltage, At the same peak value *E* of the voltage obtained, all the elements of the connections *a* can be constructed for a voltage only half as high as those of *b*.

Above the generator in the same cabinet are also housed all the switching and regulatory elements. Compared with various installations previously described, these include only one interesting feature which we shall discuss in some detail, namely an automatic regulation of the tube current. For satisfactory dosing it is necessary that the previously chosen tube current should be accurately maintained from the beginning to the end of a treatment. Now there are various circumstances which may have an unfavourable effect on the constancy of the tube current. In the first place there are always certain fluctuations of the mains voltage and consequently of the filament current of the tube. The electron emission depends so closely on the filament temperature that a mains voltage variation results in a ten percent greater variation of the tube current. This difficulty could be met by employing a stabilizer for the filament current, as is for example done in certain X-ray apparatus for diagnosis, where constant tube currents are likewise required [10]). In our case this method does not lead to the desired result because the relation between the filament current and the electron emission is not fixed to the same degree as in the apparatus for diagnosis mentioned above. This is due chiefly to the small distance between anode and filament: after switching on, the anode gradually heats up and heats the cathode by radiation, so that the emission increases even when the filament current is constant. The resulting increase in the tube current during the first few minutes after switching on is particularly unpleasant, since during the treatment, which should proceed smoothly and be finished within a few minutes (sometimes even within 10 sec.), instead of being able to devote his whole attention to the patient, the doctor would be compelled continually to regulate the tube current.

Now in order to eliminate all the causes of tube current variations simultaneously, a filament voltage regulator is employed which is governed by the tube current itself. *Fig. 12* shows the principle of the regulator. The filament current of the X-ray tube B is supplied by the transformer T_1, which in turn is fed by the transformer T_2 over the resistance R_1. In this resistance an extra voltage loss is caused by the current in the relay tube L. An increase in the average current through L will therefore cause a decrease in the average filament current. Now when we assume the potentiometer P to be in the position indicated by the dotted line, the voltage over the resistance R_3 acts on the grid of L. This voltage, which is smoothed by the condenser C_3, is proportional to the current through the X-ray tube. When the tube current increases the grid voltage of L becomes more positive, the moment of ignition of this tube in each period of

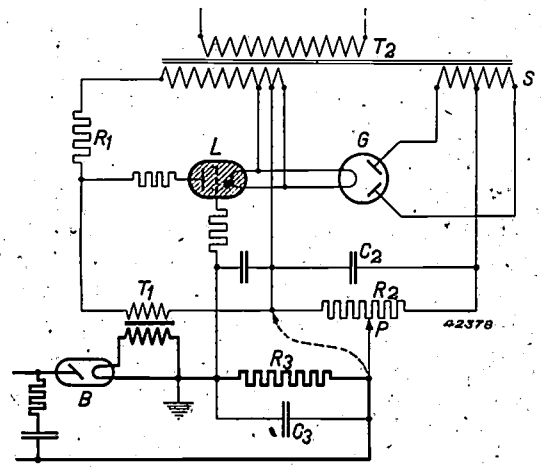


Fig. 12. Simplified scheme of the regulator for automatically keeping the tube current constant. See the description in the text of the article.

the AC voltage supplied by T_2 is advanced, the average current through the relay tube rises and the result is a fall in the filament current of the X-ray tube, so that the increase of the tube current is compensated. The regulatory mechanism thus tends to maintain the nominal tube current.

[10]) Philips techn. Rev. 6, 12, 1941.

The nominal tube current can be determined with the potentiometer P. With this potentiometer a variable part of the fixed DC voltage over R_2 (obtained by rectification of the AC voltage acting on S by the rectifier valve G and smoothing by C_2) can be subtracted from the bias on R_3C_3. The grid of the relay tube then becomes less positive with the same tube current and the mechanism only reaches on equilibrium at a higher tube current.

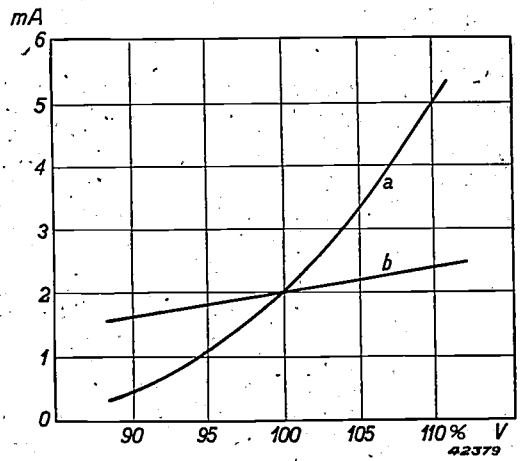


Fig. 13. Change of the tube current in mA upon fluctuations of the mains voltage expressed in percent. Curve *a* without regulator, curve *b* with regulator.

The effect of the regulatory arrangement can be seen in *fig. 13*, where the tube current is drawn as a function of the mains voltage. Without regulator the current set at 2 mA varies between 0.55 and 5 mA with approximately 10 percent variation of the mains voltage, *i.e.* between 28 percent and 250 percent of the nominal value; with the regulator the current remains between 1.6 and 2.4 mA, *i.e.* between 80 and 120 percent of the nominal value. The decrease in the effect upon heating-up is not made evident in this graph, but experiments have shown that this effect is practically inappreciable with the current regulator.

Practical application of the apparatus

We have already pointed out that due to the small distance between focus and skin very high radiation intensities can be obtained. With 50 kV tube voltage and 2 mA tube current the intensity at the surface of the "Philite" cap amounts to not less than 7000-8000 röntgen per minute. For the sake of comparison it may be mentioned that in fluorscopy of the lungs of a patient receives a dose of abour 2 *r*, that the dose necessary to cause an erytheme amounts to about 600 *r*, while 0.2 *r* is the tolerance dose, *i.e.* the dose which the operator of X-ray installations may receive daily without harm in the long run. For the therapeutic treatment of tumours of the skin total doses of, for instance, 3000 to 20,000 *r* are required. Thus a total time of irradiation of not more than several minutes is usually sufficient. This makes it possible, if desired, to "burn out" the tumour in one treatment (so-called X-ray caustic of van der Plaats). But also when for certain reasons it is not desirable to administer the total dose in a single treatment, it is in any case possible to treat the patient without hospitalization. This is in contrast to a radium treatment where the normal radiation intensity is about a thousand times smaller, so that the irradiation occupies several days. This great advantage has already contributed to the application of X-ray treatment to a much larger circle of patients. It does not, however, mean that radium treatment can be replaced by contact therapy in all cases. The choice between the two methods of treatment will often be decided by the doctor according to the ease of application in a given case.

At the beginning we showed that surface therapy, and especially contact therapy, is in a much more favourable position than depth therapy. The percentage of cures with contact therapy is very high, and in cases where there is a free choice between the two methods, namely in the case of tumours in cavities of the body, preference will usually be given to contact therapy. It is perhaps possible that in a not too distant future methods will be developed to make more deeply lying tumours accessible for the method of contact therapy with the help of operational technique.

THE MEASUREMENT OF IMPEDANCES PARTICULARLY ON DECIMETRE WAVES

by J. M. van HOFWEEGEN.

621.317.33.029.63

Several methods are discussed by which impedances can be measured at radio frequencies. At wave lengths above 1 metre it is customary to connect the impedance to be measured in parallel with an oscillation circuit and to calculate the required impedance from the detuning and damping influence experienced by the circuit. On decimetre waves a Lecher system is used as oscillation circuit. This article describes the manner in which this method has been worked out in the Philips Laboratory. As measuring instrument for the high-frequency voltage a diode voltmeter needing only relative calibration is used.

Introduction

With the increasing use of high frequencies in radio technology and television, namely waves of from several metres to a few decimetres, the necessity is more and more felt of having at one's disposal methods of measuring the impedances present at such frequencies. As one important field of application for such methods of measurement we may, for example, mention the measurement of the input and output impedances of amplifier valves, which impedances are very important for the use of those valves. A second important application is the measurement of the impedance of leakage resistances; at high frequencies the impedance of such resistances may deviate considerably from the D.C. resistance, due to skin effect and parasitic capacities and self-inductions. For use in a circuit a knowledge of the correct resistance value at the frequencies used is very important.

The characterization of impedance

The importance of a circuit or of an element in a circuit may be characterized in different ways; the simplest way is by the determination of the absolute value of the ratio between terminal voltage and terminal current and of the phase shift between these two quantities. Another manner, which often offers advantages for obtaining a clear insight, and which is therefore very commonly used in high-frequency technique, is the indication of an equivalent parallel or series connection of two elements, one of which is a pure resistance and the other a loss-free reactance. Thus for example the input impedance of a radio valve is often indicated by an equivalent connection in parallel of a resistance and a capacity, the resistance in particular depending closely upon the frequency. In measuring an impedance in high-frequency technique it generally suffices, in fact, to determine the two elements of this parallel connection.

Before discussing the measurement of impedances on decimetre waves, we shall first deal with the methods of measurement which are customary in the case of long waves.

Measurements of impedance at wave lengths greater than about 1 metre

The methods by which it is customary to measure an impedance at low frequencies (from current

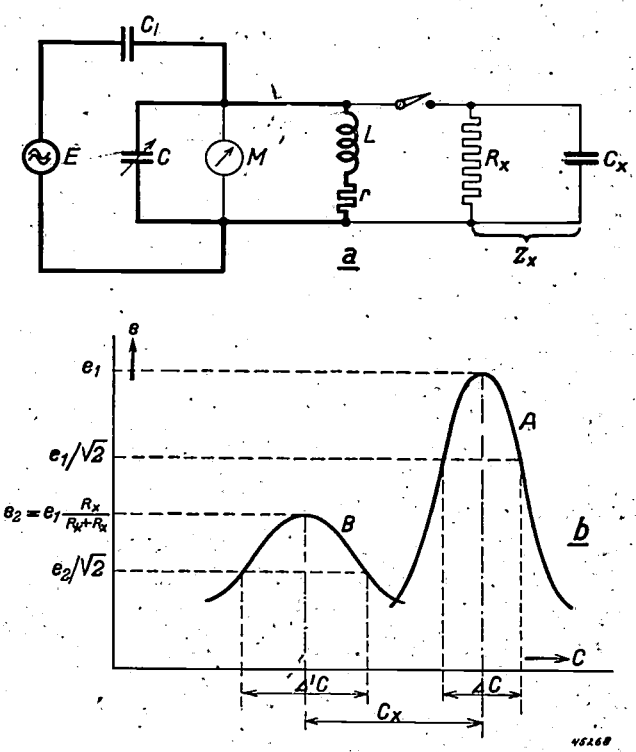


Fig. 1a) Diagram showing the principle of the method by which impedances can be determined at wave lengths above 1 metre. E is a source of high-frequency voltage. The coil L with resistance r and the calibrated variable condenser C form an oscillation circuit. M is a measuring instrument for determining the high-frequency voltage on this circuit. R_x and C_x characterize the impedance Z_x to be measured.
b) The high-frequency voltage e_m read off on M as a function of the capacity C. A without Z_x; B with Z_x.

[1]) See for instance Philips techn. Rev. 3, 357, 1940

and voltage with a Wheatstone bridge or more complicated bridge connections), cannot be used at radio frequencies without very special precautions. The chief reason for this is that due to the inevitable parasitic capacities and self-inductions in the connections it is never entirely certain whether the same current flows in two components connected in series, while also the presence of the same voltage between the terminals of two components connected in parallel is by any means not always assured. It is therefore customary to measure an impedance at high frequencies from the detuning and damping effect which that impedance exerts on an oscillation circuit. The principle of the connections used to do this is represented in *fig. 1a*. The oscillation circuit consists of a coil with the self-induction L and the resistance r and in parallel to it a calibrated variable condenser C. The circuit is coupled with the source of high-frequency voltage e, for example *via* a condenser C_k, while the voltage measuring instrument M, to which we shall revert later, is connected in parallel with the circuit. At the same time the impedance Z_x to be measured, which is for instance characterized by R_x and C_x, can be connected in parallel with the circuit.

The measurement is now performed as follows. The circuit (without Z_x) is first tuned to the frequency of the source of high-frequency voltage. The voltage read off on M, e_1 (see fig. 1b) is then a maximum. The condenser C is then increased and reduced, in both cases in such a way that the voltage on the circuit becomes smaller by a factor $\sqrt{2}$ with respect to e_1 [3]. If the difference in capacity read off on C in both these cases is equal to ΔC, the following relation exists between ΔC and the resonance resistance R_k of the circuit:

$$R_k = \frac{2}{\omega \Delta C}, \quad \ldots\ldots\ldots \quad (1)$$

where ω is the angular frequency of the high-frequency voltage applied to the circuit.

The impedance to be measured is then connected in parallel with the circuit. The condenser C must then be reduced in order to bring the circuit into tuning again. The amount by which C must be reduced is equal to C_x [4]. After the connection of Z_x and the retuning of the circuit, the voltage on the circuit e_2 is smaller than it was when the circuit was tuned to the measuring frequency without Z_x, since the resonance resistance now consists of the connection in parallel of R_k and R_x. By means of a simple calculation it can be shown that e_1 and e_2 are in the same ratio as the resonance resistances, thus:

$$e_1 : e_2 = R_k : \frac{R_k R_x}{R_k + R_x},$$

or

$$\frac{e_1}{e_2} = \frac{R_k + R_x}{R_x} \quad \ldots\ldots\ldots \quad (2)$$

R_x can now be calculated from (1) and (2).

The following is a variation of this method of measurement. After the connection of Z_x and the retuning of the circuit, C is again increased and reduced, this time by such an amount that the voltage on the circuit with respect to e_2 becomes smaller by a factor $\sqrt{2}$. If a total variation of capacity $\Delta' C$ is necessary for this (see fig. 1b), the connection in parallel of R_k and R_x is given by the relation

$$\frac{R_k R_x}{R_k + R_x} = \frac{2}{\omega \Delta' C}, \quad \ldots\ldots \quad (3)$$

It finally follows from (1) and (3) that

$$R_x = \frac{2}{\omega (\Delta' C - \Delta C)}, \quad \ldots\ldots \quad (4)$$

The first method described here, the one in which the resonance resistance after the connection of Z_x is found from the height of the peak of the resonance curve, is especially suitable for measuring impedances where the resistance component R_x is larger than the circuit resonance resistance R_k. Due to the connection of Z_x and the retuning of the circuit, the resonance resistance of the whole then undergoes only a relatively small change and e_2 is not much smaller than e_1. The high-frequency voltage supplied by the source of voltage can now be chosen so high that this small variation in voltage can be read off with a wide deviation of the voltmeter M. If the second method is applied, where the resonance resistance after the connection of Z_x is determined by the detuning of the circuit, then according to (4) the change must be measured which is experienced by ΔC through the connection of Z_x. Since ΔC is quite small as a rule, a small variation in its value can only be determined with less accuracy.

If on the other hand R_x is so small compared with R_k that e_2 is considerably smaller than e_1,

---

[2]) The capacity of the circuit is then formed by $C + C_i$.

[3]) A different factor can also be chosen. The formula for R_k, however, does not then assume the simple form represented by (1).

[4]) When the impedance to be measured may be represented by a connection in parallel of a resistance R_x and a self-induction L_x, C must be increased by an amount $\frac{1}{\omega^2 L_x}$.

the last-mentioned method is to be recommended. Since by this method it is not necessary to compare e_2 with e_1, by increasing the voltage supplied by the source of voltage provision can be made for reading off e_2 also at a wide swing of the measuring instrument M. Since in this case the difference between ΔC and $\Delta' C$ will be fairly large, the objection to this method raised in the preceding paragraph is met.

Since, as mentioned above, it is often necessary to be able to read off accurately very small values of ΔC or $\Delta' C$, while a rather large variation of C is often necessary for the measurement of C_x, in practice, instead of a single variable condenser, a connection in parallel of two calibrated variable condensers is generally used, namely a small one for the determination of ΔC and if necessary $\Delta' C$, and a large one for measuring C_x. It must further be pointed out that the absolute value of the circuit capacity is of no importance for the measurement, so that the capacity of wiring and the like plays no part. For this reason also it is not necessary to make C_k especially small. C_k may be considered as forming a part of the total circuit capacity. A large value of C_k will therefore not influence the accuracy of the measurement in the first instance. It must, however, be taken into account that with a large value of C_k the circuit is often strongly damped by the internal resistance of the voltage source. The result of this is a small value of R_k, which makes the determination of a large R_x inaccurate.

The method of measurement described is only possible due to the fact that a good variable condenser is practically loss-free at the frequencies in question, so that a variation of C does not affect the losses of the circuit. The use of a variable condenser which cannot be considered loss-free leads to incorrect results.

According to the principle described here impedances can be measured where the resistance compontent R_x is of the same order of magnitude as the resonance resistance of the oscillation circuit used (for instance from 20 R_k to 1/20 R_k).

When it is desired to measure a very small impedance, such as for example the impedance of a short metal wire, an oscillation circuit must thus be constructed with a very small resonance resistance. This would necessitate the use of a very large variable condenser (the resonance resistance is given by L/rC). Since in addition to the requirements of accuracy and freedom from losses this condenser is also bound to maximum dimensions (see later) especially at wave lengths below 10 meters, a lower limit is thereby set to the value of R_x which can be measured by this method. Very small impedances can, however, be measured by a similar method. They are then not connected in parallel with the oscillation circuit, but in series with the circuit coil. The formulae to be applied in this case are less simple than those given above. In order to work them out a knowledge is required of the absolute value of the capacity C (or of the self-induction L).

In practice the method described can be applied for wave lengths of up to 1 metre, not already at wave lengths of several meters certain precautions are necessary in order to avoid systematic errors in measurement. Special attention should be paid to keeping the connections between the various components very short in order to prevent the self-induction of the connecting wires from causing incorrect results. Where for mechanical reasons a connection cannot be made short enough, it may be necessary to compensate the self-induction of the connecting wire by connecting a capacity in series with the wire in question. Thus in *fig. 2*

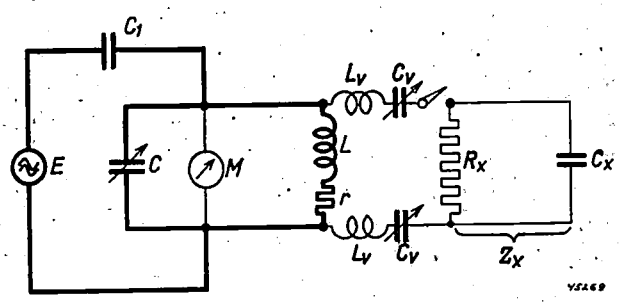


Fig. 2. The self-inductions L_v of the connecting wires of the impedance to be measured can be compensated by capacities C_v. See the text under fig. 1 for the meaning of the other symbols.

it is indicated how the self-induction L_v of the connections between the impedance to be measured and the circuit can be compensated by capacities C_v, which are in series resonance with L_v [5]), so that the following holds: $\omega L_v = 1/\omega C_v$.

Another objection which often arises is that the calibration of the condenser C, which has usually been carried out at a low frequency, is no longer correct at the high frequency at which the measurement is made, owing to the self-induction of the connecting wires. In these connections also, therefore, capacities should be included in this case. Furthermore the dimensions of the condenser C should be kept small, since otherwise the self-induction of the fixed and rotating plates may cause the above-mentioned difficulty. Since in this case the self-induction depends upon the size of C, it cannot be eliminated by a condenser in the connections.

[5]) The adjustment of these capacities to the correct value may take place in different ways. A common method is to connect a second voltage measuring instrument instead of Z_x and to adjust the condensers C_v in such a way that the voltage which is read off on the latter instrument is equal to the voltage read off on M.

Not only is it possible to determine the resonance resistance of an oscillation circuit from a variation in capacity, but it is also possible by a variation in the frequency of the high-frequency voltage source. A fixed tuning capacity can thus be used in this case. The frequency is now varied so much that the circuit voltage is decreased by a factor $\sqrt{2}$ on both sides of the resonance peak. When the total frequency variation necessary for this is Δf the resonance resistance is given by

$$R_k = \frac{1}{2\pi C \Delta f} \quad \ldots\ldots\ldots \quad (5)$$

Since at the very high frequencies as considered here the accurate measurement of small frequency variations is generally more difficult than the accurate calibration of a variable condenser, the measurement of a fixed frequency and a variable capacity is of more importance in practice.

Practically the only measuring instrument which can be used for the measurements in question is a diode voltmeter. In order to keep the connections, which carry the high-frequency AC, short, the diode is generally soldered directly to the circuit. In *fig. 3* a diagram is given of the way in which a diode voltmeter can be connected. *Via* a condenser C_d the diode D is joined with the top of the oscillation circuit. The DC voltage obtained on the diode causes a direct current in the leakage resistance R_l which is measured with the micro-ammeter M_1. The battery V_a and the potentiometer R_a make it possible to give a small negative voltage to the anode of the diode, so that the operating point can be adjusted to a favourable position on the diode characteristic. The battery V_f and the variable resistance R_f serve for the provision and regulation of the heating current for the diode.

Further it must be noted that the diode voltmeter needs only to be calibrated relatively for these measurements. Since it is only desired to measure voltage relations, the absolute value of the high-frequency voltage is of no importance. It is also unnecessary that the same high frequency voltage should act on the diode as on the oscillation circuit. The coupling between the two may therefore, if desired, be very loose, with a small value of C_d. If necessary the diode can also be coupled with the oscillation circuit in some other way, for example through an inductive coupling with the coil L.

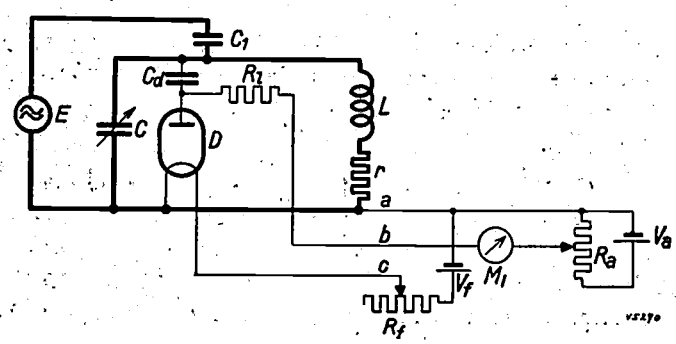


Fig. 3. Connections of the diode voltmeter for measuring the high-frequency voltage on the oscillation circuit. D is the diode, R_l the leakage resistance, M_1 is a microammeter, V_f and V_a are batteries for supplying the filament current and a small negative anode voltage which can be regulated, respectively, with the variable resistance R_f and the potentiometer R_a. The connections indicated by a, b and c do not carry H.F. current, and may therefore have any length.

As a rule the calibration of the diode voltmeter is carried out at a low frequency. Now there are two reasons why a calibration performed at a low frequency is no longer correct at very high frequencies. The first lies in the self-induction of the supply lines to the diode and the capacity of the diode. This, however, affects only the absolute calibration [4]). The second reason results from the fact that the electrons need a certain time to reach the anode from the cathode; although this time is very short, often at very high frequencies it may not be neglected, having regard to the oscillation time of the anode AC voltage. Since the transit time of the electrons depends upon the anode voltage, the frequency at which the influence of the transit time of the electrons on the calibration of the diode voltmeter begins to be felt, will be in part determined by the magnitude of the AC voltage applied to the diode. The result is that at very high frequencies there is also an effect of the frequency on the relative calibration. In order to keep the transit time of the electrons very short and thereby to make the calibration performed at a low frequency still valid at the highest possible frequency, for measuring purposes diodes are used which have been especially developed for this purpose [6]), whereby among other features, especial attention has been paid to procuring a small distance between anode and cathode.

Impedance measurements on decimetre waves

For wave lengths below 1 metre it is practically impossible to make a normal oscillation circuit with variable capacity, since the self-induction and capacity necessary are too small. It is, however, possible to construct a cavity resonator for wave lengths below 1 metre, but there are various objections to this for measurements by the method described above. The chief of these objections

[6]) See Philips techn. Rev. 7, 124, 1942.

is that the connection of the impedance to be measured would have to be inside the cavity resonator in order to keep the connections short, and a very impractical construction would result.

An obvious solution would of course be to use a Lecher system and self-induction. In the following we shall describe a method of measurement worked out in this laboratory in which use is made of a Lecher system.

A Lecher system which is short-circuited at one end and the length of which is about equal to a quarter wave length, shows much similarity to an oscilation circuit with concentrated capacity and self-induction in which the two are connected in parallel (parallel circuit). Similarly to the latter, the Lecher system exhibits a high resonance resistance [7]) with sufficient freedom from loss. Detuning of the system may take place by changing the length, thus by sbifting a movable short-circuiting bridge. We can now determine the resonance resistance in a way which is entirely analogous to the method followed in a parallel circuit. For this purpose the Lecher system is coupled with a high-frequency voltage source e, for instance by means of small capacities C_k (*fig. 4a*; the switches S are for the time being assumed to be open). The high-frequency voltage at the terminals of the Lecher system can be read off on a measuring instrument M.

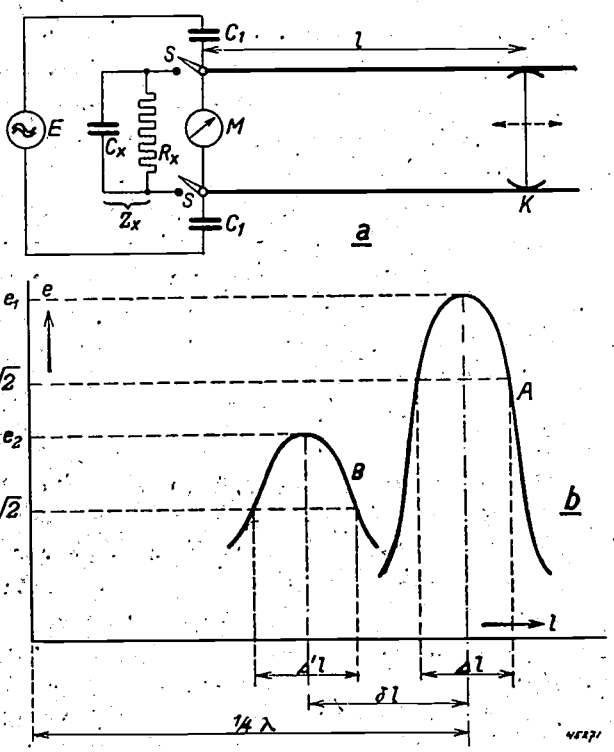


Fig. 4*a*) Diagram showing the principle of the method by which impedances can be measured with the help of a tuned Lecher system. The system, of which the length is l, is coupled by means of the small capacities C_k with the source of high-frequency voltage E. M is a measuring instrument for indicating the high-frequency voltage on the input terminals of the Lecher system. R_x and C_x characterize the impedance to be measured Z_x. K is the movable short-circuiting bridge with which the Lecher system is tuned.

b) The high-frequency voltage e read off on M as a function of the length l. A without Z_x; B with Z_x.

The resonance resistance is now measured as follows: The Lecher system (without Z_x) is first tuned to the frequency of the source of high-frequency voltage e. The voltage e_1 (see fig. 4*b*) read off on M is then a maximum. The length is then increased and decreased, in both cases in such a way that the voltage with respect to the maximum becomes smaller in the ratio $1 : \sqrt{2}$. If the distance between these two positions of the short-circuiting bridge is Δl the following relation holds between the resonance resistance R_x and Δl:

$$\frac{R_s}{\zeta} = \frac{\lambda}{\pi \Delta l} = \frac{6 \cdot 10^{10}}{\omega \Delta l}, \quad \ldots\ldots\ldots \quad (6)$$

where ζ represents the wave resistance [7]) and Δl is expressed in cm. This formula only holds as long as Δl is small compared with a quarter wave length. A normal system, however, always exhibits such a high resonance resistance R_s that Δl is small enough to permit the application of formula (6).

Formula (6) shows much similarity with (1) and this similarity can be made plainer by a slightly different form of the equation. If we call the self-induction and the capacity of the Lecher system per unit of length L^I and C^I, respectively, the wave resistance [5]) is given by

$$\zeta = \sqrt{\frac{L^I}{C^I}},$$

while there is the following relation between L^I and C^I:

$$\sqrt{L^I \cdot C^I} = \frac{1}{v} = \frac{2\pi}{\omega \lambda},$$

where v represents the velocity of propagation of the electromagnetic waves along the Lecher system.
When we obtain for the resonance:

$$R_s = \frac{2}{\omega C^I \Delta l}.$$

Since C^I represents the capacity per unit of length $C^I \Delta l$ is the variation in capacity of the system and thus the expression corresponds exactly to (1).

In the foregoing it was assumed that the Lecher system is in resonance with the voltage of the measuring transmitter at the length l, which is exactly equal to a quarter wave length (see fig. 4*b*, curve A). In practice, however, this cannot usually be realized. It is impossible to avoid the occurrence of a certain capacity between the connection terminals, *i.e.* the capacity of the

[7]) See Philips techn. Rev. 6, 240, 1941.

insulator by which the extremities of the two conductors must be supported and the capacity of the coupling condensers C_k. Due to these capacities resonance will occur with the transmitter voltage at a length l which is shorter than a quarter wave length. When the capacity in question can be kept so small that the length l at which resonance occurs differs only slightly from a quarter wave length, formula (6) can also be used for the calculation of the resonance resistance. If, however, the capacity in question is not small enough for that, a compensation of that capacity can be obtained by means of an auxiliary Lecher system (see later).

For the measurement of Z_x, this impedance is now connected in parallel with the Lecher system (fig. 4*a*, the switches S being closed). If the impedance to be measured may be characterized by the connection in parallel of a resistance and a capacity, the length of the Lecher system must be decreased in order to bring the system into tuning again with the high-frequency voltage applied. If the length by which l must be decreased is equal to δl (see fig. 4*b*), C_x is determined by

$$\omega C_x = \frac{1}{\zeta} \tan 2\pi \frac{\delta l}{\lambda} \quad \ldots\ldots\ldots \quad (7)$$

If the impedance to be measured can be represented by a connection in parallel of a resistance and a self-induction L_x, for retuning l must be increased. The self-induction L_x can then be calculated from the equation

$$\omega L_x = \zeta \operatorname{cotan} 2\pi \frac{\delta l}{\lambda} \quad \ldots\ldots\ldots \quad (8)$$

For the determination of the resonance resistance of the Lecher system and of the reactive part of the required impedance, therefore, the Lecher system is used in the same way as the parallel circuit already discussed, although with the application of somewhat different formulae. In the determination of R_x, however, account must be taken of the fact that in a Lecher system the losses are dependent on the variable element, namely the length, in contrast to the losses of the parallel circuit, which do not depend upon the variable element (loss-free condenser). While in the connections according to fig. 1*a* the inclusion of a loss-free condenser C_x (thus without R_x) and the corresponding decrease of C exert no influence on the resonance resistance, in the connection according to fig. 4*a* the resonance resistance *does* experience an alteration due to the connection of a loss-free condenser C_k and the corresponding decrease of l. It would, therefore, lead to incorrect results if one reckoned that after the connection of Z_x and the retuning of the Lecher system the resistance between the terminals would be composed of the connection in parallel of R_x and the value of R_s calculated from (6). The determination of R_x from the ratio of e_2 to e_1 or from $\Delta' l$ and Δl (see fig. 4*b*) is now also possible, but this calculation leads in general to unreliable results, since the manner in which the losses depend upon the length of the Lecher system is usually not known accurately. These losses are due to various causes, namely:

a) the resistance of the conductors and of the moveable short-circuiting bridge,
b) the resistance of the contacts between the conductors and the short-circuiting bridge,
c) losses in the insulators by which the conductors must be supported,
d) radiation of the conductors and of the short-circuiting bridge.

It now depends upon the construction of the system what relation these losses will have to each other. In the case of a Lecher system used without shielding the radiation of the short-circuiting bridge is often dominant. By the introduction of suitable shielding the radiation losses of conductors and short-circuiting bridge can be eliminated for the greater part, but this can never be complete in the case of the first-mentioned, since the shielding must be open at the side where the impedance Z_x is to be connected, and the parts of the conductors lying in the vicinity of this opening will thus radiate.

Each of the losses mentioned exhibits a different kind of dependence on the length of the Lecher system. In practice this is found to be of no importance to the application of formula (6) which relates only to small variations in the length, in the neighbourhood of 1/4 wave length. The losses of the Lecher system, even at large values of δl, have in the first approximation no influence on formulae (7) and (8). The quantities e_2 and $\Delta' l$ are, however, dependent on the way in which the losses are connected with the length of the system. It is thus in general impossible to calculate R_x in a simple way from the ratio between e_1 and e_2 or from Δl and $\Delta' l$. Only when the impedance to be measured possesses practically no reactive component, thus when δl is small compared to a quarter wave length, can R_x be determined by one of these methods.

In many cases the main cause of the losses occurring does not lie in the Lecher system proper, but in the measuring instrument M. The losses are then entirely concentrated at the connection terminals of the system. In that case the determination of R_x may take place by one of the above mentioned methods, even with a larger value of δl. In the determination from Δl and $\Delta' l$ the connection in parallel of R_x and R_s is given by the formula:

$$\frac{R_x R_s}{R_x + R_s} = \frac{\lambda}{\pi \Delta' l} \cos^2 (2\pi \frac{\delta l}{\lambda})$$

which equation in combination with (6) makes possible the calculation of R_x.

Owing to the uncertainty mentioned about the way in which the losses depend upon the length of the Lecher system, it is desirable to construct the measuring arrangement in such a way that the measurement of R_x may take place while the length l is the same as that at which R_s was measured. A practical method is to connect a second Lecher system (*fig. 5*) with the initial terminals of the above mentioned Lecher system. This second system is likewise provided with a moveable short-circuiting bridge so that its length l_{II} can also be adjusted.

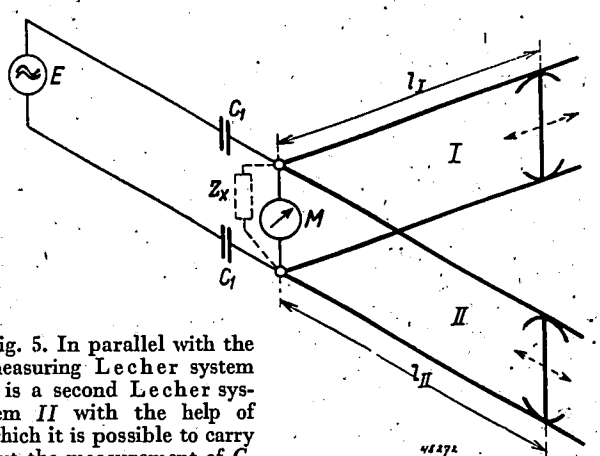


Fig. 5. In parallel with the measuring Lecher system I is a second Lecher system II with the help of which it is possible to carry out the measurement of C_x as well as of R_x, while the length l_I deviates only slightly from a quarter wave length.

Let us now first assume that l_{II} is equal to a quarter wave length. The input impedance then has no reactive component from this. As a result the measurement of C_x may take place with the aid of system I, entirely as described above. Now if the system II is exactly like system I, the measurement of R_x can be performed with the help of II. For practical reasons, however, in the installation constructed in this laboratory, system II is built differently from system I (see below). In order to determine R_x one now sets to work as follows.

After C_x has been determined with the help of system, I l_I is again set at a quarter wave length. The length l_{II} is now changed so that C_x is compensated by the reactive component of the input impedance of system II, which is ascertained from the fact that the deviation of the measuring instrument M is then at a maximum. The impedance between the terminals of the two Lecher systems is now composed of the connection in parallel of R_x, R_s and the input resistance of system II which we shall indicate by R'_s. The magnitude of this parallel connection can now be determined with system I by measuring $\Delta' l$ (δl now being equal to zero), and we can calculate R_x from this if we know the parallel connection of R_s and R'_s. Since R'_s depends upon the length l_{II} of the auxiliary Lecher system, the latter parallel connection must be measured at that length of l_{II} at which system II is in resonance with C_x. This can be done most simply by connecting with the terminals, instead of Z_x, a loss-free condenser the size of which corresponds to C_x. Since at the wave length considered here it will seldom be necessary to measure capacities of more than a few pF, this auxiliary capacity may take the form of two small metal plates a short distance apart. The adjustment of the capacity then takes place by bending these plates.

The auxiliary Lecher system has still another function in the measurement. The diode voltmeter is inductively coupled with the short-circuited end of the system. Since here also a relative measurement of the high frequence voltage is sufficient, it is unnecessary to know the absolute value of the voltage at the terminals of system I. It is thus sufficient to measure the voltage between the terminals m and n of a loop which is inductively coupled with the short-circuited end of Lecher system II (See *fig. 6*). At any length l_{II} this voltage is proportional to the terminals voltage of the two systems. By applying this possibility of setting up the diode at some distance from the terminals of the Lecher system, various structural objections are met which occur in the connection of a diode directly to these terminals. In particular it is generally very difficult to keep the connections to the impedance to be measured short when a diode is introduced at this spot. The inductive coupling of the voltmeter with the short-circuiting bridge of system II makes it impossible to consider this system as being actually short-circuited. For this reason it is desirable that the measurement not only of C_x, but also of R_x, should be carried out with the help of system I as described above.

As has already been indicated, there is always

a certain capacity between the connection terminals, owing to which it would not be possible to use a length l_I which is approximately a quarter wave length. In the variable length l_{II} of system *II* one has a means of compensating these capacities, so that all the measurements can therefore be carried out at a length l_I of Lecher system *I* which differs only little from a quarter wave length.

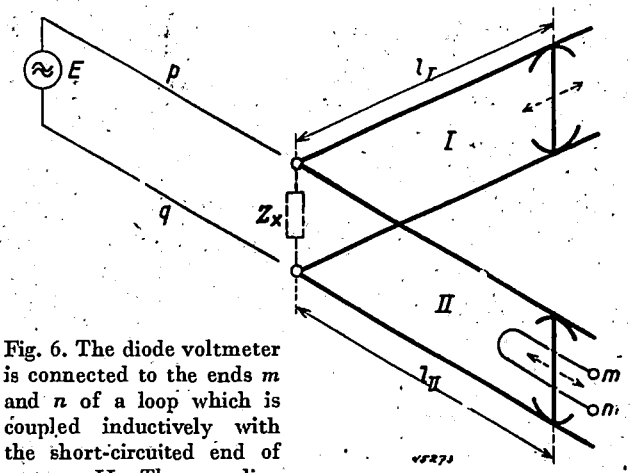


Fig. 6. The diode voltmeter is connected to the ends *m* and *n* of a loop which is coupled inductively with the short-circuited end of system *II*. The coupling with the measuring transmitter *E* is realized by the situation of the ends of the two conductors *p* and *q* at a short distance from the connection terminals of the Lecher system.

The coupling of the source of voltage (measuring transmitter) *E* with the measuring arrangements can be carried out by placing the ends of the two conductors *p* and *q* at a short distance from the connection terminals of the two Lecher systems (see fig. 6). By a variation of this distance the coupling can be regulated according to need. In order to keep the connections short the impedance to be measured is not, as is represented in fig. 4*a* for the sake of simplicity, connected by means of switches, but the object to be measured is soldered directly to the terminals.

Fig. 7 is a sketch of a measuring set-up as described in the foregoing. The arrangement shown here is used for impedance measurements at wave lengths of about 50 cm.

Just as in measurements with a parallel circuit, in those with a Lecher system there is a lower limit to the resistance R_x which can be measured[8]). At a small value of R_x according to formula (6) a large value of Δl is necessary, and for a value of

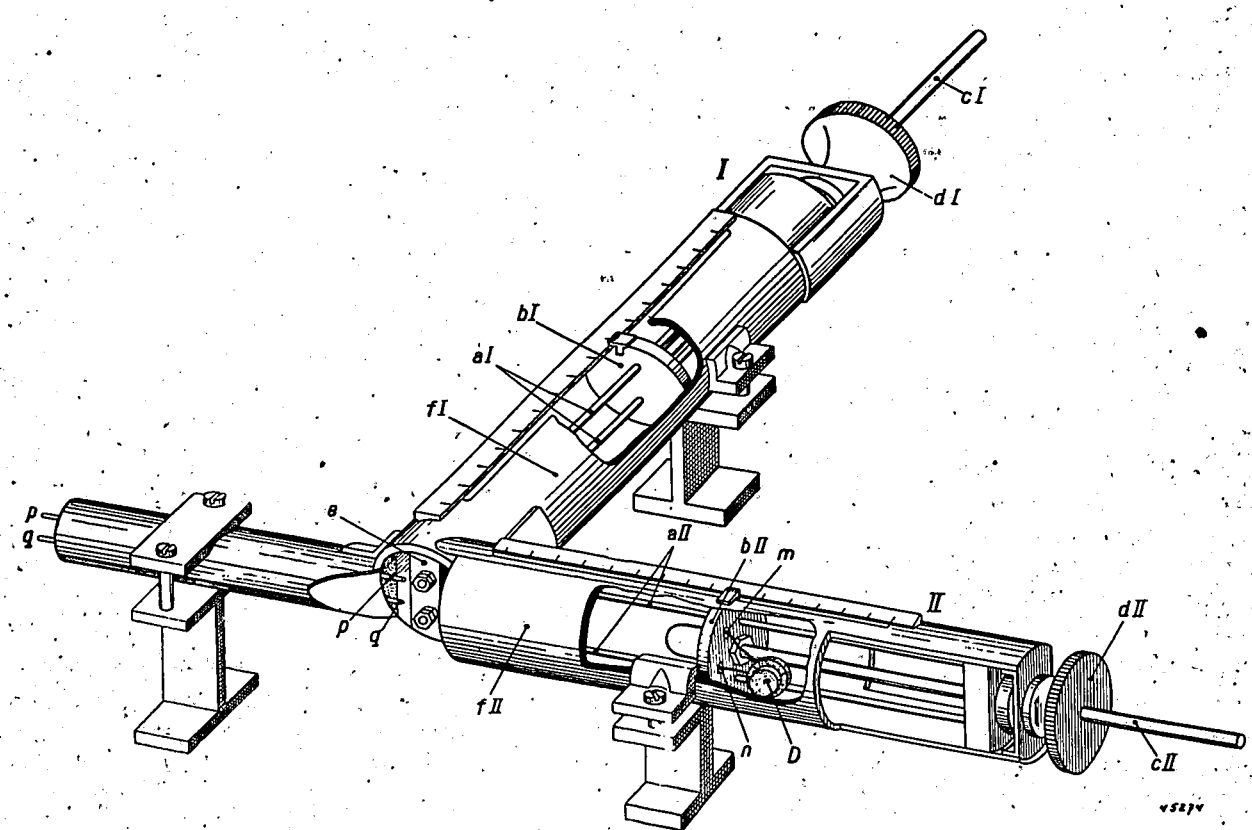


Fig. 7. Sketch of an arrangement for the measurement of impedances at wave lengths of about 50 cm according to the principle indicated in fig. 6. In order to show the construction clearly the shields *f* which surround the Lecher systems are partially cut open. For the meaning of the symbols *I*, *II*, *p*, *q*, *m* and *n* we refer to figures 5 and 6. The two conductors of each Lecher system are indicated by *a*; *b* are the short-circuiting bridges, which are moved by means of the screw rods *c* and the knobs *d*. *e* is a plate of insulation material. *D* is the measuring diode.

Δl which may not be considered as small compared with a quarter wave length formula (6) can no longer be applied. The decrease of the wave resistance ζ, which according to (6) would also make possible the measurement of a smaller value of R_x, is often incapable of being realized, since for a small wave resistance the two conductors of the Lecher system must be very close together, which usually leads to mechanical difficulties.

Resistances which are at least several times smaller than the wave resistance can be measured by connecting them with the measuring apparatus with the intermediate connection of a Lecher system having the length of a quarter wave length. As was stated in the article cited in footnote 5), such a Lecher system may be considered as a transformer which transforms an impedance Z into an impedance ζ^2/Z. A resistance which is a given number of times smaller than the wave resistance ζ is thus transformed with such a transformer into a resistance which is the same number of times larger than the wave resistance. Actually the introduction of this transformation Lecher system comes down to the use of a Lecher system closed at one end one *half* wave length long.

It is obvious that the precautions against the occurrence of systematic measuring errors, which were already mentioned in the case of measurements at wave lengths above 1 m, are even more necessary at decimetre wave lengths. Careful attention should be paid to keeping the connections very short. Where for technical reasons a connection cannot be made short enough, in interpreting the results of the measurement account must be taken of the presence of this connection. With the very compact assembly which is necessary for decimetre waves it is often impossible to introduce a condenser to compensate the self-induction of a connection.

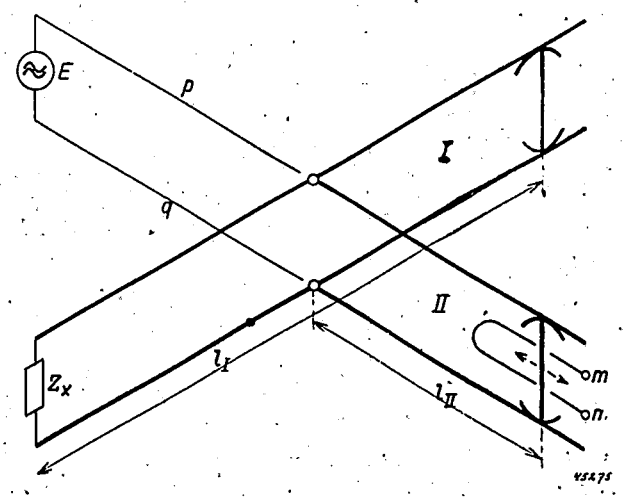


Fig. 8. An impedance Z_x which is small compared with the wave-resistance can be measured by the aid of a Lecher system the lenght of which amounts to about half a wave length. The coupling of the measuring transmitters L_1 and the connection of the auxiliary Lecher system *II* take place in the middle of the Lecher system *I*.

8) An upper limit is set by the accuracy with which Δl can be determined.

AN APPARATUS FOR STROBOSCOPIC OBSERVATION

by S. L. de BRUIN.

771.448.1:778.39

When a condenser is discharged through a gas-discharge lamp, a light flash of short duration and high intensity is obtained. In this manner it is possible to make a source of light which is suitable for stroboscopic observations and for the photography of rapidly moving objects.

A description is given in this article of a stroboscope developed by Philips, consisting of a tube filled with rare gas as gas-discharge lamp and an electrical apparatus for the condenser discharges. The latter furnish periodically repeated current impulses of 2000 A at time intervals which can be varied between 2 sec. and 1/250 sec. The light flashes hereby excited have an intensity of about 200 lumen seconds with a maximum light flux of 2×10^6 Dlm (duration 10^{-5} sec.). The period of the light flashes can be synchronized in many ways with the phenomenon to be investigated, so that numerous possibilities of application exist, which are illustrated by means of a few examples.

If one compares the performance of a piece of modern optical apparatus with that of the human eye it will be found that nature still leads technology in sensitivity, but that the technical devices offer the advantage of a much higher velocity and greater precision. The eye is unable to follow phenomena which take place in less than $^1/_{10}$ second. Rotational motions of more than 3 or 4 revolutions per second can no longer be clearly seen, while in the case of translation at a speed of more than 5 m/sec, seen from a distance of 1 m, the image already becomes so indefinite that there can be no question of exact observation.

Cameras with focal plane shutters can work approximately 100 times as fast, and in this way record phenomena which are invisible to direct observations. The velocity of about 0.001 second thus obtained still fails, however, to meet the requirements made in many technical investigations. With high-speed machines, for example, speeds up to 100 revolutions per second and translatory motion at speeds of up to 100 m/sec are not exceptional. If in the latter case it is desired to fix accurately the momentary situation of a part of a machine to within 1 mm, the observation may not last longer than 10^{-5} sec, and the moment of the observation must also be fixed accurately to within 10^{-5} sec.

A device commonly used to satisfy these requirements is the stroboscope. The action of the stroboscope is based upon the fact that it makes the objects to be investigated visible or photographable only at those moments at which the observation is desired. In the commonly occurring case in which these moments are repeated periodically there are simple devices. For example, a rotating disc with holes distributed at regular intervals around the circumference of a circle can be placed in the entrance pupil of a telescope. When this disc is rotated at the correct speed the successive holes will follow each other at such time intervals that the part of the machine which rotates or runs back and forth has executed exactly one cycle in that interval and is back again in the old position. The moving object is thus, as it were, fixed. Instead of fixing it, it may also be allowed to be displaced slowly, either forward or backward, by rotating the disc somewhat more slowly or more rapidly.

Another method of limiting the visibility of an object to certain periodically repeated moments consists in the use of a source of light which changes in intensity periodically. When a moving object is irradiated with such a light source the strongest visual impression is received at certain moments, so that by the correct choice of frequency and phase certain positions of the moving object can be emphasized. This stroboscopic effect of a source of light often occurs unintentionally when a machine which is driven by a synchronous motor is illuminated with a lamp which is fed from the same main as the synchronous motor. This phenomenon is most pronounced with certain gas-discharge lamps, which emit a very fluctuating radiation when fed with A.C. In the case of the arc light there is also a strong stroboscopic effect, while it is practically absent in the case of filament lamps.

In the case of non-periodic phenomena also the use of an intermittent light source offers improved possibilities of observation. A familiar experiment is the observation of an induction machine (Wimshurst machine) by the light of the jumping spark. The duration of the electrical flashover is so short that the discs of the machine seem to stand still. It is striking that the inertia of the eye here forms no hindrance; apparently the eye can receive a visual impression within any given short

time interval provided the total amount of light is sufficient and the eye has time enough after the flash to assimilate the impression.

The light of a spark is also very suitable for photographing a moving object. In this way it is easily possible to reduce the exposure time to 10^{-5} or 10^{-6} sec. It is less easy to fix also with this precision the moment at which the exposure takes place. It depends upon the nature of the phenomenon to be observed what devices must be used; mechanical, photoelectrical or acoustic devices can be employed.

In the following an apparatus will be described which is suitable for stroboscopic examination in laboratories and factories. This apparatus, type GM 5500, is so designed that a single flash as well as periodically repeated flashes can be obtained. The time interval between the flashes can be adjusted within wide limits, and the generator which excites the flashes can be synchronized with the part of the machine to be observed. The light intensity of the flashes is so great that the apparatus can be used at the brightness normally present in a room in the daytime without hindrance from the continuous light.

The flash lamp

As already stated above, the flash of an electric discharge is used. Now for a permanent set-up it is undesirable that the spark should occur in air, because the electrodes would be too strongly attacked. The obvious solution is to use a rare gas instead of air. The spark gap is thus constructed as a discharge lamp filled with a rare gas.

The best results were obtained with argon at a high pressure. The higher the pressure, the greater the intensity of the light of the flashover. A limit is set to the pressure by the requirement that the flashover voltage may not be too high and that a reasonable lifetime must be ensured for the lamp.

The light of the high-pressure argon lamp is a bluish white and has a continuous spectrum with a high photographic actinity. It thus shows some resemblance to a water-cooled mercury lamp, which could also be used as flash lamp. However, due to the fact that there is no necessity for water-cooling, the argon lamp is easier to handle and, moreover, possesses various advantages compared with the mercury lamp: it is always ready for use, while in order to obtain the necessary vapour pressure the mercury lamp must first be warmed, and it has a shorter flash time, since the light emission of the mercury lamp is not only an electrical phenomenon but in part also a thermal phenomenon, so that there is a certain phosphorescence due to the heat inertia of the mercury vapour.

After a detailed investigation of the most favourable combination of gas pressure, separation of electrodes and tube dimensions, a lamp with very good technical characteristics was successfully designed. The flash time amounts to about 10^{-5} sec; within this time an amount of light of 200 lumen seconds is emitted and an energy of 2 W sec. consumed, so that the specific light flux amounts to 100 lm/W. This is about the same efficiency as can be obtained under favourable conditions with a water-cooled mercury lamp of very high pressure.

The light flux during the flash is about 2×10^6 Dlm with a current of 2000 A and a power of 200 kW. In addition to these very high peak loads the lamp also receives a small continuous load in order to facilitate the breakdown. The total energy consumption of the lamp is about 75 W at, for example, 25 flashes per second.

The tube of the lamp consists of quartz and is mounted in a nitrogen-filled bulb. The nitrogen filling is necessary, since with air filling the exterior of the hot quartz tube would be attacked by the atomic oxygen which is formed due to the ionizing action of the ultraviolet radiation of the lamp.

The rear of the bulb is covered on the inside with a mirror and concentrates the light beam in a relatively small solid angle, so that an amplification of the illumination of about 75 times is obtained in the axis of the beam. At a distance of 2 m from the lamp the peak value of the illumination intensity at the axis is about 10^7 lux, which is sufficient for photography even with a strongly diaphragmed lens in the available time of 10^{-5} sec. The photographs given further on in this article were made with diaphragms F/16-F/22, so that it was possible to place the camera close to the object and still obtain great depth of focus.

In *fig. 1* the stroboscope is shown. The lamp is mounted in a rotating holder which is connected by means of a high-tension cable with a cabinet on wheels containing the electrical apparatus. The light distribution of the stroboscope lamp is reproduced in *fig. 2*.

The electrical apparatus

From the description of the flash lamp it has become apparent that the electrical apparatus must furnish current impulses of 2000 A and with a duration of 10^{-5} sec. Depending upon the appli-

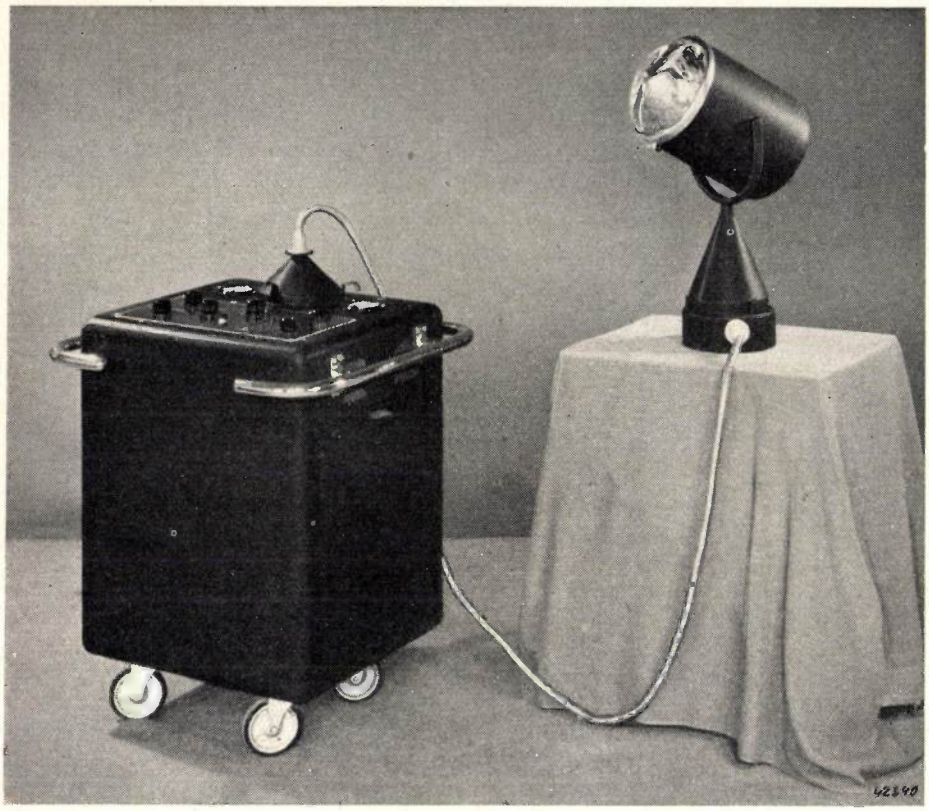


Fig. 1. The Philips stroboscope GM 5500.

cation it may be desirable to obtain single or periodically repeated discharges, while in every case it is important to be able to synchronize the moment of the breakdown with the phenomenon to be observed.

For feeding the flash lamp we use connections with a relay valve which is ignited by a voltage impulse at the moments at which the discharges are required. The frequency at which these impulses, which are excited by a separate generator, are repeated can be regulated between 0.5 c/sec (30 impulses/min) and 250 c/sec (15 000 impulses/min). Moreover, it is possible to synchronize the impulses with a voltage led in from the outside. Finally the apparatus contains an electrical frequency meter which is coupled with the control generator.

We shall now describe successively the chief components of the apparatus.

The supply generator for the flash lamp

The connections for the supply generator for the flash lamp are given in *fig. 3*. *Via* a charging resistance R_1 a condenser C_1 is gradually charged to 600 V in order to be discharged again at certain moments *via* the relay valve and the flash lamp *A*, producing thereby the required current impulses.

As already discussed, the relay valve must be able to deal with several thousand amperes, while the average current amounts to less than 1 ampere. Taking into account these very abnormal conditions of use, a valve of special construction was designed and is shown diagrammatically in *fig. 4*. In connection with the low average load the dimensions of this valve were chosen relatively small, so that at each breakdown the electrodes are heavily overloaded. This overloading however, has no harmful results since both electrodes consist of mercury. Only an intense evaporation of the mercury occurs during the current impulse. The vapour

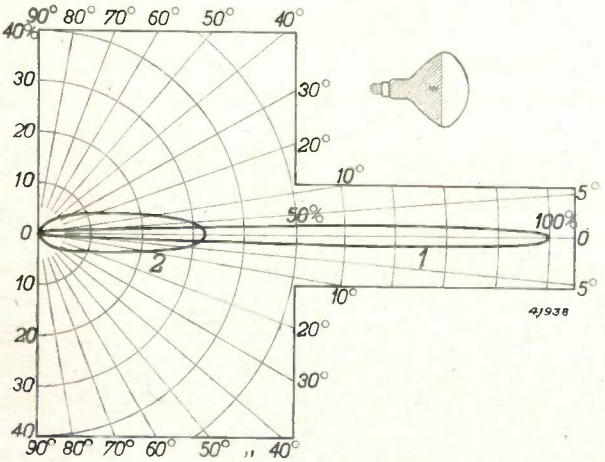


Fig. 2. Distribution of the light intensity of the stroboscope lamp, *1*) with a clear, *2*) with a frosted bulb. The rear wall of the bulb is covered with a mirror in both cases. The scale of curve *2* is 10 times as large as that of curve *1*, so that the maximum light intensity for the clear bulb is about 30 times as great as for the frosted bulb.

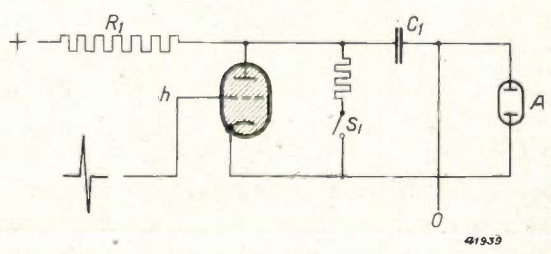


Fig. 3. Connections for the supply of the flash lamp *A* by means of a condenser and a relay valve.

condenses in the glass domes over the electrodes and the condensed mercury drips back to the electrodes.

For the ignition of the relay valve an auxiliary electrode h is introduced at the cathode side, which upon application of a voltage impulse (about 8 kV), causes the occurrence of a cathode spot and thus initiates the discharge. The condenser is then discharged *via* the relay valve and the flash lamp, whereupon the relay is extinguished. If due to the current supplied *via* the charging resistance R_1 the relay valve should continue to burn, it may be extinguished by practically short-circuiting it during a short time by means of the switch S_1.

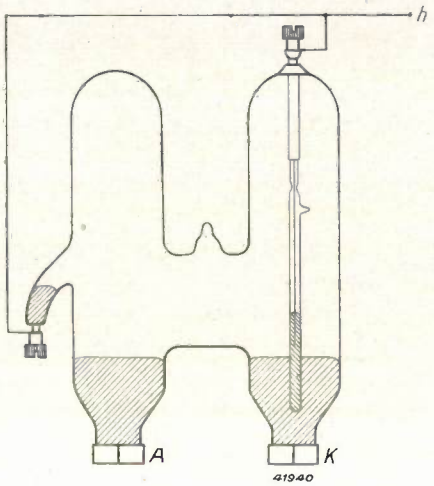


Fig. 4. Construction of the relay valve for the supply of the flash lamp. The electrodes are of mercury and can tolerate short-lived current impulses of 2000 A. For the ignition of the valve a voltage of a few thousand volts must be applied to the auxiliary electrode h.

According to the employment of the stroboscope there are different ways of synchronising the voltage impulses which serve to ignite the relay valve, namely:

a) with the AC voltage of the supply mains,
b) with an AC voltage or voltage impulse of about 50 V from the outside,
c) mechanically, by breaking a contact,
d) mechanically, by making a contact; this is also used for electrical synchronization with a single voltage impulse.

In order to obtain the synchronization voltages in the case of non-periodic phenomena, use may be made of a microphone or a photocell. Since these apparatuses only give minimum voltages or currents, it is of course necessary to employ a suitable amplifier.

As an illustration of this method of procedure a set-up for the photography of the collision between

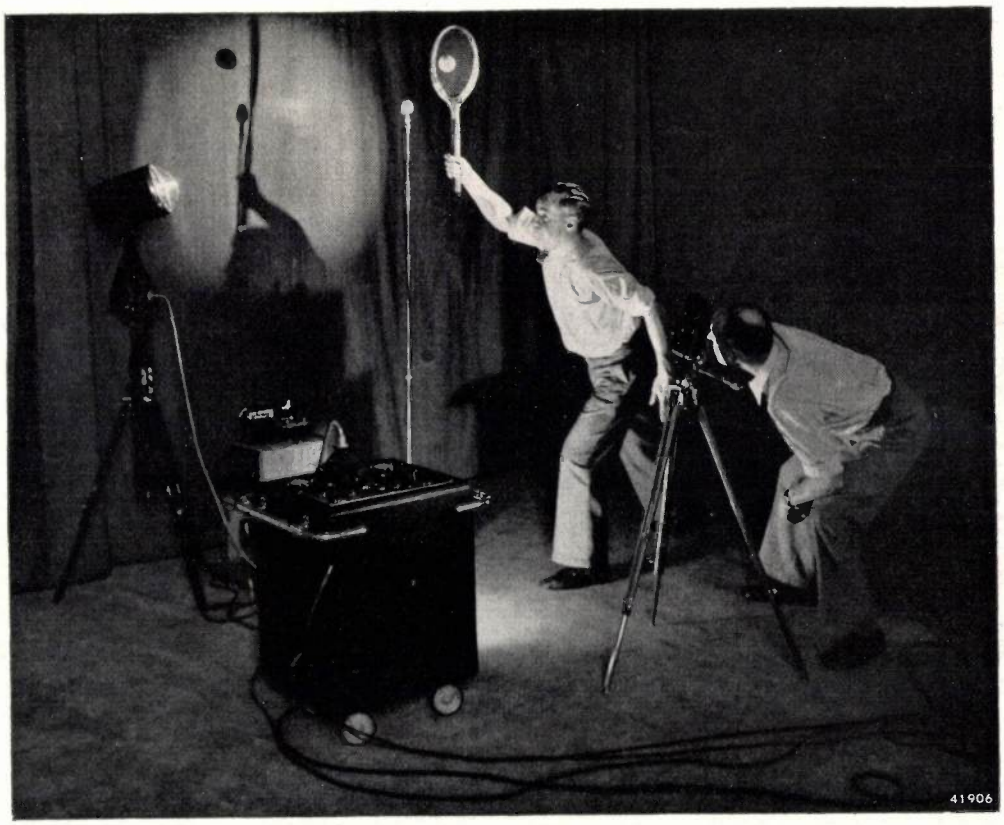


Fig. 5. Arrangement for photographing the collision between tennis ball and racket with the help of the stroboscope. The illumination is synchronized by means of a microphone.

a tennis ball and a racket is reproduced in *fig. 5*. The synchronization is here obtained by means of a microphone which reacts to the ping of the collision. The greater the distance between racket and microphone the greater the difference in time between collision and moment of exposure. This may be seen in *fig. 6*: exposure *a*) was made with a distance of only a few decimetres, while in the case of exposure *b*) the distance was more than a metre. In the first case the moment of contact was recorded, while in the latter case at the moment of exposure the ball had already rebounded several centimetres.

The frequency metre

In order to be able to control accurately the number of flashes per second furnished by the stroboscope, the apparatus is provided with a frequencymeter. This instrument measures the frequency indirectly by indicating the average charging current of a condenser which is discharged at each flash through the gas triode, to be charged again afterwards to a certain voltage. *Fig. 7* shows the principle of the connections.

The gas triode of the frequencymeter is controlled by voltageimpulses which are taken from the same impulsevoltage of 8 kV with which the relay valve of the supply generator is controlled. In addition to this impulsevoltage there is also on the grid of the gas triode a negative bias of such a magnitude that in the absence of an impulse no breakdown can occur. Since at every breakdown the condenser is almost completely discharged, the average current through the gas triode is equal to the charge of the condenser multiplied by the frequency of the voltage impulses. This current is indicated by a rotating coil instrument which is calibrated directly in number of flashes per minute.

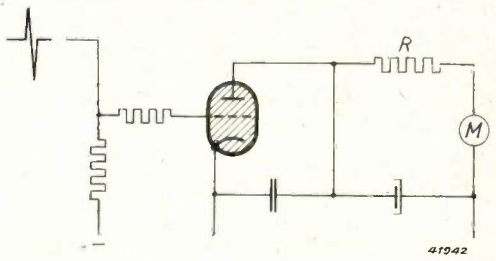


Fig. 7. Connections of the frequencymeter. The gas triode is blocked by a high negative voltage and only breaks down when a voltage impulse acts on the terminals which are connected with the output of the control generator. The meter *M* measures the number of discharges per second of the condenser.

Applications

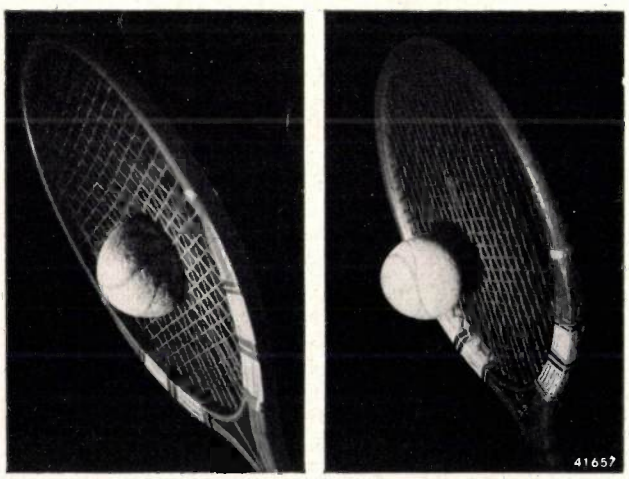
In visual observation the stroboscope is especially suitable for rapid, often orientating observations. With the help of the frequency meter the frequencies of vibrations or other periodic movements can be determined. Interesting phenomena which occur in a certain phase of the motion being investigated can be recorded photographically, while photography is also capable of furnishing information about events which occur only once.

It is often possible immediately to interpret the photographs quantitatively. In the case of the examples discussed in the following, however, we shall omit this and only use the photographs for the characterization of the object observed.

Cavitation of ships' screws

In shipbuilding it is important to adapt the dimensions and shape of the screw to the entrance velocity and the static pressure of the surrounding water in order to avoid "cavitation". Cavitation is the phenomenon which takes place on the surface of the screw when locally and temporarily such a low pressure occurs that vapour or gas bubbles occur. These bubbles may implode on the surface of the blade of the screw, which is then exposed to the water hammer effect which may result in serious erosion of the material of the screw.

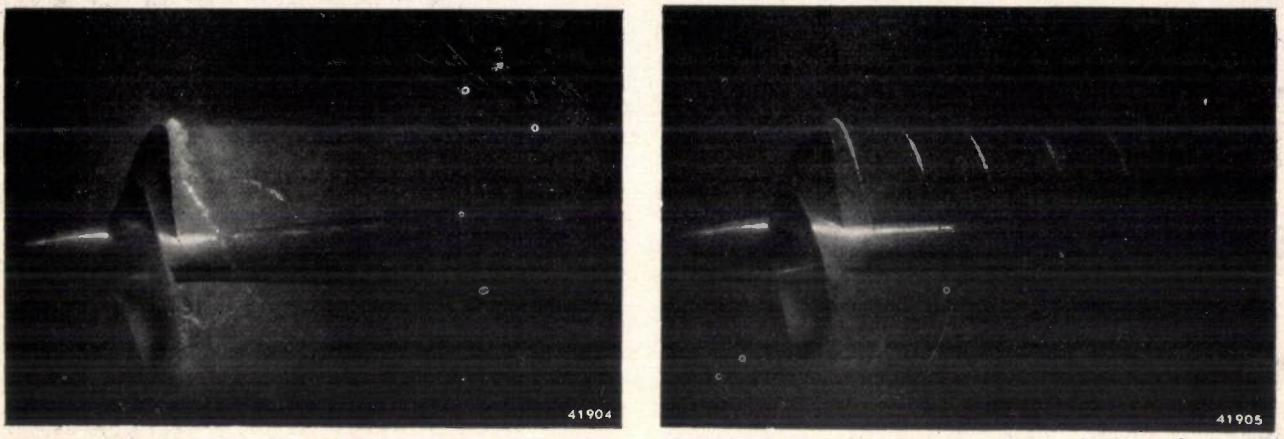
In order to study this phenomenon and to determine experimentally the best shape for the screw, models of screws are studied in a flow channel in which, by regulation of the velocity of the water, the water pressure and the number of revolutions,



a *b*

Fig. 6. Two photographs of the collision between tennis ball and racket.
a) Distance of the microphone a few decimetres;
b) Distance of the microphone several metres.

the phenomenon of cavitation can be faithfully reproduced and its influence on the driving properties of the screw can be measured. This "cavitation tank" is provided with glass observation windows through which it is possible to illuminate the model srew stoboscopically and to observe and photograph the phenomenon of cavitation. *Fig. 8a* shows a ship's screw with strong cavitation, while in fig. 8*b* a screw with no cavitation is shown. The bright spiral line seen in the latter case is the top spiral of the screw.



a *b*

Fig. 8. Ship's screw in cavitation tank. *a*) Strong cavitation, *b*) no cavitation. The bright line in *b*) indicates the top spiral of the screw. The photographs were taken in the Naval Shipbuilding Testing Station in Wageningen.

The cutting by the teeth of a planing machine

In testing a planing machine it is important to know whether all the teeth take an equal share in the cutting process. This can easily be ascertained by observing the machine in action with the help of a stroboscope lamp. *Fig. 9* is a reproduction of a photograph which was made during such an investigation. This single photograph, which shows only one of the teeth in action, is naturally insufficient to permit the drawing of the desired conclusions.

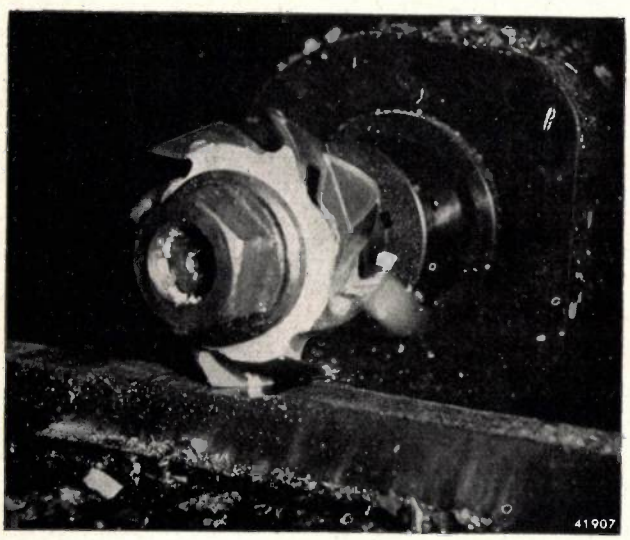


Fig. 9. Instantaneous photograph during the observation of a planing machine with the help of the stroboscope.

Investigation of internal combustion motors

In the case of high speed machines the stroboscope may be useful in checking certain parts. As an example a photograph is given in *fig. 10* of the valve

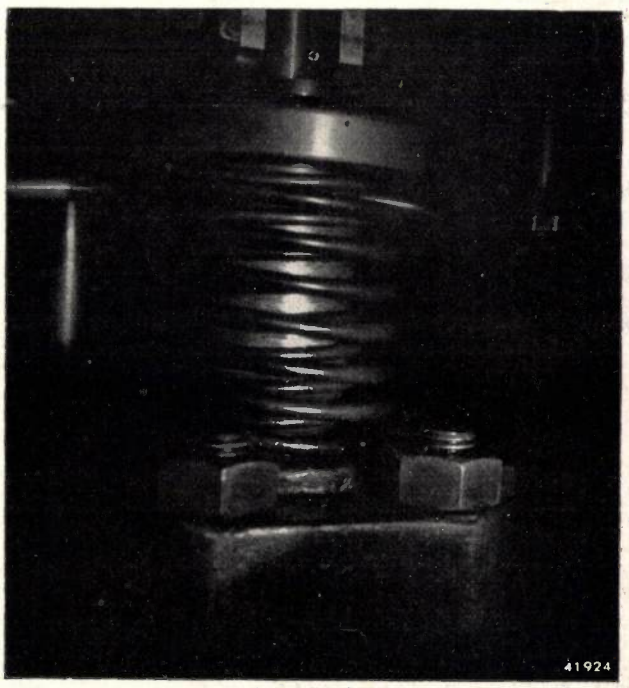


Fig. 10. Photograph during the observation of the valve springs of a Diesel motor with the help of the stroboscope.

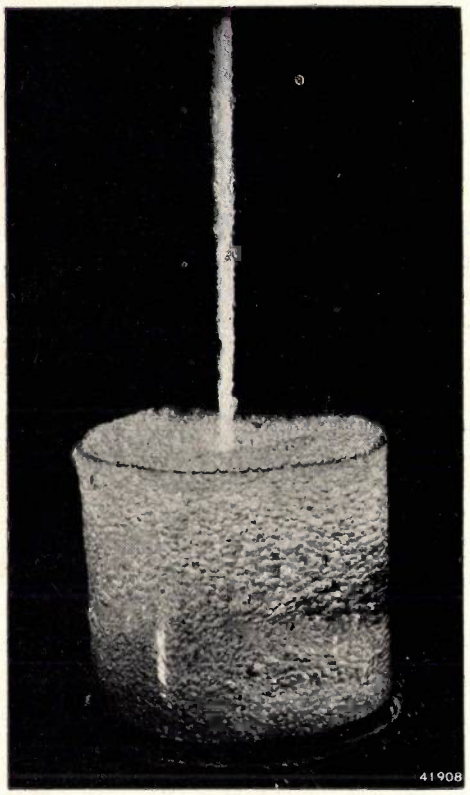


Fig. 11*a*

Fig. 11. Photographs of water jets with light flashes from the stroboscope lamp. *a*) Jet from a faucet, *b*) jet from a douche.

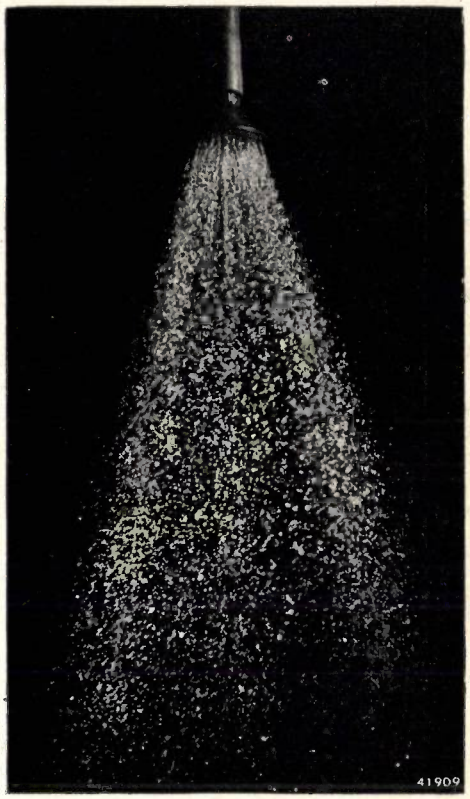


Fig. 11*b*

Fig. 12. This photograph was made for the sake of comparison with fig. 11*a* with the help of a photoflash lamp. The flash time of the photoflash lamp is found to be many times too long for a sharp photograph of the foam.

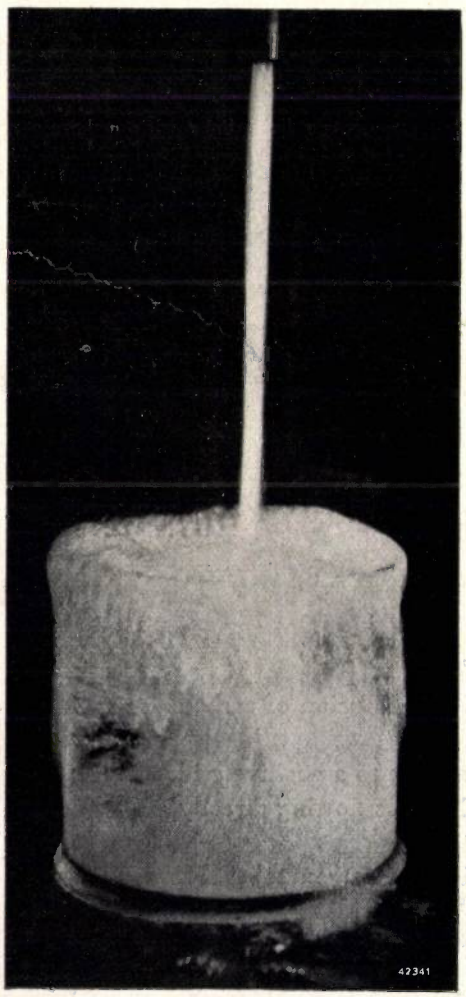


Fig. 12

springs of a Diesel motor. The stroboscopie observation of the valves springs is important, because, due to phenomena of inertia, the phase of the opening and closing of the valves may deviate at high speeds from the phase adjusted with the motor stationary.

Observation of liquids

In certain applications of the stroboscope no synchronization of the light flash is necessary and particular advantage is drawn from the short exposure time. An example of this is the investigation of water jets, illustrated in *fig. 11a* and *b* by the comparison of the jets of a faucet and a douche. It can clearly be seen that the foaming water from the faucet consists entirely of bubbles, which cannot be distinguished in the photograph reproduced in *fig. 12* taken for the sake of comparision with a photoflash lamp.

A practical application is found in the investigation of jets of liquid in the lubrication of machines for metal working. As an example in *fig. 13* a photograph of the lubrication of a centreless grinding machine is reproduced. It may be seen that the lubrication liquid is blown aside by the wind from the grindstone and thus does not flow over the grindstone as was intended. This observation led to a change in construction.

Other applications

In addition to these examples there are, of course, numerous other possible applications of the stroboscope. An extensive field is formed for example by the high-speed machines of the textile industry, such as spinning machines, looms, sewing machines, etc. A less obvious possibility of application is in

Fig. 13. Cooling of a centreless grinding machine. The lubricant is blown aside by the wind from the rotating grindstone.

material testing, where advantage can be taken of the possibility of photographing at exactly the right moment, with the help of the stroboscope lamp, the very rapid process of the breaking of a piece of work.

In order to illustrate this possibility we reproduce as a final example in *fig. 14* an electric light bulb being broken with a hammer. The synchronization in this case was mechanical, by means of a thread which gives the synchronization impulse at the moment that it is broken by the descending hammer.

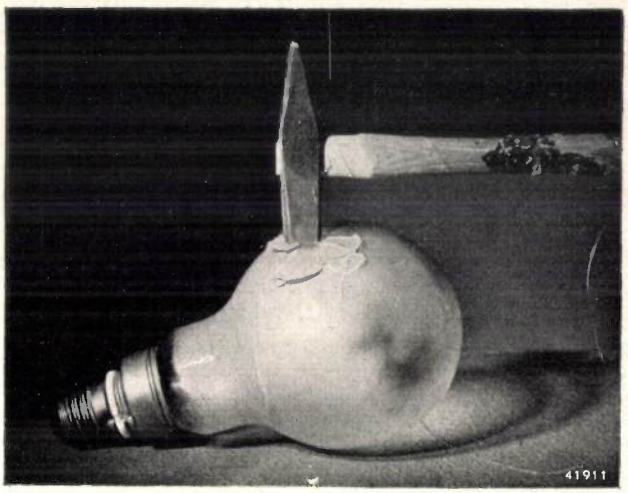


Fig. 14. Smashing an electric light bulb with a hammer. The synchronization of the photo was by means of the breaking of the thread which is fastened to the handle of the hammer.

Contents of *Philips Research Reports* 1, No. 1, 1946

VOL. 8 No. 2, p. 33-64 FEBRUARY 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

X-RAY TUBES WITH ROTATING ANODE ("ROTALIX" TUBES)

by J. A. van der TUUK.

621.386.1

X-ray tubes with a rotating anode permit a much higher specific loading of the focus than tubes with a stationary anode, a property which is of great importance for the röntgenography of moving organs, such as the heart and lungs. While, however, for photographs of heart and lungs only low tubes voltages (< 70 kV) and short loading times are required, for other diagnostical applications (especially for examinations of the stomach) the X-ray tube with rotating anode must also be able to be operated at very high tube voltages (up to 100 kV) and be loaded continuously with 150 W. A type of tube with a rotating anode is here described, which satisfies all these requirements very well. Due to the fact that the tube, which is entirely of hard glass, is immersed in oil, its dimensions and weight could be reduced considerably compared with previous constructions. The anode consists of a disc of solid tungsten, while in the original „Rotalix" tubes the anode was a block of copper with a thin layer of tungsten on its front surface. The detailed discussion of the advantages and disadvantages of the two constructions presents an opportunity of dealing with a number of typical problems connected with the rotating anode.

When in 1929 there first appeared on the market a practically useful X-ray tube with rotating anode — the "Rotalix" tube developed in the Philips laboratory [1]) — it meant a big step ahead in the quality of the X-ray photographs which could be made in the medical diagnosis of moving organs such as the heart and lungs. In the case of moving objects short exposure times are of course desired. Thus for a given blackening of the film a high X-ray intensity is required, and that requires a high specific loading of the focus of the X-ray tube, since that focus must be kept small for the sake of a small half shadow (geometrical lack of sharpness) in the X-ray picture. The specific focus loading is limited by the temperature of the focus, which rises during the time of loading and which may not exceed a certain value because of the evaporation of the anode material. Because of the fact, that, in the case of the rotating anode, each point on the surface of the anode serves as focus only for a small part of the total exposure time, the specific loading of the focus can be made considerably higher than in the case of a stationary anode. With exposure times of less than about 0,04 sec, as was discussed previously [2]), the theoretically attainable improvement in the specific loading amounts to a factor

$$p = \sqrt{2\pi rnt/f} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

r is here the radius of the path of the middle of the focus, n the number of revolutions, t the time of exposure and f the width of the focus. The practically attainable improvement factor is somewhat smaller for various reasons, than follows from the formula: with a "Rotalix" tube of the original construction, with $r = 2$ cm, $n = 2900$ r.p.m. and $f = 2$ mm, for an exposure time of 0.04 sec., $p = 8$ to 9.

The result of this improvement in the specific focus loading in the case of photography of moving organs meant approximately a reduction by one half of the lack of definition of the X-ray picture (with the same contrast), since this lack of definition is practically [3]) proportional to $\sqrt[3]{p}$, when one works under the optimum exposure conditions. Diagnos-

[1]) A. Bouwers, Verh. deutsch. Röntgen-Ges. 20, 102, 1929.

[2]) See: J. H. van der Tuuk, The "Rotalix" Tube for X-ray Diagnosis, Philips techn. Rev. 3, 272, 1938. The arguments given here very briefly will be found in more detail in that article.

[3]) See for example the explanation of the optimum exposure conditions in: A. Bouwers and G. C. E. Burger, X-ray Photography with the Camera, Philips techn. Rev. 5, 255, 1940.

tically this improvement in definition was so valuable that the photographs obtained with a tube provided with a rotating anode could be said to be of a higher class.

It would appear from formula (1) that the factor p can be made still larger, for instance by giving the anode a larger diameter and making it rotate more rapidly. However, the additional improvement, which can practically be obtained in this way is relatively small compared with that already achieved, since p increases only proportionally with $\sqrt{rn}$. A large increase in the specific loadability is not readily possible, while the necessity of a slight increase cannot at present be considered to be actual, since in order to guarantee optimum exposure conditions the photographic industry would have to increase considerably the sensitivity and the fineness of grain, respectively of the combination X-ray film plus intensifying screen [3]). It is thus understandable that the different types of tube with rotating anode which have been developed since the beginning [4]) do not aim primarily at a further increase in the specific loadability of the focus. The new constructions are characterised rather by a tendency towards greater simplicity and security of operation and towards an extension of the possibilities of application. As concerns the latter, it may be noted that the „Rotalix" tube was originally intended for the photography of heart and lungs, where it is mainly a question of a high specific focus loading, while the tube voltages may be relatively low (45 to 70 kV) and the tube is always loaded only for a short time. It soon became evident, however, that the tube with rotating anode should also be able to be used for other diagnostical purposes, especially for stomach exposures and for fluorscopy, since doctors usually prefer to carry out all those examinations with the same tube [5]). The tube therefore had to be made suitable also for the high voltages (up to 100 kV for example or even higher) necessary for stomach photography and for relatively heavy continuous loading (100 W and sometimes even 200 W), while in connection with easy adjustability for the different kinds of exposure, the dimensions and the weight of the tube had to be limited as far as possible.

In this article we shall give a description of a new „Rotalix" tube developed in recent years. It will be shown how and to what extent the above-mentioned requirements are satisfied.

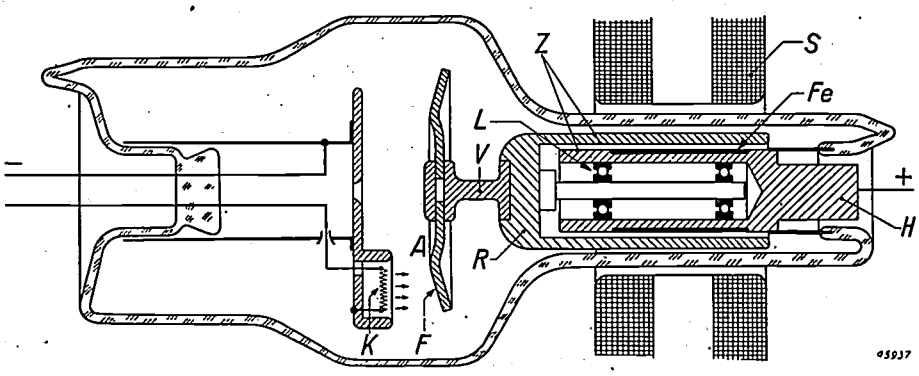


Fig. 1. Diagram of the cross section of the new „Rotalix" tube. K filament, F focus on the anode A which consists of a solid tungsten disc. The latter is fastened by means of a molybdenum connecting piece V to the copper rotor R which turns in two ball bearings L in the anode support H and is driven by the stator S. Fe is an iron ring for the purpose of concentrating the lines of force in the rotor: Z blackened surfaces for better radiating the developed heat. The anode support and the ball bearings remain practically cold, which is essential for a good working of the bearings. Behind the cathode a barium getter is applied.

Main characteristics of the new „Rotalix" tube

Fig. 1 shows a simplified cross section of the tube. The anode in this case consists of a disc of solid tungsten fastened by means of a short molybdenum connecting piece to a copper rotor. This part of the tube is surrounded externally by a stator fed with polyphase current which furnishes the motive power. In fig. 1 may also be seen the filament excentrically opposite the outer edge of the rotating anode disc. The (primary) electrons emit-

[4]) See for example A. Ungelenk, Fortschr. Röntgenstr. 49, 162, 1934.

[5]) In stomach examinations it is essential that the tube can also be used for fluorscopy. Fluorscopy and photography here follow one another immediately, as will be discussed later.

ted by the filament are focussed on the anode. The tube wall is made of a special kind of hard glass which is very resistant to the bombardment by the secondary electrons which are emitted by the anode [6]). The whole tube is housed in an earthed metal jacket which protects the operator from contact with parts carrying high voltage and helps to cut off the X-rays in undesired directions. (see *fig. 2* and *3*). The jacket is filled with oil.

When we compare this construction with that of the older "Rotalix" tubes, which have already been described [2]), there are two obvious differences. The older tubes were surrounded by air instead of by oil; and the anode had an entirely different construction. It consisted of a heavy copper block upon the front of which was fixed a thin layer of tungsten (lozenge) and whose rear side was provided with special cooling jackets.

The immersion in oil has certain important advantages [6]). Due to the high break down resistance of the oil it was possible to use relatively small insulation distances between the poles and between the latter and the earthed housing. While the older tube with housing was 83 cm long and had a maximum width of 22 cm, in the tube according to fig. 2 these dimensions could be reduced to 46 en 14 cm, respectively. The new tubes are adopted in all respects to voltages of up to 100 kV. It is also of importance that the breakdown resistance is ensured independent of atmospheric conditions. Tubes in air, which are intended for use near sea level, cannot in general be loaded with the highest voltages in the mountains (sanatoria), since at the lower atmospheric pressure the breakdown resistance of air is lower. In spite of the oil filling the total weight (17.5 kg.) of the new tube with housing is still appreciably lower than that of the tube with air insulation.

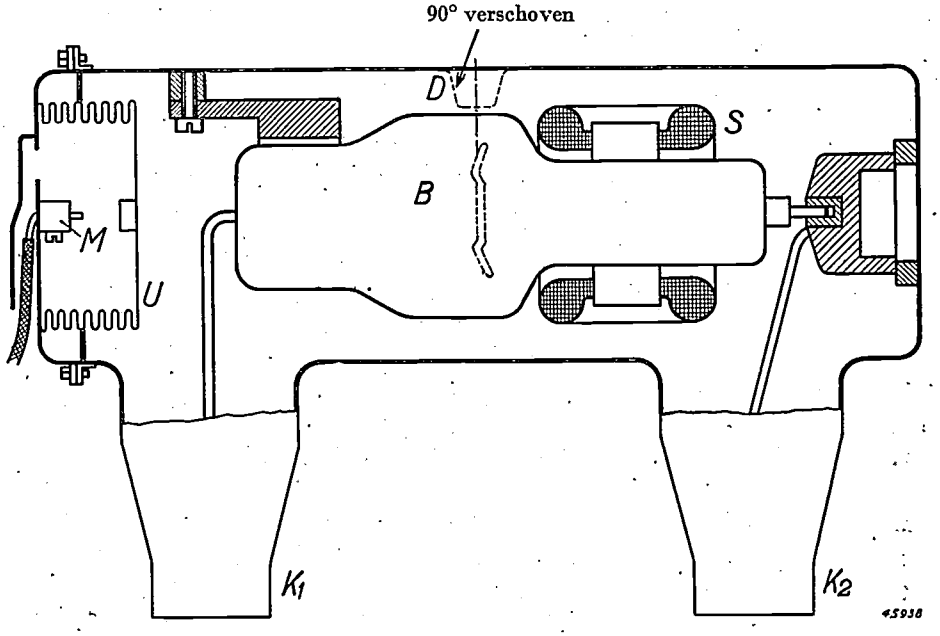


Fig. 2. The *B* tube of fig. 1 mounted in the metal housing *S* is the stator. At K_1 and K_2 are the connections for the high-voltage cables. *U* accordion shaped box for sealing off the oil-filled space. Because of this shape the oil can expand when warm without appreciably increasing the pressure. If the temperature of the oil rises above a certain value, the compression of *U* closes a switch *M* which switches off the supply voltage of the X-ray tube. *D* is a cup made of insulation material (shifted about 90° in position in the drawing) which keeps the oil away from the spot where the X-rays leave the tube. In this way the absorption of the useful X-rays by the oil is kept as small as possible. By making the tube wall and the thickness of *D* small, the filter of the tube with housing is kept small, namely less than corresponds to a thickness of 1 mm of aluminium. The housing is made of iron with a thin lead covering on the inside to prevent the passage of X-rays.

The insulation by means of oil also has advantages in the driving of the rotating anode. In order to use as little power as possible for that purpose, the „air slit" between the stator and the rotor at high voltage should be as narrow as possible. A possible solution [4]), and the one used in the case of air insulation, is to have the stator with its windings at the same high potential as the anode. A special transformer is then necessary for the feeding: one

[6]) *Cf.* also: J. H. van der Tuuk, Hard glass X-ray Tubes in Oil, Philips techn. Rev. **6**, 309, 1941.

with a secondary winding which is insulated for high voltage. Placing the tube in oil, however, makes this complication entirely unnecessary, since even with the stator earthed, the air slit can be made very narrow with no danger of flashover from the glass wall, so that only 500 W during 0,5 sec. are necessary to start the rotor. In the case of the previously described „Rotalix" tube, thanks to the practically earthed metal intermediate section, an „earthed" stator could also be employed, but in that case about 1100 W during 0,8 sec. were necessary just to reach the desired number of revolutions (because of a greater moment of inertia of the rotor and a wider air slit).

Finally the oil immersion facilitates the cooling during continuous loading to no small degree. Because of the good heat transfer by the oil, the temperature of the tube wall and the insulation material remains low enough, even though the dimensions of the tube are small. Furthermore, due to the natural flow in the oil, the heat is uniformly distributed over the whole housing so that with the relatively small dimensions of the type shown fig. 2 a power of 150 W, continuous, can be dissipated. Even for the intensive diagnosis of the stomach, this is practically always sufficient.

Let us now pass on to the second distinguishing feature of the new type mentioned above, the construction of the anode. The use of a tungsten disc instead of the combination of copper, with a thin lozenge of tungsten means a simplification in various respects. However, a comparison between the two types of anode is not complete, simply by mentioning this fact. The transition to the construction now employed involved a give and take, and it is quite comprehensible that the older construction will continue to be preferable for certain special diagnostical purposes. It seems worth while to go somewhat more deeply into the problems here concerned, especially because the specific problems of the rotating anode are thereby strikingly emphasised.

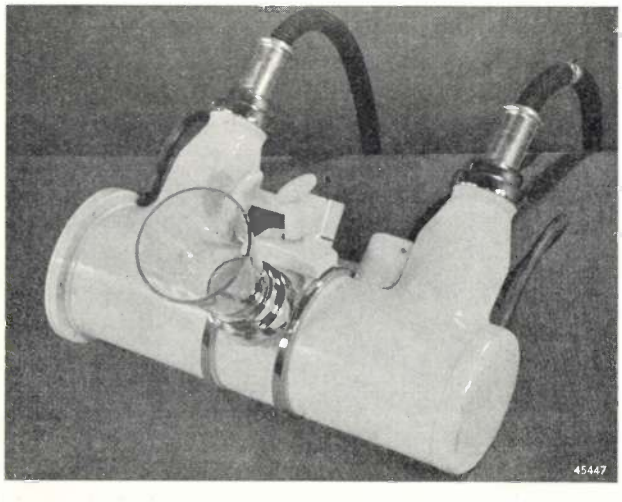


Fig. 3. The new „Rotalix"-tube in its housing. the whole having a length of 46 can. The housing is in a clamping-strip turnable around its axis. The leadglasstube, through which the X-rays pass, is visible. The housing is of special finish with built in cooling spiral, by which the power could be increased to 600 W. Besides the contacts for the high tension cables rubber tubes can be seen for the leading in and ont of the cooling water. The normal construction, intended for 150 W, does not possess a cooling spiral.

The tungsten-copper anode

The main reason, why the combination described of tungsten and copper was chosen for the anode of the original „Rotalix" tubes, was the desire to attain the highest possible specific loadability of the focus. The accelerated electrons must in any case impinge on tungsten, since that metal combines a very high emission of X-rays with a very high melting point. The highest focus temperature permissible for a short time is about 2500 °C for tungsten. When it is a question of the heaviest possible loading, it is not only that the focus must be able to stand a high temperature, but the heat developed on the focus must be conducted away, as quickly as possible, to the inner, cold parts of the anode mass. The anodematerial must therefore be a good heat conductor. Tungsten itself is not a bad conductor. Copper, however, is more than twice as good a conductor, so that the loadability will be increased when the anode material, as close as possible, behind the focus is not of tungsten but of copper.

Closer consideration shows that this reasoning it still somewhat too simple. The tungsten layer cannot be made arbitrarily thin, since the copper directly behind it must not become hotter than a maximum of 1000 °C (meltingpoint 1080 °C). The minimum required thickness of tungsten can be determined by considering the variation of temperature in a longitudinal cross section of the anode, when the extremety of the cross section lying in the front surface of the anode ($x = 0$) is exposed to the beam of electrons from the moment $t = 0$ to $t = t_1$. Here t_1 is the time, during which the anode rotates one width of focus farther; thus with the above mentioned values of r, n and f, t_1 is 0.00033 sec. At each point the temperature will first rise and then, after the loading has ceased, it will fall again as the heat is conducted farther away toward the inside of the anode. If we now plot the maximum increase in temperature T, which is reached in every separate point, we obtain, with a solid tungsten anode, a curve like that reproduced in *fig 4a*. The load is there assumed to be such,

that the maximum temperature increase of the focus itself ($x = 0$) which occurs at the moment $t = t_1$, possesses exactly the highest permissible value. It is now clear, that the anode can only consist of copper beginning at a certain depth x_1, for which the maximum temperature ever occurring is not more than 1000 °C. Since, as will appear later, we must count on a basis temperature of for example 450 °C for the copper, we can find x_1, in the figure as the point where the curve *a* has fallen to $T = 550$ °C. In that way we find that, according to this curve, the tungsten layer must extend at least to a depth of 0,2 mm. If we now draw the temperature distribution in the longitudinal cross section at the moment t, at which the loading is interrupted and the focus reaches its highest temperature, we obtain curve *b* of fig. 4. This shows that, at that moment only, very little (5%) of the total heat applied penetrated deeper than 0.2 mm into the body of the anode. The maximum temperature of the focus will therefore scarcely be affected, when the rest of the anode behind the first 0.2 mm is made of copper instead of tungsten.

We thus reach the unexpected conclusion that the employment of the copper would offer no advantage with respect to loadability. This is, however, true only when the exposure time is so short that the anode does not make more than one complete revolution during the loading, thus for exposures of less than $t_2 = 0.02$ sec. If the loading lasts longer each point on the surface of the anode is struck not once, but repeatedly by the beam of electrons. Since in the time interval between the first and the second loading the temperature of each point in the orbit of the focus does not decrease from the maximum $T_1 = 2500$ °C to the initial temperature $T_0 = 450$ °C, but only to a certain value $T_2 > T_0$, the second loading must be lower, the third still lower, etc. The advantage of the copper now becomes evident. In the time t_2, which, with the above-mentioned value of r and f, is about 60 times as long as t_1, the greater part of the heat (>80%) penetrates deeper than 0.2 mm into the anode. The temperature increase $T_2 - T_0$ remaining after the first revolution is thus practically entirely determined by the properties of the more deeply lying material, and is therefore, when copper is chosen, only half as great as with a solid tungsten anode, namely about 60°. Curves *c* in fig. 4 give the temperature distribution at the moment t_2: with solid tungsten (fullline curve) and with copper from a depth of 0.5 mm (broken-line curve).

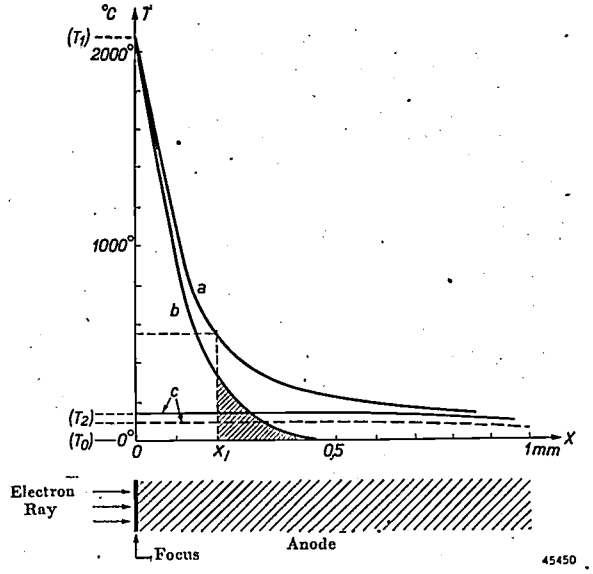


Fig. 4. Temperature distribution over a longitudinal cross section of the anode. The increase of temperature T above the basis temperature (T_0) of about 450° C is plotted as a function of the depth x beneath the focus. *a*) Maximum temperature ever occuring during or after the loading; anode of solid tungsten, loading from $t = 0$ to $t_1 = 0.00033$ sec. (*i.e.* the time during which the focus is displaced by its own width), such that the focus ($x = 0$) takes on the highest permissible temperature, ($T_1 = 2500$° C). *b*) Temperature at the moment t_1, at which the loading ceases. *c*) Temperature at the moment $t_2 = 0.02$ sec. *i.e.* after one complete revolution; full-drawn line: anode of solid tungsten, broken line: anode of copper from a depth of 0.5 mm. (All curves borrowed from W. J. Oosterkamp, Thesis, Delft 1939).

With exposure times of the order of 0.05 to for instance 3 sec., the copper in this way gives an appreciable improvement in the permissible loading. Much longer exposure times scarcely occur in the use of tubes with rotating anodes in ordinary X-ray diagnostics. It may, however, happen that a large number of successive exposures of a moving object must be taken at short intervals (röntgen cinematography), which amounts to total loading times of for example 20 sec. or more. With such long times, the heat developed on the front surface of the anode penetrates not only to the inside, but even to the rear side of the anode, in other words, the necessarily limited dimensions and the consequently limited heat capacity of the anode now begin to make themselves felt. The residual temperature of the focus, which steadily increases with increasing loading time, will therefore increase even more rapidly and the permissible loading will decrease even more rapidly. If, in order to be able to use the tube for cinematography, it is desired to make the heat capacity of the anode fairly large, it is therefore fundamentally of advantage to use copper. Copper is much cheaper than tungsten and, for the same heat capacity, only 3/4 of the volume and 1/3 of the weight of copper is needed.

[7]) Those considerations are borrowed from: W. J. Oosterkamp, thesis Delft, 1939.

If we go a step farther with the loading time we enter the region of fluorscopy, where we may speak of continuous loading. The power which can here be applied is no longer determined by the focus temperature but by the temperature taken on by the anode as a whole [8]). It is now found that, when copper is used, there are definite limitations: for a copper anode covered with tungsten the permissible temperature lies much lower than for a solid tungsten anode. While for the copper immediately behind the lozenge it amounted to about 1000 °C for a very short exposure in the case of continuous loading of the whole anode mass a temperature of not higher than about 450 °C (the basis temperature assumed above) is permissible, since at higher temperatures all kinds of undesired things take place: recrystallisation phenomena occur in the copper, the mechanical strength decreases, and the requirements are fairly high in that respect because of the rapid rotation, too much vaporized copper enters the tube (the vapour pressure of copper at 850 °C is already 10^{-5} mm; the vapour pressure of the copper oxide from various sources is even noticeable at 500 °C) and finally traces of residual gases may be freed from the anode. The latter may be ascribed to the fact that, during the evacuation and outgassing of the tube, the temperature may not be raised too much because of this very restriction in the permissible temperature of the copper. Now since the dissipation of heat by the anode, which rotates in a vacuum, can only take place by radiation, and since the radiation increases with the fourth or fifth power of the temperature, the restriction of the basis temperature of the anode to 450 °C makes it difficult to reach an adequate power for fluorscopy. Only by employing the fairly complicated construction, which has been described in the articles already cited [2]), and which consists in an artificial enlargement of the radiating surface of the anode and an improvement of its emissivity, was it possible to increase the heat dissipation and thereby the fluorscopic power to reasonable value.

The combination of tungsten with copper was accompanied by another complication. The brief, very high specific loadings result in corresponding high temperature gradients and mechanical tensions in the rotating anode. These make necessary a special solution of the problem of joining the tungsten lozenge to the copper (as was also discussed in the article cited) a solution which made very heavy demands on construction and fabrication.

The tungsten anode

The simplifications resulting from the use of tungsten alone for the anode are obvious. No precautions are necessary to keep the basis temperature of the anode low with continuous loading. The anode can be rigorously outgassed during fabrication, which promotes particularly the safety from high voltage of the X-ray tube[9]). The problem of attaching the tungsten lozenge no longer exists.

From the foregoing it will have become clear that certain sacrifices must be made for these simplifications. We shall discuss these in somewhat more detail.

With exposure times shorter than the period of one revolution of the anode (< 0.02 sec.) the loadability of a tungsten anode, as we have seen above, will not be appreciably smaller than that of a tungsten-copper anode of the same radius and speed of revolution. With longer exposure times where the permissible loading falls as a result of the residual temperature upon repeated loading of the focus, the tungsten anode cannot compare with the tungsten copper anode. With for instance 0.1 sec. the loadability would amount to only about 80% percent of that of the tungsten-copper anode, as calculation will show. Since exposure times of this order of magnitude are in practice the most important, the tungsten anode has been given a larger radius ($r = 3.0$ cm); the loadability, which according to equation (1) increases proportionally with $\sqrt{r}$ (this is approximately valid also for times of the order of 0.1 sec. although strictly speaking equation (1) is no longer valid here), is therefore the same for the two types of tube for exposures of from 0.1 sec. to about 2 or 3 sec. For shorter exposure times the tungsten anode, because of its larger radius, is now somewhat at an advantage, as may be seen from *fig. 5*.

It must also be mentioned that the transverse dimensions of the tube with housing, inspite of the larger diameter of the anode in the new „Rotalix" tube, are not larger, but are even smaller than those of the old tube, thanks to the employment of oil insulation.

When we consider the further course of the curves in fig. 5 we see that the tungsten anode, which is now equal to the tungsten copper anode (of smaller diameter) from about 0.1 sec. to about 3 sec. is

[8]) See the communication of G. C. E. Burger, Hand. Ned. Nat. Gen. Congr. 1943, p. 286.

[9]) It may be mentioned that this has aided to make possible the driving of the „Rotalix"-tube on A.C. We hope to return on this question at another opportunity.

much inferior at still longer exposure times. This is due to the much lower heat capacity of the tungsten anode. The tungsten copper anode of the old "Rotalix" has a weight of more than 1 kg, for the same heat capacity the tungsten anode would, as stated above, need to be three times as heavy. Apart from the fact that the employment of 3 kg. of tungsten is unnecessary for our purpose and would accord very ill with the aspiration toward limitation of expense and weight, the present position of technology would make it practically impossible to make a block of tungsten of that weight with the desired good mechanical properties. The high melting point which among other properties renders tungsten so particulary suitable as material for anti cathodes, however, also makes it impossible to prepare massive tungsten objects by the ordinary metallurgical processes, namely fusing and casting. The methods of powder metallurgy must here be applied [10]). Pure tungsten in powder form, prepared chemically, is pressed into a rod which is then heated to a high temperature for some time (presintered and sintered). After this sintering, which is accomplished by heating with an electric current conducted by the rod, the ductility necessary for further working and application is obtained by means of a mechanical treatment. Finally the material can be rolled to a sheet from which the anode disc can be cut. The mass of the disc thus obtained is therefore determined by the amount of tungsten powder which is originally used for pressing and sintering. The amount which can be treated at one time is limited for various reasons,

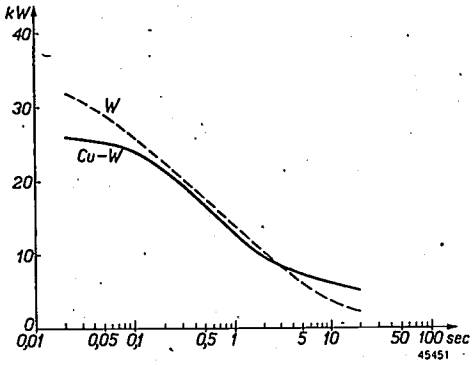


Fig. 5. Loadability (in kW) as a function of the loading time for a tungsten-copper anode (full-drawn line) and a tungsten anode (broken line). Loaded with pulsating direct current obtained with four values. The diameter of the anode at the middle of the focus is 4 cm with the tungsten-copper anode and 7 cm with the tungsten anode. The basis temperature is 450° C in the first case and 800 °C in the second. The speed of revolution is 2900 r.p. m. in both cases, the width of focus 2 mm.

[10]) See: J. D. Fast, The preparation of Metals in a Compact Form by Pressing and Sintering, Philips techn. Rev. 4, 309, 1939.

for example because of the fact that with increasing mass the very brittle presintered rods become increasingly difficult to handle and further the fact that the currents required for sintering become extremely large, etc. In the methods of fabrication employed by us the finished tungsten discs have a weight of about 225 g. We shall return to this subject later.

The heat capacity of the tungsten anode is therefore only 1/12 of that of the copper anode. On the other hand, however, the temperature of the tungsten disc may be much higher than that of the copper, but this advantage does not compensate for the much smaller heat capacity. In this way therefore the loadability of the tungsten anode for very long exposure times (röntgencinematography) is much lower than that of the tungsten-copper anode.

Although the small mass of the tungsten anode is indubitably a disadvantage in this respect it may not be forgotten that in all kinds of other respects it is an advantage, particularly in connection with the bearings and driving of the anode. Since the tungsten anode together with the copper rotor only weighs about half as much as the tungsten-copper anode of the old tubes, it is easier to limit sufficiently the friction and wear on the ball bearings on which the anode turns [11]). Furthermore, since the moment of inertia of the anode is only about 1/3 of that of the tungsten-copper anode it reaches the required speed of revolution with the same driving power much more quickly after being started. This is important for stomach examinations where the doctor uses fluorscopy and requires immediately to fix the picture photographically at a definite moment. Since fluorscopy takes place as a rule with stationary anode or at least with the anode not rotating at full speed, the necessary delay in photographing is determined by the time necessary for the anode to obtain full speed and for the heating of the filament which operates at a lower temperature for fluorscopy (lower tube current of for instance 2 tot 3 mA) than for photography. Reaching full speed from a state of rest, which took 0.8 sec. for the tungsten-copper anode, easily takes place in less than 0.5 sec. with the tungsten anode, thanks to the smaller moment of inertia. In order to profit by this, of course, the time necessary for

[11]) The ball bearings are lubricated with a soft metal, in the Philips tubes lead is used. The vapour pressure of lead is extremely low so that the high vacuum in the tube remains so complete that the extremely high requirements of reproducibility demanded by modern diagnosis with automatic tube-current regulation are satisfied (see H. A. G. Hazeu aud J. M. Lederboer. A universal Apparatus for X-ray Diagnosis, Philips tcchn. Rev. 6, 12, 1941.)

heating the filament had to be reduced to 0.5 sec. also, and this was achieved by making the filament longer and thinner, since the times, necessary to become hot, are proportional to the diameters of the filaments, given the same power.

Let us finally consider the employment of the tungsten anode for fluorscopy. We have already pointed out, that it is no longer necessary to keep the basis temperature of the anode low. During fluorscopy, there is no objection to the anode becoming red hot. The heat radiation of the anode then becomes so great that, without any further artificial means such as increasing the radiating surface, a very considerable power can be dissipated. In the case of the tube, shown in fig. 1, with a continuously applied power of 200 W the tungsten anode takes on a temperature of 800 °C, while tungsten can be made still much hotter without objection, especially since in the fabrication of the tube it was outgassed at a considerably higher temperature. The fluorscopic power is therefore actually not determined by the anode temperature, but by the temperature taken on by the metal housing of the tube, which must finally give of the heat it receives to its surroundings. If it is desired, that the housing shall not become hotter than for instance 60 to 70 °C, so that it can be touched without burning the fingers, the fluorscopic power for the very small housing according to fig. 2 is limited to about 150 W, as stated above.

We have already pointed out that, with a higher basis temperature, the margin to the highest permissible focus temperature becomes smaller. Therefore, if immediately after fluorscopy one passes over to photography, as is the rule in stomach examinations, mentioned above, and is generally assumed as a basis for the determination of the loadability of every X-ray tube, the hot tungsten anode is at a disadvantage, compared with the cooler copper anode. This fact has been taken into account in making the tungsten-anode tube the equivalent of the old „Rotalix" tube by a corresponding extra increase of the radius of the anode. The loadability plotted in fig. 5 holds for a tungsten copper anode with the basis temperature 450 °C and a tungsten anode with the basis temperature 800 °C[12]).

It is of course unavailable that, when the tungsten disc is red hot, a certain flow of heat will also take place along the axle bearing the disc toward the rear and to the copper rotor (see fig. 1). Care must thus be taken, that the rotor does not become too hot. To that end the heat transference by radiation from the turning rotor to the stationary anode support is increased by providing good radiating surfaces (blackened). An attempt might also be made to keep the flow of heat as small as possible, by making the axle long and thin and chosing as material a poor heat conductor. However, one is not absolutely free in this respect, as on the one hand the axle must be sufficiently stiff, *i.e.* sufficiently short and thick, to transmit the driving couple and not to fall into transverse vibrations, on the other hand also the axle (at least the front end of it) must be able to be outgassed at a high temperature. A not too high heat resistance of the axle has the advantage, that the heat capacity of the copper rotor can contribute to the improvement of the loadability of the anode in exposures with very long exposure times.

Prospects for further development

Summarising, we may say that the combination tungsten-copper offers indisputable advantages when it is desired to take full advantage of the gain involved in the principle of the rotating anode; that, however, the employment of tungsten alone makes it easier, as far as fabrication is concerned, to satisfy the requirements of general use in normal X-ray diagnosis. Only for those special purposes, where exposure times longer than a few seconds are necessary, namely for röntgencinematography, is the tungsten-copper anode so plainly superior, that it is clear, that in this field it will certainly maintain its usefulness, perhaps in an improved form. In the meantime the development of the tungsten anode is by no means at an end.

In spite of the fact that, as has already been explained, there can be no question of its competing with the tungsten-copper anode as far as heat capacity is concerned, it is quite possible to achieve a certain progress in that direction and thus to improve considerably the loadability of the tube for long exposure times. One path in that direction is the working out of methods of sintering larger tungsten discs. Recently indeed we have succeeded in making tungsten discs of more than 350 g instead of the abovementioned weight of 225 g. The loadability for long exposure times is increased by more than 50 percent in this way. A tube with such an anode is now at least the equal of tube provided with a tungsten - copper - anode of 1 kg, up to about 8 sec.

Finally, we should like to consider the possibility that a relatively slight increase of the specific focus loading, at very short exposure times, will

[12]) It is also possible, as is sometimes done, to lower the basis temperature somewhat by making the tungsten radiate more easily, for instance by making it black.

become desirable. Although we stated in the introduction, that at present there is no great need of this, further development might rapidly alter this fact. The possibilities of increasing the loadability at very short times were briefly indicated in the introduction: increase of the diameter and of the speed of revolution of the anode. As to the first, that trump has already been pratically played in passing over from the tungsten-copper to the tungsten anode, since we gave the tungsten disc a larger diameter to compensate for the poorer heat conduction and the smaller margin between basis and focus temperature. As to the second, the very much lighter tungsten disc is again very much in the advantage, since it will be much sooner possible, in that case, to increase appreciably the already high speed of revolution (about 2900 r.p.m.).

FREQUENCY MODULATION

by Th. J. WEIJERS.

621.396.619

After a short discussion of the ways in which a carrier wave of high frequency can be modulated, for instance with an audible frequency, so that the latter is made suitable for radio transmission, the method of frequency modulation is explained more fully. The principles of the necessary transmitter and receiver are described. Finally special attention is devoted to the so-called non-quasi stationary phenomena and their part in frequency modulation. It must be kept well in mind, that for the frequency in the different expressions it is not permissible to use simply the momentary frequency.

An audible sound can never be transmitted by means of radio by being converted directly into electromagnetic oscillations, because much higher frequencies are required for radio transmission. It is thus necessary to transmit a high frequency oscillation as carrier wave and to impress upon this wave, in some way or another, the low-frequency rhythm of the sound to be transmitted. This is generally done, by modulating the amplitude of the high frequency oscillation according to the low-frequency sound rytbm, and one then speaks of amplitude modulation. Even at the beginning of the development of practical radio-transmission, however, it was a well known fact, that, in principle, modulation could also be reali ed, by keeping the amplitude of the high-frequency carrier wave constant and varying the frequency according to the low-frequency rhythm to be transmitted. It was at first thought, that in such a system of so-called frequency modulation, there would be the advantage of being able to transmit signals with a smaller band width in the high-frequency region, so that more transmitters could be included in the same frequency region of carrier waves.

However, after Carson [1]) had shown in 1922 that this was incorrect, no more serious attempts were made in the following years to apply a system of frequency modulation for radio broadcasting.

New life was lent to attempts in this direction in 1936, when Armstrong [2]) showed, that frequency modulation can offer the advantage, that the influence of disturbances on the reception can be more strongly reduced, than is the case with the customary amplitude modulation. Since then frequency modulation has indeed aroused a steadily increasing interest and the characteristics and practical possibilities of this system have been thoroughly investigated. As we intend to deal often in the future with problems, which are connected with frequency modulation, the principle of frequency modulation will be further explained in this article, and then the characteristics will be discussed of the transmitters and receivers for such a system. A separate section will be devoted to the rôle of the non-quasi-stationary phenomena in frequency modulation. In a later article the two systems of amplitude and frequency modulation will be compared with respect to their degree of suitability for different applications.

The modulation of oscillations

A simple harmonic oscillation, such as is shown in *fig. 1a*, can be represented by:

$$V = A \cos(\omega_0 t + \varphi), \quad \ldots\ldots \quad (1)$$

where A, ω_0 and φ are independent of t. A is called the amplitude and $\omega_0 t + \varphi$ the phase of the oscillation. In order to make possible radio transmission of this oscillation, it is necessary, that the frequency $f_0 = \omega_0/2\pi$ should be much higher, than in the audible region. The transmission by radio of an oscil-

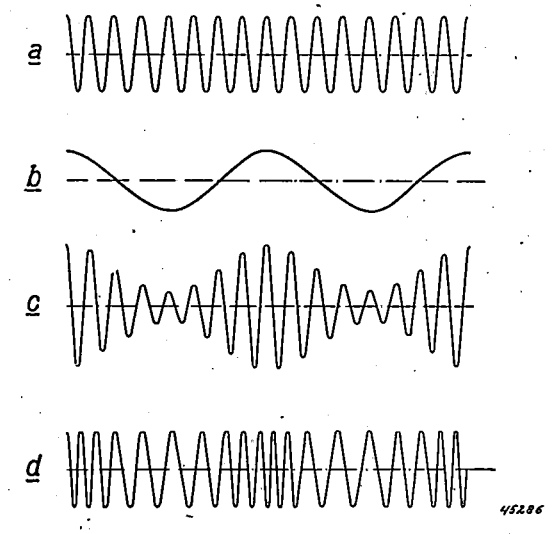


Fig. 1. *a*) Unmodulated high-frequency oscillation.
b) Modulating low-frequency oscillation.
c) Amplitude-modulated oscillation.
d) Frequency-modulated oscillation.

[1]) J. R. Carson, Notes on the theory of modulation, Proc. I. R. E. 10, 57. 1922.

[2]) E. H. Armstrong, A method of reducing disturbances in radio signaling by a system of frequency modulation, Proc. I. R. E. 24, 689, 1936.

lation with an audible frequency can thus never take place directly. A high-frequency oscillation is distorted in the rhythm of the low-frequency oscillation to be transmitted, in such a way, that radio transmission remains possible, *i.e.* that the resulting oscillation may be considered as the sum of several sinusoïdal components with sufficiently high frequencies lying in a relatively sufficiently narrow frequency region. In principle this can be done in two different ways: the amplitude A can be distorted and amplitude modulation is the result, or the phase $\omega_0 t + \varphi$ can be distorted and phase modulation results [3]). As will be seen later, frequency modulation is a special case of this.

For the sake of simplicity, we shall assume for the time being, that it is desired to transmit a single sinusoidal low-frequency vibration $B \cos qt$, as drawn in fig. 1*b*. Thus the amplitude B and the phase qt of this vibration must be transmitted.

When this is done by means of amplitude modulation, the constant A in (1) is replaced by $A\ (1 + m \cos qt)$, m being proportional to B. In order to avoid distortion upon detection in the receiver, care is taken that m remains smaller than 1.

The vibration to be transmitted is represented by

$$V = A\ (1 + m \cos qt) \cos \omega_0 t \quad . \quad . \quad . \quad (2)$$

The phase constant φ in (1) is of no importance here, and is therefore omitted in (2) for the sake of simplicity. This vibration is reproduced in *fig. 1c*. The varying amplitude of this curve is, in amplitude as well as in phase, an image of the vibration to be transmitted $B \cos qt$ (curve *b*). It must here be noted, that the zero points, *i.e.* the moments, when the vibration passes through zero, are not affected by the modulation.

Equation (2) can be resolved into three sinusoïdal components:

$$A\ (1 + m \cos qt) \cos \omega_0 t = A \cos \omega_0 t + \\ + A \frac{m}{2} \cos (\omega_0 + q)t + A \frac{m}{2} \cos (\omega_0 - q)t.$$

$A \cos \omega_0 t$ is called the carrier wave, the two other terms the side-bands. The frequencies of these three components may satisfy the above conditions for radio transmission; the frequencies $\omega_0 - q$, ω_0 and $\omega_0 + q$ are high enough and lie in a frequency region which is relatively narrow enough when ω_0 is chosen sufficiently large. By relative width of the frequency region in this case is meant the quotient $2q/\omega_0$.

[3]) In principle, both amplitude and phase can be deformed, but little or no practical use is made of this possibility and we shall not discuss it here.

The three components in question can be represented in a vector diagram (*fig. 2*). The carrier

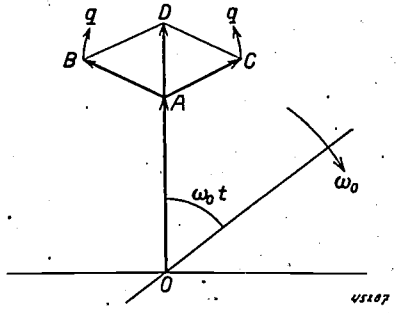


Fig. 2. Vector diagram of a sinusoïdal amplitude-modulated signal OD. OA is the high-frequency carrier wave with the angular frequency ω_0. AB and AC represent the side-band oscillation with the angular frequencies $\omega_0 - q$ and $\omega_0 + q$.

wave $A \cos \omega_0 t$ is represented by the stationary vector OA, the momentary value of which is equal to the projection of this vector on the line, which rotates around O with an angular frequency of ω_0; the two other components are represented by the vectors AB and AC, which rotate about the point A in opposite directions with an angular velocity of q. The amplitude modulated oscillation (2) is represented by the sum of these three vectors, *i.e.* by the stationary vector OD of varying length.

Such an oscillation can be generated in a transmitter, consisting of an oscillator with an amplifier. The oscillator voltage $A \cos \omega_0 t$ is applied to the grid of the amplifier valve. The amplification of this valve varies proportionally with the grid D C voltage or with the anode D C voltage. If one of these voltages is varied proportionally with $1 + m \cos qt$, the transmitter produces the required vibration proportional to $A\ (1 + m \cos qt) \cos \omega_0 t$.

One then speaks of grid or anode modulation, respectively, according to the method by which the amplitude-modulated signal is reproduced in the transmitter.

In the receiver, this oscillation is applied to a detector, for example a diode detector with leakage resistance. When an unmodulated oscillation, with a sufficiently large amplitude, is applied to this detector, a D C voltage occurs over the leakage resistance, which is approximately equal to the amplitude of the oscillation supplied. The frequency of the applied oscillation has practically no effect in this process of detection, provided it is sufficiently high. If the amplitude of the oscillation changes slowly, the voltage over the leakage resistance changes correspondingly. If the amplitude of the vibration is $A\ (1 + m \cos qt)$, the voltage given by the detector is $A\ (1 + m \cos qt)$ and the desired low-frequency oscillation $A\ m \cos qt$ is obtained.

The D C voltage component A is of no interest, it is kept away from the following amplifier valve by a condenser.

If it is desired, to transmit the low-frequency oscillation $B \cos qt$ by means of phase modulation, the part φ of the phase in (1) is replaced by $m \sin qt$, so that (1) becomes [4])

$$V = A \cos (\omega_0 t + m \sin qt) \quad . \quad . \quad . \quad . \quad (3)$$

Like the amplitude-modulated oscillation, this oscillation also does not change sinusoïdally with the time, and thus has no definite frequency. It is, however, possible to speak of a varying instantaneous frequency. To define this instantaneous frequency, we consider a non-sinusoïdal oscillation:

$$y_1 = A \cos \{ f(t) \}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where $f(t)$ changes arbitrarily with the time, and we look for a sinusoïdal oscillation:

$$y_2 = B \cos (\omega t + \varphi), \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

which corresponds, as nearly as possible for the value of t considered, with the oscillation y, y_2 is then called the instantaneaus frequency of y_1. What is to be understood by „corresponds as nearly as possible", is to a certain extent arbitrary.

Of the oscillation y_2, one can choose the amplitude B, the frequency ω and the phase angle φ suitably; thus the oscillation can be made to satisfy these conditions. The most suitable choice is obtained, when one lets $A = B$, $y_1 = y_2$ and $\frac{dy_1}{dt} = \frac{dy_2}{dt}$ and then calculates B, ω and φ. One then finds:

$$\omega = \frac{df(t)}{dt} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

and this frequency is defined [5]) as the instantaneaus frequency ω_m of y_1. When this definition is applied to the phase-modulated oscillation (3), one finds for the instantaneaus frequency:

$$\omega_m = \omega_0 + mq \cos qt = \omega_0 + \Delta\omega \cos qt. \quad (7)$$

This varies therefore between $\omega_0 - \Delta\omega$ and $\omega_0 + \Delta\omega$. $\Delta\omega = mq$ is called the frequency sweep and $m = \Delta\omega/q$ the modulation index. Fig. 1*d* represents the phase-modulated oscillation (3). The amplitude does not change due to the modulation, but the zero points are shifted: the latter now lie sometimes closer together and sometimes farther apart than in the unmodulated oscillation.

[4]) The reason why sinqt and not cosqt is used in (3) will appear later.

[5]) The same definition is used by Helmholtz for acoustic vibrations.

The oscillation (5) is drawn in the form of a broken line curve in *fig. 3* for the vicinity of one point. It is a sine curve (b) with the same amplitude

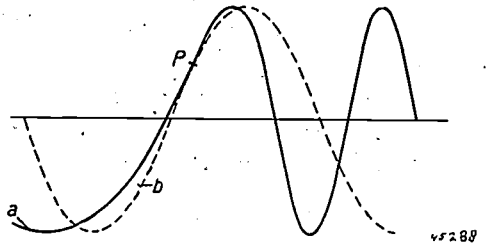


Fig. 3. Curve a represents a phase-modulated signal, while b is the sinusoïdal oscillation whose frequency is equal to the instantaneous frequency ω_m of a at point P.

as the modulated oscillation (a), it passes through the point P in question and possesses the same tangent at this point.

Conversely, one may reason as follows. From (6) follows: $f(t) = f\omega dt$. An oscillation with an instantaneous frequency $\omega_m(t)$ has the form $y = A \cos f\omega\, dt$. If it is desired to tranmit a low frequency oscillation $B \cos qt$, this can be done by an oscillation, whose instantaneous frequency is $\omega_0 + \Delta\omega \cos qt$, $\Delta\omega$ being proportional to B. This oscillation is thus $y = \cos \int (\omega_0 + \Delta\omega \cos qt)\, dt = \cos (\omega_0 t + \Delta\omega/q \sin qt)$.

This shows, why sinqt was taken in (3) and not cosqt.

In the transmission of speech or music, it is not a single frequency, but a whole frequency region, which must be transmitted. The modulation system (transmitter and receiver) can now be so arranged, that the index m is independent of the frequency q of the modulating oscillation. Then the frequency sweep $\Delta\omega = qm$ is proportional to this frequency. Then one usually speaks of phase modulation in the narrower sense. It can, however, also be arranged so that not m, but the frequency sweep $\Delta\omega$ is proportional to the amplitude and independent of the frequency q of the modulating oscillation. The index $m = \Delta\omega/q$ is then inversely proportional to this frequency. One then speaks of frequency modulation. It has been found, theoretically as well as practically, that frequency modulation is preferable to phase modulation in the narrower sense. We shall therefore concern ourselves further only with frequency modulation.

Frequency modulation

A sinusoïdal frequency-modulated oscillation can be resolved into sinusoïdal components. In contrast to amplitude modulation, where this resolution gives one carrier wave and two side-band

frequencies, in frequency modulation an infinite series of oscillations is obtained, in which the frequencies of the successive components differ by an amount q and lie symmetrical with respect to the frequency of the unmodulated oscillation, which in this case also called the carrier wave.

Fig. 4 gives an idea of the amplitudes and frequencies of the components for different values of the modulation index m. It will be noted here, that although the instantaneous frequency ω passes continuously through the whole frequency region from $\omega_0 - \Delta\omega$ to $\omega_0 + \Delta\omega$, only a finite number of sinusoïdal components occur in this frequency region; and, moreover, that components occur, whose frequencies lie outside the region, which is covered by the instantaneous frequency. With a constant frequency sweep $\Delta\omega$ the region covered by the instantaneous frequency is more densely occupied by components, the smaller q is, *i.e.* the larger the modulation index m. This is not, however, surprising,

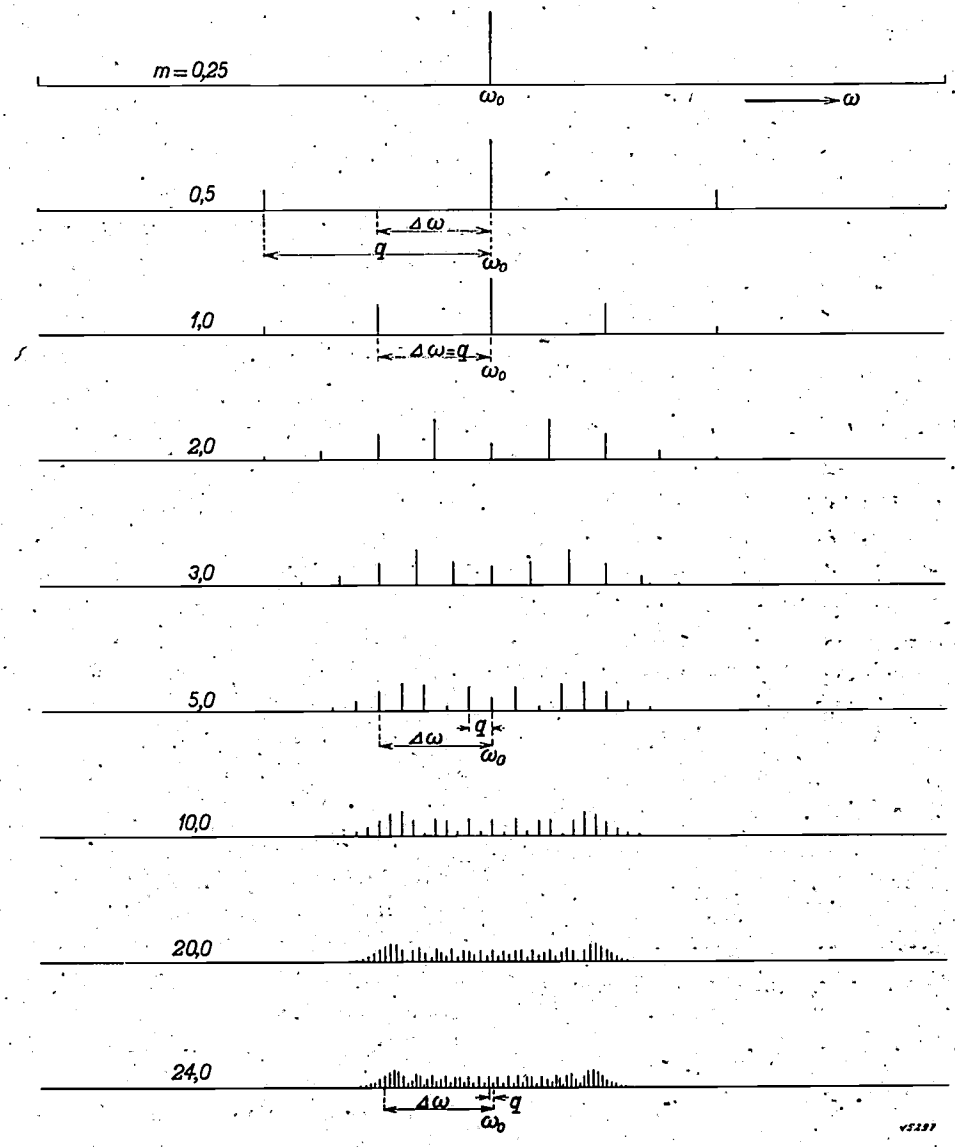


Fig. 4. Frequcncy spectra of frequency-modulated signals[6]) with different modulation indices $m = \Delta w/2$ and constant frequency sweep Δw.

when it is kept in mind, that the concept of instantaneous frequency is defined somewhat arbitrarily. Furthermore, it is clear from fig. 4 that, when $m > 1$, and thus $\Delta\omega > q$, the amplitudes of the components, whose frequencies differ more than about 3/2 $\Delta\omega$ from the carrier frequency are very small and may, be neglected practically. If $m < 1$ and thus $\Delta\omega < q$, the first, at most the first two components on either

[6]) Balth. van der Pol, Proc. J. R. S. 18, 1201, 1930.

side of ω_0 are of importance. In this case, however, no components at all, caused by the modulation, lie in the region covered by the instantaneous frequency. If m is small (for instance $m < 0.25$), the signal is practically the same, as the carrier wave with two side-band frequencies. These three components are represented in *fig. 5* by vectors in an ana-

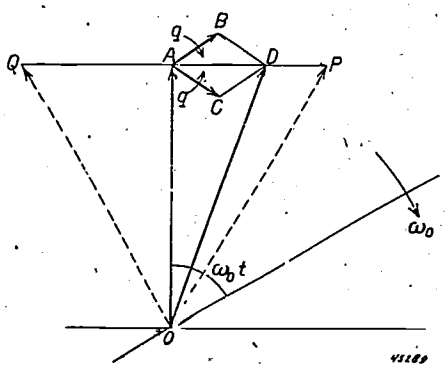


Fig. 5. Approximate vector diagram for a sinusoidal frequency-modulates signal whose modulation index m is a maximum of $^1/_4$. OA is the carrier wave with frequencies ω_0; AB and QC side-bands with frequencies $\omega_0 - q$ and $\omega_0 + q$; OD modulated signal.

logous manner to that in fig. 2 for amplitude modulation, but there is the difference, that the components of the side-band frequencies are rotated through an angle of $\pi/2$ with respect to the carrier wave. The sum of these three vectors is a vector OD, whose extremety moves along the straight line PQ, which is perpendicular to the vector OA of the carrier wave. The amplitude of the vector is thus not entirely constant. This is due to the disregarding of the other side-band frequencies. If on either side of the carrier wave an additional side-band frequency is included in the calculations, which can be represented by two smaller vectors, which rotate at double the angular velocity of the first two side-band frequencies, it may be seen from *fig. 6* that the sum of these five vectors gives a better approximation of the true vector, which, maintaining a constant length, fluctuates between the positions OT and OU, and which is obtained by including all the side-band frequencies (an infinite number) in the calculation.

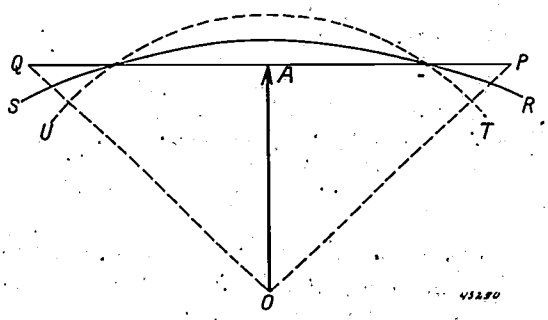


Fig. 6. The vector OA represents the carrier wave. On the straight line $PQ \perp OA$ lies the end point of the vector, which represents the sum of the carrier wave and one set of side-bands. On the curve RS lie the end points of the vectors, which represent the sum of the carrier wave and two sets of side-bands. On the arc of the circle TU lie the ends of the vectors, which represent the sum of the carrier wave and all (an infinite number) of side-bands.

In amplitude modulation a band width is necessary, which is equal to twice the highest audio frequency to be transmitted; in frequency modulation, for ideal transmission, strictly speaking one needs the whole infinite frequency region, practically, however, a frequency region equal to about three times the frequency sweep is necessary, if $\Delta\omega > q$, and equal to $2q$, if $\Delta\omega < q$.

Thus by the employment of frequency modulation no space in the frequency spectrum is saved, as was previously thought, but on the contrary more space is needed!

Transmitter for frequency modulation

A transmitter for frequency modulation can be arranged in different ways. Armstrong set about it as follows. An oscillator controlled by a crystal furnishes the carrier wave. This carrier wave is amplitude-modulated, by the low-frequency oscillation to be transmitted, in a push-pull modulator, so arranged, that it furnishes only the two side bands without the carrier wave. The carrier wave is then added again with a phase rotation of $\pi/2$ and a large amplitude, so that an oscillation like that of fig. 5 is formed. Care is taken that the angle $POA < \pi/12$, so that the amplitude variation remains small enough to be disregarded. The argument of $\cos(\omega_0 t + m \sin qt)$ represented by the varying vector OD, changes between $\omega_0 t + m$ and $\omega_0 t - m$. From this follows, that $\angle POA = m$. The condition that $\angle POA < \pi/12$ therefore expresses, that $m = \Delta\omega/q < \pi/12$ or $\Delta\omega < \pi/12 \cdot q$. For a low modulation frequency, for instance $q = 2\pi \cdot 30$ sec^{-1}, $\Delta\omega < \pi/12 \cdot 2\pi \cdot 30 \approx 50$. For higher modulation frequencies $\Delta\omega$ is proportionally, but even for the highest modulation frequencies, which occur in praxis *e g.* $q = 2\pi \cdot 15\,000$, corresponding with the highest audible frequency, the frequency sweep is not greater than $2\pi \cdot 25\,000$ sec.$^{-1}$

By the method just described, phase modulation in the narrower sense would be obtained, because the modulation index m was assumed to be constant. In order, however, to obtain frequency modulation (*i.e.* the frequency sweep $\Delta\omega$ independent of the modulating frequency), before the amplitude modulation, the low-frequency signal is passed through a network, in which the amplitudes of the different frequencies are amplified proportionally

with the modulating frequencies. In order to obtain a large frequency sweep, which is desirable as will be discussed in a later article, the frequency of the modulated signal is multiplied in several successive stages, before it is transmitted. The carrier frequency ω_0 as well as the frequency sweep $\Delta\omega$ are thereby multiplied. The advantage of this method is, that a crystal-controlled oscillator is used, so that the carrier frequency is very constant.

In another method of modulation much used in transmitters, the capacity of the oscillation circuit is varied. If a condensor microphone is connected in parallel with the tuned circuit of an oscillator, the capacity of the circuit varies upon speaking into the microphone. The intantaneous frequency of the oscillator thereby also varies. If the variation of the capacity is small, compared with the total capacity of the circuit, the variation of the instantaneous frequency is approximately proportional to the capacity variation. Practically, however, there is an objection to including the microphone directly in the oscillator circuit. In order to avoid this, a so-called reactance valve is preferably connected in parallel with the oscillator circuit, which valve serves as variable capacity or variable self-induction[7]). The action of the reactance valve is as follows. Between anode and cathode an A C voltage is applied with the same frequency as, but shifted in phase with respect to, the anode voltage. The grid voltage can be derived in a simple way from the anode voltage, for instance by introducing a resistance R between anode and grid and a capacity C between grid and cathode (*fig.* 7). Here

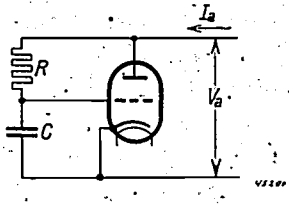


Fig. 7. Connections for the reactance valve. With the help of the resistance R and the condensor C the grid voltage is derived from the anode voltage V_a.

$V_g = V_a/1 + jR\omega C$. The anode A C is approximately $I_a = SV_g = V_aS/(1+jR\omega C)$, if S represents the slope of the characteristic. The impedance measured between anode and cathode is now $V_a/I_a = 1/S + jR\omega C/S$, and therefore equivalent to a connection in series of a resistance $1/S$ and a self-induction RC/S. It is here assumed, that the impedance of the series connection of R and C is so high, that the current through this impedance may be neglected compared with the anode A C By varying the negative grid voltage of the valve the slope S is varied and thereby also the effective self-induction RC/S and thus the frequency of the oscillator in parallel with whose circuit the reactance valve is connected (*fig.* 8). The positions of R and C

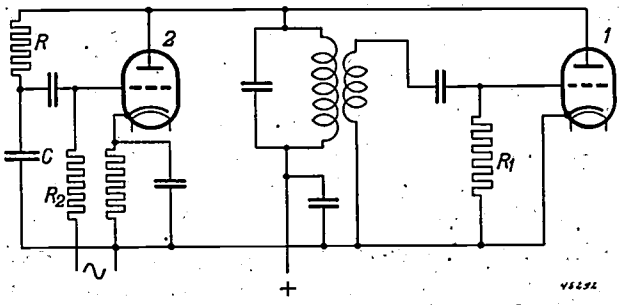


Fig. 8. Oscillator circuit with oscillator valve 1 and reactance valve 2/R and C are introduced according to fig. 7. The modulating signal is fed to the grid over the large resistance R_2.

can also be interchanged, where upon the reactance valves becomes equivalent to a resistance and a capacity in series. It is also possible to use a self-induction L, instead of the condenser C.

When the modulating low-frequency voltage from the microphone is now applied to the grid over the resistance R_2, the desired frequency-modulated signal is obtained in the oscillator circuit. The frequency of the frequency variation of this is equal to the frequency of the modulating signal. With suitable characteristics of the valve employed and suitable dimensions of the other elements of the connections, the frequency sweep is sufficiently nearly proportional to the amplitude of the modulating signal, provided it does not become too large.

Receiver for frequency modulation

In the receiver the frequency modulated signal must again be converted into a low-frequent oscillation, which is a faithfull copy of the voltage at the microphone. The detector (diode), employed in the amplitude-modulation receiver, reacts only to the amplitude and not to the frequency of the high-frequency signal applied and cannot therefore be used directly in frequency modulation, where the amplitude is constant. Therefore in the frequency modulation receiver the frequency-modulated signal is first converted into an amplitude-modulated signal and then applied to a diode detector. This conversion of a frequency-modulated into an amplitude modulated signal is accomplished by means of a frequency detector *i.e.* a network, which furnishes a high-frequency voltage, whose amplitude is sufficiently nearly a linear function of the frequency of the applied voltage,

[7]) Compare also: Philips techn. Rev. 4, 85, 1939.

if the amplitude of this applied voltage is constant. For this purpose may be used, for example, connections which consist of a resistance, a self-induction and a capacity in series. The resonance curve, which represents the voltage E over these connections as a function of the frequency, with constant current through the circuit, possesses a section, which is approximately straight (AB in *fig. 9*). If the frequency of the current varies over the region corresponding to this almost straight part, the variation of the voltage E over the circuit is proportional to this frequency variation. The sine curve a represents the instantaneous frequency ω_m of the modulated signal as a function of the time t; the sine curve b represents the instantaneous amplitude E of the high-frequency voltage over the circuit as a function of the time t. Except for a constant, this amplitude is proportional to the instantaneous frequency of the modulated signal minus the carrier frequency. This voltage, which thus contains both frequency modulation and amplitude modulation, is applied to a diode detector, which does not react to the frequency modulation, but to the amplitude modulation of this voltage. In this way, from a frequency modulated signal $\cos\{f(t)\}$ with a carrier frequency ω_0, one obtains a low-frequency signal, which is proportional to the instantaneous frequency minus the carrier frequency *i.e.* to $\{df(t)/dt\}-\omega_0$.

If we prolongue the nearly straight portion AB of the characteristic in fig. 9 until the frequency axis, this axis is cut in a point, lying at a distance q_1, of the central frequency ω_0. Moreover, we see in fig. 9, that the voltage E possesses an amplitude modulation with a modulation depth $\Delta\omega/q_1$. The ideal would be, that the characteristic followed indeed the whole of the straight dotted line; in this case the signal, arising after frequency-detection, would have obtained an amplitude modulation of 100% as a consequence of the frequency-sweep q_1, of the original frequency-modulation.

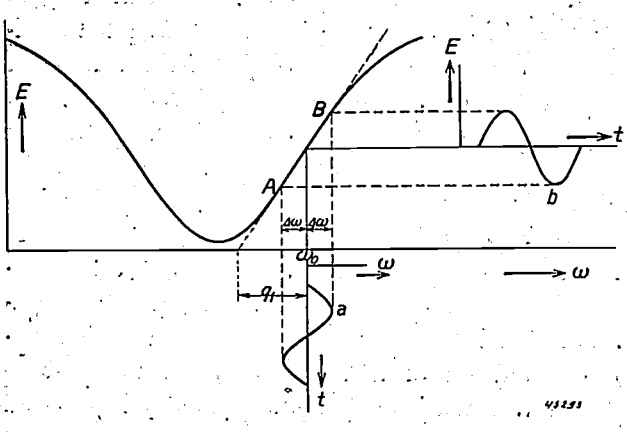


Fig. 9. High-frequency amplitude of the voltage E over a connection in series of a resistance, self-induction and capacity, as a function of the frequency ω. If we assume, that the instantaneous frequency varies sinusoïdally according to a with the time t, we obtain the relation given by b between the instantaneous amplitude E of the modulated high-frequency signal and the time t.

Another often used circuit for frequency-modulated signals is shown in *fig. 10*. In the anode circuit

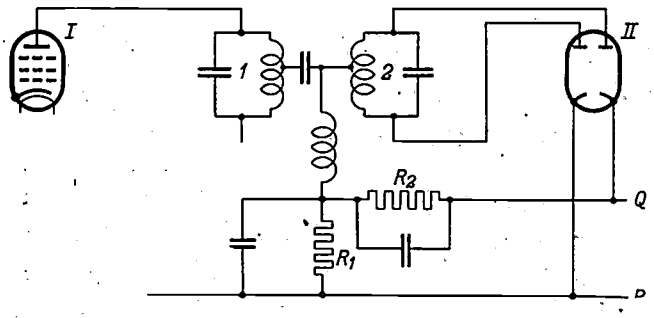


Fig. 10. Detector circuit for frequency-modulated signals. I last intermediate-frequency valve; II duodiode, *1* and 2 are the primary and secondary circuits of the frequency detector. Between the points P and Q we obtain the detected low-frequency signal.

of the last intermediate frequency, valve I of the receiver a network is included, which consists of two inductively coupled circuits, each tuned to the frequency of the carrier wave of the frequency-modulated signal to be received. For this frequency, the voltages over the primary circuit *1* and the secondary circuit *2* differ in phase by $\pi/2$; for higher frequencies this phase shift changes in one direction, for lower frequencies in the other. The middle of the secondary circuit *2* is connected with a point of the primary circuit *1*. To each of the two diodes, of which valve II consists, a voltage is applied consisting of the sum of a part of the voltage of the primary circuit *1* and half of the voltage of the secondary circuit *2*. These voltages are represented in *fig. 11* in a vector diagram. If the signal is unmodulated, OA is the primary voltage, AB and AC are the two halves of the secondary voltage.

The voltages on the two diodes are thus OB and OC. The D C voltages over the resistances R_1, and R_2 of fig. 10 are equal, since the vectors OB and OC are equal in length. Between the points P and Q of fig. 10 one obtains the difference between these two voltages, which is zero in this case. If, however, the frequency, of the signal applied, changes, the vectors AB and AC rotate, for instance to AB^1 and AC^1. The voltages over R_1 and R_2 are no longer equal, but $OB' > OC'$. The voltage between points P and Q is now equal to the difference between the lengths of the vectors OB' and OC'. If the frequency of the signal applied varies in the opposite direction the sign of the voltage between P and Q changes.

If the applied signal is frequency-modulated, there therefore occurs between P and Q a low-

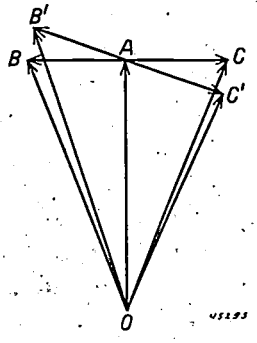


Fig. 11. Vector diagram for the frequency detector according to fig. 10. OA is the voltage, on the primary circuit *1*. OB and OC are the voltages, which the carrier wave furnishes to the two diodes, of which valve *II* consists, OB' and OC' are furnished at a frequency, which deviates slightly from the carrier frequency ω_0.

frequency A.C. voltage, which is an image of the modulation. By suitable choice of the coupling between the two circuits, of the ratio of the primary to the secondary voltage and of the damping of the circuits, the voltage between P and Q can, in a sufficiently wide frequency region, be made very nearly proportional to the frequency deviation from the carrier frequency. The desired low-frequency signal is therefore obtained free of distortion. *Fig. 12* gives for such a frequency detector the voltage between P and Q as a function of the angular frequency ω of the signal applied. This curve approximates to a straight line, closely enough between the frequencies ω_1 and ω_2, so that the

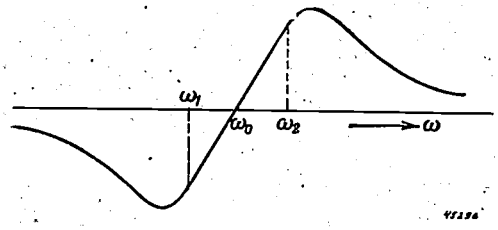


Fig. 12. The detected voltage between the points P and Q of fig. 10 as a function of the frequency ω. In the region around the carrier frequency ω_0 the curve is practically a straight line between ω_1 and ω_2.

detector can be used satisfactorily for a carrier frequency ω_0 and a frequency sweep $\Delta\omega$ which must be slightly smaller than $\omega_0 - \omega_1 = \omega_2 - \omega_0$, as will be discussed in the following section.

Non-quasi-stationary phenomena

In the foregoing, it has always been tacitly assumed, that the voltages and currents in networks behave, under the influence of signals with varying frequency, in the same way as at constant frequency, thus that the instantaneous frequency may simply be substituted for the frequency in the expression for the impedance. The fact, that this is at least doubtful, follows immediately from the somewhat arbitrary definition of instantaneous frequency. Carson and Fry[8]) have shown, how it is possible, to calculate the behaviour of arbitrary networks under the influence of a voltage with variable frequency. This exact calculation is quite complicated, but simple practical conclusions can be drawn from it.

When an A.C. voltage $V = V_{max} \cos \omega t$ is connected at the moment $t = 0$ with a resistance r and a self-induction L in series, a current

$$i = i_{max} \cos(\omega t - \varphi) - i_{max} \cos \varphi \cdot e^{-\frac{r}{L}t}$$

flows. This current can be divided into two parts: a "quasi-stationary" part $i_{max} \cos(\omega t - \varphi)$, which continues to exist as long as the external voltage acts, and a transient current

$$-i_{max} \cdot e^{-\frac{r}{L}t} \cos \varphi,$$

which practically disappears after a certain, usually very short time. In the case of a more complicated network also the current always consists of a stationary part and a transient current, which decreases to zero. Also when the amplitude or the frequency of the voltage applied changes, a temporary component of the current appears. With arbitrarily changing voltages therefore, one is concerned with such temporary components. If the change in frequency takes place so slowly, that the temporary component has already become very small before any appreciable frequency change has again taken place, *i.e.* in such a way, that the network has time to adjust itself to the new situation, these exponentially decreasing components are scarcely noticeable, and the currents and voltages in the network can be measured practically with sufficient accuracy, as if one were dealing with a constant frequency, namely as if the frequency at every moment were equal to the instantaneous frequency. In this case the network behaves as if quasi-stationary. In any case this method, of considering it quasi-stationary, gives a first approximation of the phenomena occuring. The extent, to which this approximation produces sufficiently accurate results in a practical case, depends upon the modulating frequency q, the frequency sweep $\Delta\omega$, and thus also on the modulation index m, as well as on the dam-

[8]) J. R. Carson and T. C. Fry, Variable frequency electric current theory, B. S. T. J. **6**, 513, 1937.

ping of the network. When in this connection the spectra of fig. 4 are again considered, it is seen that although the instantaneous frequency covers a region $2\Delta\omega$ continuously, the oscillation is composed of components with discrete frequencies. With a large values of m, *i.e.* with a slow variation of the frequency, the region of the instant neous frequency is indeed more densely occupied, than at small values of m, and thus more closely approaches a continuous spectrum, as would be expected according to quasi-stationary considerations. Moreover, with large values of m a relatively smaller region lies outside the region of the instantaneous frequencies than with smaller values of m. If $m < 1$, except for the carrier wave, not even a single component lies in the region covered by the instantaneous frequency. From this it may be seen again, that the quasi-stationary method of consideration leads to more accurate results the larger m is, and thus with a given value of $\Delta\omega$, the smaller q is.

In the amplifier stages of transmitter and receiver tuned circuits are employed, as a rule band filters, consisting of two coupled circuits. From the quasi-stationary method of consideration, it would follow, that the band width of these filters is sufficient, when a frequency region $2\Delta\omega$ is uniformly amplified. From the spectra of fig. 4, however, it follows that this band width is unsufficient. If $m < 1$ a band width of a least $2q$ is necessary, otherwise the whole modulation might even disappear. If $m > 1$ a band width of about $3\Delta\omega$ is necessary; otherwise too many side-band frequencies are cut off, which would cause distorsion in the modulation and therefore also in the low frequency signal detected.

From this it is evident, that the quasi-stationary method of consideration, which led to the concept of instantaneous frequency, can easily lead to incorrect conclusions in the case of frequency modulation. The incorrect concept, mentioned in the introduction, that a frequency-modulated signal with a small frequency sweep would require only the region in the frequency spectrum, which is covered by the instantaneous frequency, was a result of the quasi-stationary method of consideration, which does indeed give a useable approximation, when the modulation index m is large, but which leads to absurd results when m is small.

After the above discussion of the way, in which frequency modulation can be realized for the transmission of low-frequency signals on a high-frequency carrier wave, there remains the problem, as to the cases in which such a system can be succesfully used for radio transmission. To what degree frequency modulation offers advantages over the usual broadcasting system of amplitude modulation will, however, be discussed in a later article.

THE FORMATION OF STEREOPHONIC IMAGES

by K. de BOER. 534.621.3

It is possible to construct an installation for sound amplification or reproduction, such as is used for orchestra music, sound film or stage plays, in such a way that the „acoustic image" heard by the audience always coincides with the visual image which is seen. In particular the conditions are here discussed under which the sound must be recorded in order to attain the desired result. Several more particulars are also given about the technique of reproduction which has been dealt with previously, while it is also pointed out that stereophonic reproduction, without the source of the sound being visible, already constitutes a considerable improvement.

Equipment for sound amplification is installed in a hall in which a small orchestra performs. In the usual way a microphone is hung somewhere in front of the orchestra and the sound received by the microphone is amplified and reproduced by a loudspeaker. When the attention of a member of the audience is especially attracted first by the violon at the left and then by the piano more to the right, he will often feel troubled, that the music which he hears, does not come from the direction in which he sees the violin or piano, but always from the same direction, that of the loudspeaker. Sound reproduction at the cinema is usually subject to this objection. Here also, the sound always comes from a single direction, even though the source of sound (the speaker) moves about on the screen. This often produces an unnatural effect.

It is also a remarkable fact that in those cases where sound reproduction is employed, without the audience being able to see the original sources of the sound (or pictures of them), a considerable improvement of quality can be attained by applying stereophonic reproduction. Even the reproduction by means of two „ordinary" loudspeakers connected in series, placed at some distance from each other, already provides an improvement, which is sometimes made use of in the case of the radio. For the present, we shall confine ourselves to mentioning this fact, and in the following devote our special attention to those cases, where the audience can see the source of sound or a picture of it and where it will be expected, that the sound reproduced will come from the direction, in which the source of that sound is seen.

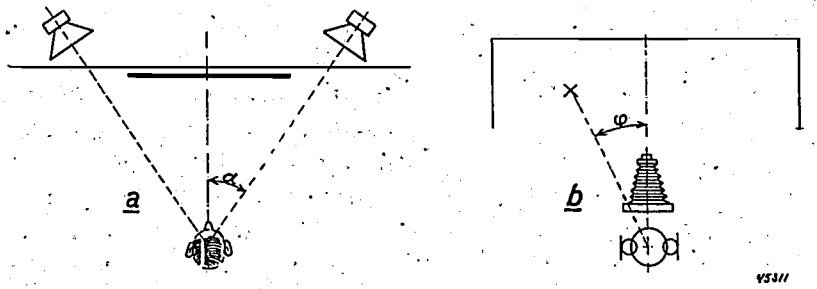


Fig. 1. *a*) Arrangement for stereophonic reproduction in a cinema. Two loudspeakers are placed one on either side of the projection screen and are observed within an angle 2α by the listener.
b) Arrangement of the artificial head in the studio. The perpendicular bisectoring planes of camera and artificial head are coincident. The speaker stands for instance in the direction φ.

With the stereophonic method of sound reproduction these objections can be removed and the sound reproduced is always heard from the same direction, as that in which its source is seen. It is important, that even on those persons, who are scarcely aware of the objections mentioned, stereophonic reproduction, nevertheless, often produces a more pleasing impression than ordinary reproduction.

Principle of the arrangement for stereophonic image forming

The principle by which the coincidence of sound image and visual image can be attained has already been sketched in a previous article [1]) for the case of the cinema, particular attention being devoted

[1]) See the article: Stereophonic sound reproduction, Philips techn. Rev. 5, 107, 1940.

to the technique of reproduction. We shall review it briefly here, see *fig. 1*. For sound recording an "artificial head" (a sphere) is set up in the studio, which has two microphones at the position of the ears. The sound recorded on the film for each of the two microphones separately is reproduced in the well-known manner *via* separate channels, in which an amplifier is included, by a loudspeaker. The two loudspeakers are placed at either side of the screen in the projection room. It must be pointed out particularly, that this method of reproduction is simpler and pleasanter than that in which each member of the audience must wear headphones, while by calculation and experimentally it was found, that the accuracy of the sound image need not suffer appreciably. In this article therefore, we shall not devote our attention so much to reproduction, but to the requirements made of the method of sound recording. The question thus is, how must the sound be recorded in the case of the orchestra and in analogous cases (stage, film), in order to obtain an acoustic image, which is true to nature when the above-mentioned method of reproduction is employed. It is found, that for this purpose both the size and the position of the artificial head must be chosen according to a definite rule. Before we state this rule and discuss its application, we shall outline the experiment from which it results.

In a recording room, see *fig. 2*, an artificial head has been placed. The microphones are connected with two loudspeakers in the projection room by two separate channels, each obtaining an amplifier.

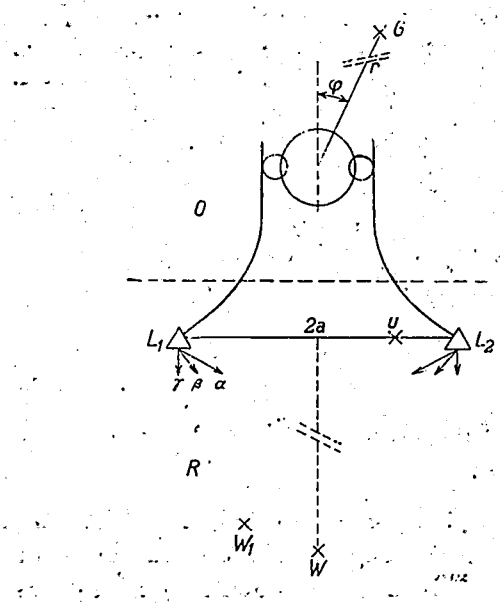


Fig. 2. In the recording room O an artificial head and a moveable source of sound G, are set up. In the recording room R, the observer W observes the sound image of G between the two loudspeakers L_1 and L_2 at a distance u from the middle.

The distance $2a$ between the loudspeakers, the base, can be chosen differently, for instance normal stage width of 4.5 m, width of the projection screen, etc.

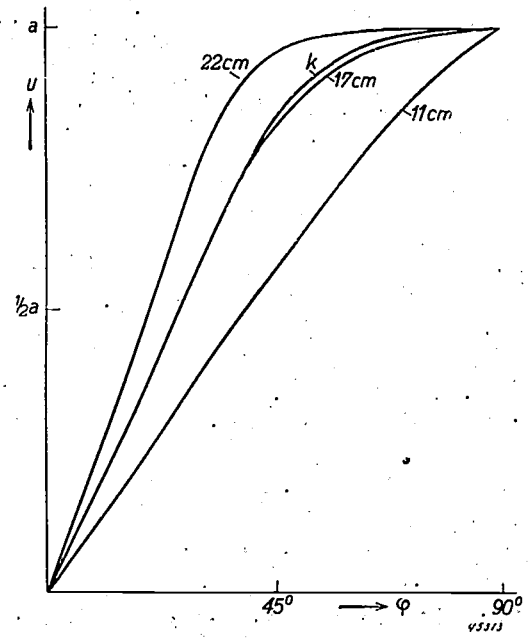


Fig. 3. Upon recording the sound from a source G (fig. 2) placed at an angle φ with the plane of symmetry of the artificial head u represents the deviation from the middle of the sound image as observed by W in the projection room. The various curves are for the different sizes of artificial head as indicated in the figure. k stands for a perfectly shaped head.

About in the middle of the hall, in our case at a distance of 9 m, the observer W is seated. In the recording room a source of sound is placed and its position fixed by the distance r and the angle φ. In the projection room W then hears the sound image between the loudspeakers at a distance u from the middle. We now determine this deviation u at different values of φ and r.

It is found, that the deviation u of the sound image depends only little upon r, so that we obtain in fact only a relation between φ and u, thus between the angle, measured from the plane of symmetry, at which the artificial head „hears" the source of sound and the deviation of the sound image from the middle.

This relation, which differs for different sizes of artificial head, is given in *fig. 3* for three different cases:

1. For an artificial head (sphere) 22 cm in diameter; which is slightly larger than a normal head;
2. For a well modelled head, and for a sphere of 17 cm diameter, which is about the same size as the head;
3. For a sphere 11 cm in diameter.

The graph shows that, for the 22 cm sphere, the relation between φ and u is linear as long as φ is not greater than 45°. This linear relation, in which equal angular rotations φ correspond to equal displacements u, gives a sound image, whose accuracy and naturalness are more than sufficient for practical purposes. As soon, however, as the source of sound has a greater angular deviation than 45° from the centre, in the case of the 22 cm sphere, the sound image always lies close to one of the loudspeakers, so that we obtain a much too compressed and distorted sound image which does not give a pleasing impression. (For example all the musicians of a large orchestra who are observed by this artificial head in directions deviating by more than 45° from the centre would in the acoustic image appear to be sitting in a lump, while the others would be at normal distances apart).

It is found further that, with the small sphere, we can reproduce a larger range of angles, namely to approximately 90° from the centre, undistorted on the base. This can easily be understood, because with a given angle φ, the ratios of intensities and differences in time are smaller with the small sphere than with the large, as that the corresponding deviations u are smaller and a wider range of angles can be accomodated on the given base [2]).

For the relation between the size of the artificial head and the width of the range of angles, which can be reproduced undistorted on the base, we can further deduce from fig. 3 the following rule of thumb: (diameter of artificial head in cm) × (angle to be projected in degrees) = 2000.

This rule of thumb is very useful in solving problems like the following. Sterephonic sound amplifying apparatus is installed in a hall in which an orchestra consisting of five musicians is performing. A loudspeaker for the sound reproduction is mounted on either side of the orchestra. How must the sound now be recorded, *i.e.* which diameter must be chosen for the artificial head and where should it be placed?

The correct position is simply found. Is is the same as the position of the single microphone in ordinary sound amplification, and certain rules for this are known from practice. When the artificial head, which is placed according to these rules „sees" the orchestra within an angle of 100°, for correct image forming we must, according to the rule of thumb, chose a diameter of 20 cm, see *fig. 4a*. We then obtain an image like that shown in *fig. 4b*. Our problem in thus solved sufficiently accurately for practical purposes; the objection described at the beginning of this article has been met. Analogous considerations hold for the formation of the acoustic image of a large orchestra, an open-air play and such cases. In the meantime stereophonic reproduction has been employed in practice only for the cinema on a large scale. while for the other cases it is still very much in its initial stages.

In connection with these figures we shall discuss several particulars in more detail.

Since fig. 3 is valid for the observer W, it might be thought that, by an observer at any other point in the hall, the sound image would also be observed at another point u. It would lead us too far afield to deal with this question in detail and we shall confine ourselves to a single aspect. When W moves along the line of symmetry the sound image scarcely moves. The displacement may, however, become somewhat greater, when W moves to one side to a point W_1 (see fig. 2). At this point the influence of L_1, is dominant and the image shifts somewhat to the left. This could, however, be avoided by giving L_1 and L_2, a certain directional effect, *i.e.* by providing that the same sound energy is not radiated in the directions α, β, γ (fig. 2), but that, as the length of the arrows indicates, there is a decrease in the intensity radiated in the order given. In this way the dominating influence of L_1 (and the diminished influence of L_2) can be compensated, and an undistorted image can be obtained throughout practically the whole hall.

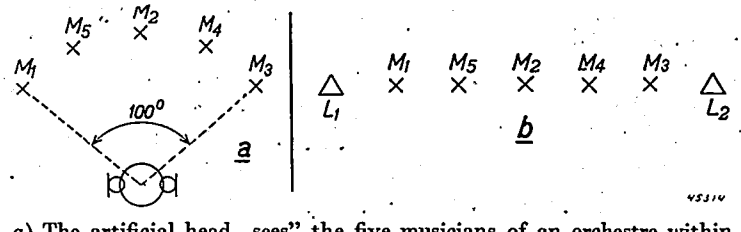


Fig. 4. *a*) The artificial head „sees" the five musicians of an orchestre within an angle of 100°. Upon correct choice of the diameter of the artificial head an acoustic image is formed as sketched in fig. 4 *b*) whose accuracy is sufficient in practice.
The loudspeakers L_1, and L_2 are actually placed on either side of the orchestra.

[2]) See the article: Stereophonic sound reproduction, Philips techn. Rev. 5, 107, 1940.

Width regulator

It is also evident, from the experiment described above, that if the base $L_1 L_2$ had been chosen longer than the width of the orchestra in question, we would still have obtained an undistorted, but at the same time an enlarged acoustic image.

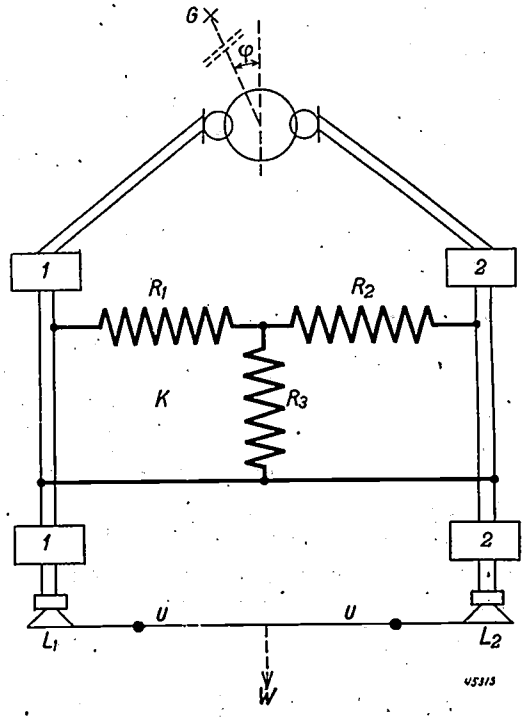


Fig. 5. Diagram of an arrangement for stereophonic reproduction with adjustable width of sound image. The sounds received by two microphones on either side of an artificial head in the recording room are reproduced by the loudspeakers. L_1, and L_2. The observer W then observes an "acoustic image" of the recording room between L_1 and L_2. By means of the coupling element K, indicated by heavy lines, the width of this image can, if necessary, be changed. In this way it is always possible to attain satisfactory correspondence with a visual image also observed by W.

Thus, when it is desired to „project" images of objects of different sizes (large orchestras, stage plays or films, small orchestras) in the same hall, the loudspeakers should in each case be placed at a different distance apart. This difficulty can now be entirely eliminated by a slight modification in the electrical connections, which are indicated by heavy lines in *fig. 5*. We shall show that, even though the loudspeakers are located permanently at the greatest distance apart which will ever be necessary, with the connections of fig. 5 the apparent distance between the loudspeakers can be decreased in a simple way, so that the sound image can always be adjusted to the natural size.

The coupling element K indicated by heavy lines consists of three variable resistances, namely R_1 and R_2, which are always taken equal, and R_3. (At the same time it is also indicated in fig. 5 that for practical reasons, which we shall not go into, the amplifier in each channel is constructed in two parts, a preamplifier and an end amplifier.) Due to the presence of K part of the current from preamplifier *1* enters the „wrong" channel (namely *2*), while at the same time the same fraction of the current from preamplifier *2* also enters the wrong channel (namely *1*). Suppose now that the sound radiated by a source G is received by the left-hand microphone at the moment t_1, and a moment later at t_2 by the right-hand microphone. Part of the sound energy taken up at moment t_1, by the left hand channel, after amplification to a certain amount I_1 is radiated by the left hand loudspeaker also at the moment t_1. (The transit time of the sound converted into electrical vibrations can be neglected). Another part, however, is fed, *via* the coupling element, to the second end amplifier, whereupon the right hand loudspeaker radiates an energy pI_1, also at the moment t_1, the fraction p being regulated by the varying of R_1 $(= R_2)$ and R_3.

In *fig. 6* this situation is sketched. A time axis is indicated in the vertical direction. The arrows pointing to the left are proportional in length to the energies radiated by the left hand loudspeaker, those pointing to the right to the energies radiated by the right hand loudspeaker. At the moment t_1, therefore we have I_1, on the left and pI_1, on the right. For the sound, which is received by the right hand channel at t_2, the case is analogous. We then find that at t_2 I_2 is radiated by the right hand loudspeaker and pI_2 by the left hand one, p being the same fraction as before, because of the complete symmetry of the connections. This is also indicated in fig. 6; at t_2 we have thus I_2 to the right and pI_2 to the left. We now determine the position of the sound image observed by W. The ratio I_1/I_2 is the

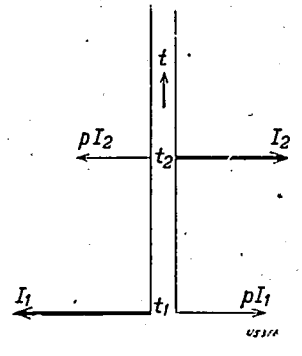


Fig. 6. The arrows toward the left indicate the intensity of the sound emitted by the left-hand loudspeaker at the moments t_1 and t_2. The arrows toward the right are for the right-hand loudspeaker. The arrows drawn with a thin line represent the effect of the coupling K of fig. 5.

same as would be radiated by L_1 and L_2 in the absence of the coupling K, since I_1 and I_2 have both lost the same fraction. The time difference $t_1—t_2$ is also unchanged by K. The combination (I_1/I_2, $t_1—t_2$) alone, would therefore be observed by W as if it came from a sound image which shows a deviation u to the left, corresponding to the angle φ according to fig. 3. However, W also observes the intensities pI_1 and pI_2 which are in the same ratio I_1/I_2 and have the same time difference $t_1—t_2$, but, in contrast to the above, the highest intensity (pT), comes from the right hand loudspeaker. The combination (p^rI_1/pI_2, $t_1—t_2$) is thus observed by W as an image, also with a deviation u, but now to the right, and weaker in intensity by a factor p. W thus finally observes the sound image which would occur if two loudspeakers at a distance u to the left and right of the middle radiated intensities with the ration p, of which the left-hand one is the stronger. According to the measurements already published [3]) the sound image deviates from the middle toward the left by a fraction of u determined only by p. Wet hus reach the conclusion, that due to the „cross talk" caused by K the deviation from the middle of every sound image is decreased in the same ratio (determined by p), so that we observe a reduced but undistorted image.

Fig. 7 gives some examples, how the image of an artificial head can be affected by a width regulator. The deviation of the sound image from the centre has been recorded as a function of the change of φ, with regard to the pl ne of symmetry of the artificial head. The dotted line *I* is the image, measured without a width regulator, for an artificial head and was drawn as smooth as possible through a great number of observations.

From the experimental curve *I* we have, for two distinct positions of the width regulator, calculated the drawn curves *II* and *III* according to the here explained theory. The measurements for these cases coincide satisfactorily with the drawn curves.

The regulation of this reduction is carried out in practice with a single knob which in a number of positions switches in suitable values of R_1 ($= R_2$) and R_3, while these values are, moreover, so chosen that the total intensity of the sound does not change upon regulation (width regulator).

Left-right regulator

It is clear that the two reproduction channels *1* and *2* of fig. 5 must not cause any alteration in the intensity relations and the time differences of the sounds received. As for the time differences, there is no danger of this happening with the connections employed; special measures must, however, be taken against a change in the correct intensity ratios. The two fixed amplifiers and the two loudspeakers can be made mutually identical once for all. A difference in the sensitivity of the recording microphones due to which a displacement of the whole sound image toward the left or right would occur upon reproduction, is usually combatted in a different way. Since different microphones are often used with different recordings, because for example of the different sizes of the artificial head, it is in practice simplest to include a potentiometer in each channel, and to increase the energy taken up in one channel somewhat, if necessary, and at the same time to decrease that in the other channel. This has the same effect as a corresponding alteration of the sensitivity of the microphones. The sound image can then be put back in the right position without a change in the total energy (*left-right regulator*).

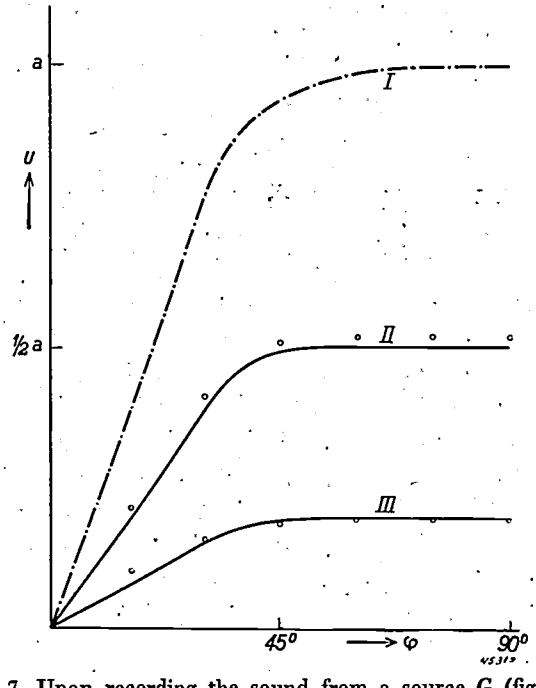


Fig. 7. Upon recording the sound from a source G (fig. 2), which is placed at an angle φ to the plane of symmetry of two separately mounted microphones, u gives the deviations of the sound image from the middle, as it is observed by W in the projection room. The various curves are for the different distances between the microphones indicated in the figure.

Recording of the sound with two separate microphones

Let us now return for a moment to the experimental setup of fig. 2. Instead of recording the sound in this case with two microphones mounted on an

[3]) See the article: Stereophonic sound reproduction, Philips techn. Rev. **5**, 107, 1940.

artificial head, it is also possible to use two separate microphones. Practically the same stereophonic effects are obtained, provided the separate microphones are placed about three times as far apart as is the case with the microphones on the artificial head. In a manner analogous to the one used above, the following rule of thumb is found: (distance between separate microphones) × (angle to be projected in degrees) = 6000, *i.e.* with a constant three times as large as in the former case. See *fig. 8*.

The fact, that the constant must be larger, is easily understood, when it is kept in mind, that two separate microphones at the same distance apart as the microphones of an artificial head would register a considerably smaller ratio of intensities and would thus give considerably less stereophonic effect. In the case of the artificial head the ratio of the intensities at the position of the microphones is considerably increased by reflection and diffraction.

Although the separate microphones have the advantage, that their mutual distance can be changed quickly upon passing from one object to be projected to another, while otherwise a different artificial head must be chosen each time, the artificial head is, nevertheless, preferable, because it is found to give sharper sound images. This is in agreement with the experimentally found result, that a sound image is sharper, the more it is determined by the intensity ratio rather than the time difference. With the separate microphones which must stand farther apart, time differences play a relatively large part and the result is somewhat less sharp images.

The characteristics here mentioned have been investigated with microphones whose sensitivity did not depend upon the direction from which the sound comes. The peculiarities, which occur when the microphones possess a certain directional effect, will not be considered, nor will the generally slight effect of the dimensions of the microphones.

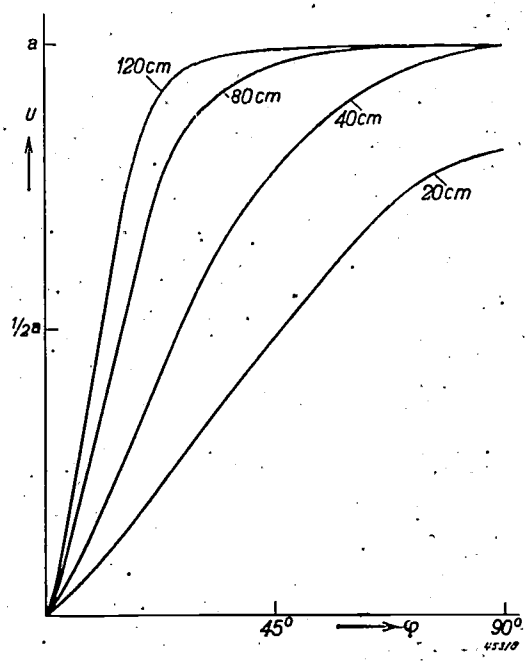


Fig. 8. In recording the sound of a source G (fig. 2), which is placed at an angle φ, with regard to the plane of symmetry, u represents the direction of the sound image from the centre observed bij W in the reproduction room. The different curves belong to the distances recorded in the figure between the two microphones.

In conclusion, we should like to point out, that in several very special cases the reproduction of the sound may be obtained, not by means of two loudspeakers, but by means of two headphones, one for the right and one for the left ear. For this case, which until now, has only been applied in the case of persons with poor hearing, analogous rules for the projection can be set up. We shall not go into it here, since this method of reproduction has a much less general application.

SEVERAL AFTER-EFFECT PHENOMENA AND RELATED LOSSES IN ALTERNATING FIELDS

by J. L. SNOEK and F. K. du PRÉ.

537.226.31

The combatting of the dielectric and magnetic after-effect is often an important problem, especially in the development of new materials for electrotechnology. After-effect phenomena are the cause of the „losses" in dielectrics, while at least part of the losses in ferromagnetic substances must also be ascribed to that cause. In this paper the phenomena are described, which occur in the after-effect and their causes are discussed, while at the same time the close analogy and the connection are pointed out between electrical and magnetic after-effect and the phenomena of elastic after-effect.

The search for materials, which possess new properties or new combinations of already known properties, which often make possible entirely new constructions, is becoming more and more important. In electrotechnology this search is manifested for example in the extensive investigation in the field of dielectrics and ferromagnetic substances. However, no matter what new properties it may be desirable for these electrotechnical materials to possess, one fondamental requirement will always be made, namely, that they must show very small „losses".

This requirement is understandable, when it is kept in mind, that in a dielectric or ferromagnetic substance, which is not free of losses, a certain heat development will occur, when it is placed in an alternating electric or magnetic field. This heat development may, in the case of strong alternating fields, such as may for example occur in transmitters, mean a considerable loss of energy, while at the same time the material may be damaged by the too rapid rise in temperature. Also in those cases, where the energy loss itself is no objection, the heat development often has a harmful effect, it is for example known, that the very important selectivity of oscillation circuits or filters is diminished by the use of materials, which are not free of losses. It is for this reason, that the combatting of losses or the preparation of new substances with few losses is such an important technical problem.

The nature of the losses in dielectrics and ferromagnetic substances may be varied. In a dielectric, when the direct-current conductivity of the substance may be neglected, the losses must practically always be ascribed to after-effect phenomena.

In a ferromagnetic substance, on the other hand, the losses are caused partly by the conductivity of the substance (eddy-current losses), while hysteresis may be mentioned as a second important cause. Finally in these substances also a part of the losses must be aseribed to after-effect, although this part is often relatively small. In the development of new magnetic materials, however, it has repeatedly been found, that a substance with otherwise favourable properties could not be employed, because the strong after-effect phenomena occuring in it could not combatted effectively. To mention a single example, the transformer iron, prepared electrolytically, is still very little used in heavy-current technology, only because of the disturbing after-effect, which is difficult to combat.

The investigation of the after-effect phenomena is therefore important not only in the case of dielectrics, but also in that of magnetic substances.

It is now our intention to show, that matter may exhibit phenomena of after-effect, not only under the influence of electrical or magnetic forces, but also under the influence of mechanical forces, demonstrating at the same time, that the after-effets are caused in many cases by a diffusion phenomenon.

Elastic after-effect

As an example, we have chosen a case of elastic after-effect, because the phenomena here are easy to explain and at the same time show a close analogy with other after-effect phenomena.

The following experiment is performed: A rod clamped at one end is suddenly bent by a force acting at the other end. It would be expected, that when the bending force is kept constant, the bending would also remain constant. This is not, however, the case; the bending gradually increases somewhat, to reach a new true equilibrium value: elastic after-effect occurs [1]) *(fig. 1)*.

This elastic after-effect is very general and in different materials, it is often due to the same cause, namely, heat conduction. If we assume, that the material has a positive coefficient of expansion,

the side, which is suddenly compressed will, as we learn from thermodynamics, become slightly warmer, and the other side slightly colder. The compressed, warm side will then exhibit extra resistance

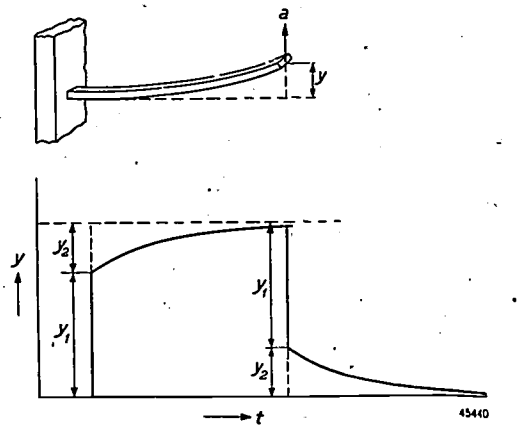


Fig. 1. At the end of the above sketched rod a constant force a is suddenly applied, and after the elapse of a long time suddenly removed again. The course of the bending y as a function of the time is indicated in the lower half of the figure. The after-effect is very much exagerated, ordinarily y_2 amounts to less than 1% of y_1.

to further compression, because it has a tendency to expand due to the rise in temperature. Similar reasoning is valid for the side that is stretched, which has a tendency to shrink. As soon as the difference in temperature is equalized by the conduction of heat, the elastic force decreases slightly and the rod bends farther. The phenomenon is known as the thermo-elastic effect.

In the exceptional case of a negative coefficient of expansion the warm side is the stretched side, but otherwise the reasoning is the same.

Closer analysis shows, that the equalization of temperature takes place in such a way, that the change in temperature ΔT at every point in the rod decreases as $\Delta T e^{-t/\tau}$. In this connection the bending y due to a constant force a can be represented by:

$$y = y_1 + y_2 (1-e^{-t/\tau}) \quad . \quad . \quad . \quad . \quad . \quad (1)$$

This equation expresses the fact, that at $t = 0$ a bending y_1 occurs, while after some time the bending has increased to $y_1 + y_2$. The extra bending y_2 occurs gradually, corresponding with the change in temperature. The quotient y_2/y_1 which is usually very small (for instance 1%), may be called the strength of the effect. The quantity τ is called the relaxation time.

Upon the sudden removal of the external force, the deviation at first suddenly decreases by an amount y_1, while the additional y_2 gradually disappears according to $y_2 e^{-t/\tau}$. From this, we conclude, that the substance suffers no permanent alteration due to the after-effect. On the contrary, the original condition is completely restored.

When the force a acting on the end of the rod is no longer constant, but periodical:

$$x = a \cos \omega t, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

the deviation can be represented by:

$$y = b \cos (\omega t - \delta), \quad . \quad . \quad . \quad . \quad . \quad (3)$$

as long as $\omega \ll$ the characteristic frequency of the rod. In the case of a periodical force, therefore, the result of the after-effect is, that a phase difference occurs between force and deviation. This important property can be derived in a simple way from the above facts, as will be shown below.

The following holds for the phase shift:

$$\tan \delta = \frac{y_2}{y_1} \frac{\tau\omega}{1+\tau^2\omega^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

while the amplitude b lies between y_1 and $y_1 + y_2$, as follows from:

$$b = y_1 + \frac{y_2}{1+\tau^2\omega^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

The quantities y_1 and y_2 are the same as those appearing in formula (1) for the case of a constant force a.

The value of δ reaches a maximum for that frequency for which $\tau\omega = 1$, see *fig. 2*. Expression (4) is valid only for small values of y_2/y_1, which is almost always the case in practice.

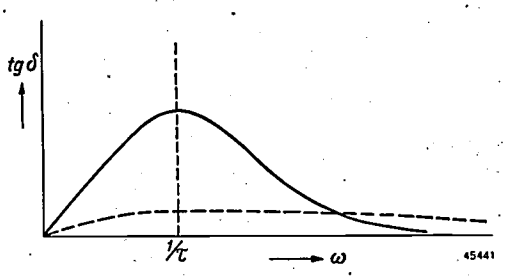


Fig. 2. With periodic loading of the rod shown in fig. 1 the likewise periodic deviation exhibits a phase shift δ (retardation), with respect to the force. The quantity $\tan \delta$ is here plotted as a function of the angular frequency ω. The full-line curve holds for the simplest case of after-effect, the broken line curve for a more complicated case. τ is the relaxation time.

[1]) In this experiment in order to prevent the rod from vibrating about its equilibrium position, thus to avoid effects of inertia, several precautions must be taken. The characteristic time of oscillation of the rod must be very short compared with the time during which a noticeable after-effect occurs. In that case the „sudden" force can be applied in a time which is very short compared with the after-effect time, but at the same time very long compared with the characteristic oscillation time of the rod, so that the after-effect is not disturbed by the occurence of characteristic vibrations.

It must be noted that, due to the after-effect, in this periodical movement heat is developed in the rod. The work performed by the force x per period, $\oint x\,dy$, is, due to the phase shift, not equal to zero. The after-effect causes a loss of energy. Furthermore it may be seen from fig. 2, how the influence of the after-effect can be combatted, not only by removing the cause, but also by employing very high or very low frequencies.

If the after-effect does not have the simple form indicated by (1) and fig. 1, the after-effect curve can, nevertheless, always be analyzed and expressed in the form:

$$y_1 + \sum_n y_n\,(1-e^{-t/\tau_n}) \quad \ldots\ldots \quad (6)$$

We then again find for the deviation under the influence of a periodical force:

$$y = b\cos(\omega t-\delta), \quad \ldots\ldots \quad (7a)$$

where now, however,

$$\operatorname{tg}\delta = \sum_n \frac{y_n}{y_1}\,\frac{\tau_n^2\omega}{1+\tau_n^2\omega^2} \quad \ldots \quad (7b)$$

and

$$b = y_1 + \sum_n \frac{y_n}{1+\tau_n^2\omega^2} \quad \ldots \quad (7c)$$

The variation of tan δ with th frequency may be much flatter than in the case of a single relaxation time, see for instance the broken line in fig. 2.

We shall now briefly derive the above results. For the sake of simplicity we assume, that the after-effect curve can be described by a single power of e:

$$y = y_1 + y_2\,(1-e^{-t/\tau}).$$

Since y_1 as well as y_2 will be proportional to the constant force a, we may also write:

$$y = a\,[\,\varepsilon_1 + \varepsilon_2\,(1-e^{-t/\tau})\,],$$

or more briefly:

$$y = a\,[\,\varepsilon_1 + \varepsilon_2 f(t)\,].$$

In order to derive from this the fact, that in the case of periodical loading the above-mentioned phase shift occurs, we must still call attention to an experimental peculiarity, namely the superposition effect, which is found not only here, but in all other after effect phenomena in solids.

Let us assume, that a suddenly applied loading is altered after a certain time t_0 by a different suddenly applied loading (for example in the opposite direction). For the deviation now occuring we find, as sketched in *fig. 3*, a form, which is given by the superposition of the deviations, which each of the processes by itself would have caused. If the first deviation would have been represented at a given moment t by $\varepsilon_1 a + \varepsilon_2 f(t)\cdot a$, and the second by itself would have been $\varepsilon_{10} + \varepsilon_2 f(t-t_0)b$, then the final result can be found by adding these expressions.

If the loading is given not by „impulses", but by a continuous function $x(t)$, then an increase in loading dx, occuring

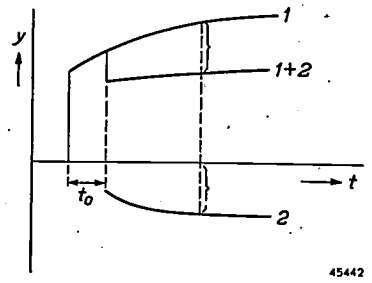


Fig. 3.

at the moment t, will be manifested at a later moment t_1 by a contribution to the deviation of the amount:

$$\varepsilon_1\,dx + \varepsilon_2 f(t_1-t)\,dx.$$

In order to deal with the above-described case, we now assume that from the moment, $t =$ „zero" the following will, hold: $x = a\cos\omega t$. As *fig. 4* shows, at $t =$ „zero" we must begin by introducing a jump of magnitude a in the loading. According to the superposition principle the deviation at any given moment t_1 is then given by:

1) The contribution of the jump, thus $\varepsilon_1 a + \varepsilon_2 f(t_1)a$;

2) The contribution of the continuously varying loading:

$$\varepsilon_1 \int_0^{t_1} dx + \varepsilon_2 \int_0^{t_1} f(t_1-t)\,dx, \text{ with } x = a\cos\omega t.$$

These contributions can be calculated in a simple way. We find:

$$y = a\left(\varepsilon_1 + \frac{\varepsilon_2}{1+\tau^2\omega^2}\right)\cos\omega t_1 + a\,\varepsilon_2\,\frac{\tau\omega}{1+\tau^2\omega^2}\sin\omega t_1 - \frac{a\varepsilon_2}{1+\tau^2\omega^2}\,e^{-t_1/\tau}.$$

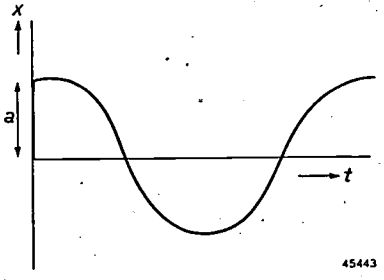


Fig. 4.

The last term may always be neglected in practice, while the other two can easily be reduced to the form already discussed, where the arbitrary moment t_1 is again called t. In the case of more complicated after-effect according to (6) the derivation is analogous.

In connection with a case of magnetic after-effect to be discussed later, we should like to consider very briefly one of the many other phenomena of elastic after-effect, for which purpose the following experiment is performed.

An iron wire is suddenly given a torque by means of a constant couple. The angle of torsion shows a variation like that represented in (1),

thus first a sudden increase and then a gradual further increase [2]).

Although the phenomenon proceeds in a manner entirely analogous to the previous one, the cause must certainly not be sought in the thermo-elastic effect, because no change in volume and thus no temperature differences occur upon torsion. The cause only became clear, when it was found, that the phenomenon was only observed in the case of iron, which contains small quantities (0,01%) of carbon or nitrogen in solid solution. The "dissolved" particles are situated at intermediate lattice spots. The way, in which these dissolved particles can cause the after-effect, can be explained as follows.

Upon torsion each volume element of the iron experiences a gliding, which amounts to a compression of the lattice in one direction and an expansion in the perpendicular direction. The carbon or nitrogen particles, situated between the iron atoms, therefore diffuse from the compressed spots, where little space is left, to the stretched spots. Due to this, however, a decrease in the elastic tensions occurs at the same time, so that a constant couple is able to give the wire gradually a further torque, corresponding to the progress of the diffusion.

In the two examples, mentioned here, the cause of the after-effect was a diffusion phenomenon, in the first a diffusion of heat, which restored a disturbance of the uniform temperature, in the second a diffusion of material particles, which compensated for a disturbance of the concentration. This property is not confined to the examples given. On the contrary, in almost all cases of after-effect in solids investigated until now, it has been found that the after-effect is based on a diffusion process, which restores a disturbed thermodynamic equilibrium.

The phenomena are always quite analogous to the ones described above, as will appear later in the case of the electrical and magnetic after-effect, to which we shall now devote our attention.

Incidently, it may be noted, that the establishment of a thermodynamic equilibrium will always take a certain time. After-effect will only be observed experimentally, when the equilibrium is established wholly or in part at a sufficiently slow rate (*e.g.* in a time comparable with the characteristic time of the considered experiment).

[2]) This phenomenon was first observed by Becker, and Kornetsky in carbonyl iron; every other kind of iron, however, shows the same behaviour.

Dielectric after-effect

The experiments, in which the dielectric after-effect was originally first observed many years ago, are entirely analogous to those used to demonstrate the elastic after-effect. We consider a dielectric, which is suddenly placed in an electric field. The dielectric displacement D which thereupon occurs, exhibits in almost all cases a variation analogous to that of (1), *i.e.* first a sudden and then a gradual increase. This, often very rapid, phenomenon is shown diagrammatically in *fig. 5*, as well as the (less simple) variation of the somewhat more graphic quantity $I = 1/4\pi \cdot \mathrm{d}D/\mathrm{d}t$,

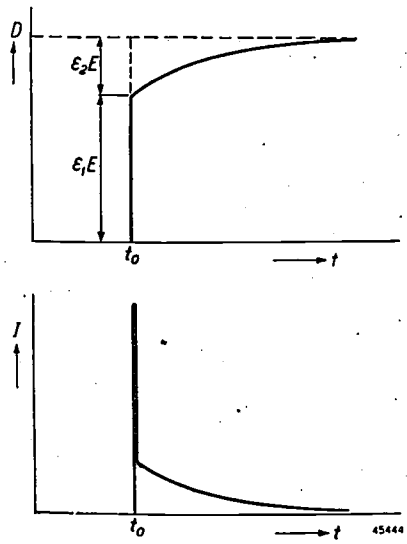


Fig. 5. Variation of the dielectric displacement D and of the displacement current $I = 1/4\,\pi \cdot \mathrm{d}D/\mathrm{d}t$ in an insulator, which at the moment $t = t_0$ is suddenly placed in an electric field.

the displacement current in the dielectric. The behaviour of D upon the sudden disappearance of the field will be immediately clear. Although the variation of the after-effect is often not according to a single power of e, but one which must be described by (6), we shall first confine ourselves to the simple case.

From the observed after-effect we conclude, as above, that in a dielectric placed in a periodical field:

$$E = E_0 \cos \omega t \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

a dielectric displacement takes place according to:

$$D = D_0 \cos (\omega t - \delta) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

with:

$$D_0 = E_0 \left[\varepsilon_1 + \frac{\varepsilon_2}{1 + \tau^2\omega^2}\right] \text{ and } \tan \delta = \frac{\varepsilon_2}{\varepsilon_1} \frac{\tau\omega}{1 + \tau^2\omega^2},$$

where ε_1 and ε_2 are constants.

The fact, that the dielectric displacement D can be expressed, in the manner here given, is the most important practical result of its interpretation by means of after-effect.

The following is valid for the current density:

$$I = \frac{1}{4\pi}\frac{dD}{dt} = -\frac{D_0}{4\pi}\omega \sin(\omega t - \delta) = \frac{D_0}{4\pi}\omega \cos\{\omega t + (90° - \delta)\}.$$

Since the current is not exactly 90° in phase ahead of the voltage, but only 90°—δ, heat development occurs in the substance, the amount being determined by the magnitude of δ. When we express graphically the relation between D and E according to (8) and (9) we obtain a slender ellipse as drawn in *fig. 6*. Its width is determined by δ, while its area indicates the energy development per period.

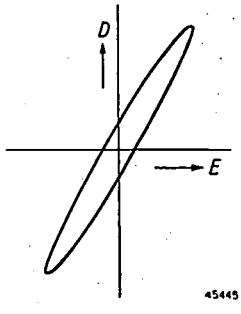


Fig. 6. Relation between the dielectric displacement D and the field strength E for a dielectric, which is placed in a periodic electric field and which is not free of losses.

When we are not interested in the energy development alone, it is often clearer to describe the influence of the angle of loss in a different way.

It must be noted, that the same phase difference 90°—δ would have occured between current and voltage, if a condensor with a loss-free dielectric were connected in series or in parallel with a certain resistance. This apparent series or parallel resistance, which presents difficulties in various electrical connections, is a graphic measure of the harmful effect of the losses.

Attempts have been made, for more than fifty years, to make the phase shift δ as small as possible. The requirements, which must be made, are often very severe. Thus a value of $\tan\delta$ of 0.01 must already be considered large, although this corresponds to a phase angle of less than one degree. In the best solid dielectrics, now known, a value of δ of about 0.5 minute has been attained ($\tan\delta \approx 10^{-4}$).

It must, however, be noted that just in these technically much used substances the after-effect must often be described by expressions analogous to (7), thus by:

$$D = D_0 \cos(\omega t - \delta),$$

with:

$$D_0 = E_0\left[\varepsilon_1 + \sum_n \frac{\varepsilon_n}{1+\tau_n{}^2\omega^2}\right] \text{ and}$$

$$\tan\delta = \sum_n \frac{\varepsilon_n}{\varepsilon_1}\frac{\tau_n\omega}{1+\tau_n^2\omega^2}.$$

The variation of $\tan\delta$ with the frequency is then often, such as that indicated by the broken line in figure 2, in other words, $\tan\delta$ depends little on the frequency.

Cause of the dielectric after-effect

It is assumed, that the after-effect losses are generally caused by „imperfections", namely either by regions in the substance which posses a certain (slight) electrical conductivity, or by dipoles, present in low concentration in the substance, which can be oriented by an electric field with a certain time lag.

It is not difficult to understand, that a substance with these properties would exhibit just the phenomena observed. A suddenly applied electric field first causes spontaneously the „ordinary" dielectric displacement E, while afterwards the quantity $D(=E+4\pi P)$ gradually increases farther, either due to the orientation of the dipoles, which increases the polarisation P, or due to the fact, that in the above-mentioned regions the positive charge accumulates on one side and the negative on the other, likewise resulting in an increase of P.

A medium, in which all the dipoles or all the regions possess equal properties, will then exhibit a simple after-effect curve according to (1), while the complicated case occurs, when the dipoles are mutually different as regards their time of adjustment, or the regions in their conductivity.

In both case, the after-effect is based upon processes, which are closely related to diffusion. Because when a dipole, consisting as it does of a positive and a negative part, changes its orientation, at least one of the two parts must change its place in the lattice, which displacement takes place only gradually under the influence of the electric field and thermal agitation, as is characteristic of diffusion. In the other case we are concerned with a diffusion of ions or „diffusion" of electrons.

Ferromagnetic after-effect

While the existence of the dielectric after-effect is quite generally known, the same can certainly

not be said of the magnetic after-effect, which is, nevertheless, also of great importance in alternating current technology. In magnetic materials, aside from the eddy currents, the more obvious phenomenon of hysteresis has drawn the main attention.

In connection with the development of new materials, however, it is important to pay especial attention to the after-effect.

Because of the fact that, in a ferromagnetic substance all three of the effects mentioned usually occur at the same time, the first important task is to be able to separate these effects from each other. Fortunately in this respect, with a suitable form of the ferromagnetic substance, the influence of the eddy currents on the phenomena in an alternating current circuit can be determined theoretically. This influence can thus be separated from the other two effects. Moreover, because this influence can always be kept small by employing the substance in lamellar form, we shall in the future disregard the eddy currents.

We shall now compare the two other effects, hysteresis and after-effect with each other, which comparison will furnish us with a means of determining the influence of the two phenomena separately. We assume, that the substance is placed in an alternating magnetic field. The hysteresis characterized by the fact is that the phase of the magnetic induction B is behind that of the field H, The relation between B and H being given by the hysteresis loop (fig. 7*a*).

In hysteresis, as well as in the after-effect, energy is developed in the substance, and this is given by the area of the hysteresis loop or of the ellipse, respectively. There are thus various points of correspondence.

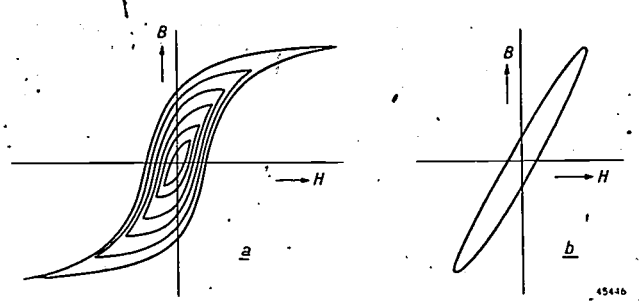


Fig. 7. *a*) Hysteresis curves of a ferromagnetic material. The variation B is drawn, when the field is varied from $+H$ to $-H$ and back again. The different curves are for different amplitudes of the variations of the magnetic field *b*) Relation between the magnetic induction B and the field strength H for a ferromagnetic substance placed in an alternating magnetic field and which exhibits after-effect only. Usually hysteresis and after-effect occur together, which makes the phenomena less clear.

An important difference between the two effects occurs, however, when we make the amplitude of the field smaller. The ellipse, as well as the loop, then naturally become smaller. The ellipse, however, retains its original proportions. The hysteresis loop, on the other hand, is deformed and approaches a straight line; its area thus decreases much more rapidly than of the ellipse. With very small amplitudes of the field, we thus retain only the influence of the after-effect, which consists in a phase shift and consequent heat development. With the help of this property we can separate the two effects.

Another phenomenon may also occur when the frequency of the field, instead of its amplitude, is changed. The after-effect ellipse may then become broader or narrower, as in the dielectric case, because, at least in certain frequency regions, the size of the phase angle depends upon the frequency. The hysteresis loop in this case remains unchanged.

Summarizing, we may therefore say, that the relation between B and H in the case of hysteresis and after-effect exhibits a difference, firstly in the shape of the loop and then also in the way in which the loops alter upon change in amplitude or frequency of the alternating field.

In order to be able now to investigate the phenomena of magnetic after-effect separately, we shall always work with extremely low field strengths, where hysteresis is a minimum, while a correction will be introduced for the influence of eddy currents.

Suppose now, that we perform the following experiment. The substance is suddenly placed in a weak magnetic field. It is then found, that the induction B shows a variations with time like that, which we have already encountered several times, and which can be represented by a curve like that of fig. 5, *i.e.* a sudden increase followed by a gradual increase. (The variation of dB/dt, which determines the induction voltage in neighbouring conductors, is entirely analogous to that of dD/dt, which is also drawn in the figure).

From this we again conclude in the familiar way that with a periodical variation of the field:

$$H = H_0 \cos \omega t, \quad \ldots\ldots \quad (10)$$

the following is valid for the induction:

$$B = B_0 \cos (\omega t - \delta) \quad \ldots\ldots \quad (11)$$

The validity of this expression, for B, is the most important practical result of the after-effect. The relation between B and H, as may be seen from (10) and (11) is given by the ellipse of fig. 7*b*, as already mentioned. The occurrence of the phase shift δ results in the fact, that in a coil in which the substance in question is employed the alternating

current will no longer be 90° in phase behind the voltage, but slightly less. Just as in the dielectric case, it may be said here also that the same phase difference would occur if the ferromagnetic substance were free of losses and the coil connected with a certain resistance in series or with a certain resistance in parallel. This apparent series or parallel resistance, which in practise is often objectionable, is a measure of the harmful effects of the losses.

Cause of the ferromagnetic after-effect

Before we discuss the cause of the after-effect for a single thoroughly investigated case, we shall first briefly recall the remarkable nature of the structure of a ferromagnetic substance with not too large a coercive force.

Such a substance is built up of little regions each of which is completely magnetized, but whose directions of magnetization are very different. In a single crystallite these dir ctions are at angles of 90° and 180° with each other, since the magnetization occurs according to the cube directions in the crystal. Where two such regions, whose dimensions are about 10^{-4} cm, border on each other, a gradual transition occurs from the magnetization direction of the one to that of the other region. If the substance is now placed in a weak magnetic field, this transition region, the „wall", is slightly displaced in such a way, that the region whose direction of magnetization corresponds more closely to the field becomes slightly larger and the other slightly smaller, whereby the substance becomes magnetized. The direction of magnetization is thus altered at certain spots and this causes a unilateral deformation of the crystal as has been demonstrated by experiments on magnetostriction. The elastic forces tend to prevent this deformation, which stretches the lattice in one direction and compresses it in the perpendicular direction. It is therefore, remarkably enough, the elastic forces which limit the extent of the magnetization brought about; the way in which this is done is, however, not known with certainty.

When the field is increased to such a value, that the wall would be displaced over a certain critical distance, the wall suddenly shoots forward and the magnetization in the remainder of the „defective" region as a whole changes over to the direction of the „better" region. This effect lies at the bottom of hysteresis. In connection with the following it must be noted, that a change of 90° or 180° in magnetization causes no change in the mechanical tensions in a given crystallite. The changes in tension upon magnetization thus do not appear in the regions, but only in the transitional layers, so that only hysteresis accompagnies the „change-over" in direction.

After this outline of the structure of ferromagnetic substances, we return to the causes of the magnetic after-effect. The most thoroughly investigated case is that of the above-mentioned iron wire, where the following conception has been arrived at.

When this substance is suddenly placed in a weak field, or more generally, when the field H_0 in which the substance is situated is suddenly increased by h, the „walls" are suddenly shifted, and therefore local deformations in the lattice suddenly occur: a stretch in one direction, a compression in the other. The carbon and nitrogen particles situated between the iron atoms therefore diffuse from the compressed spots, where there is little room left, to the stretched spots. This, however, causes at the same time a decrease in the elastic tensions and the field can then gradually displace the wall somewhat farther, increasing the magnetization thereby, which is exactly what is observed in the after-effect. The remarkable feature of the description here given of the magnetic after-effect [3]) is that the actual cause lies in the elastic after-effect of the material.

It must be noted that here also it is not always the simple relation (4) which is observed for tan δ, but often an expression like (7) occurs, with a much flatter shape as a function of the frequency. In this case also, perhaps the cause must be sought in analogous diffusion phenomena.

Thus as we have just seen, it is not necessary that the after-effect in magnetic materials should be based on „purely magnetic" effects; usually indeed this is not the case. If there is any coupling between the effect being investigated and another which exhibits after-effect, after-effect phenomena will also appear in the first effect. This situation is often encountered in the magnetic case.

Another example is the following. When a so-called powder core, *i.e.* an aggregate of magnetic particles separated from each other by a thin insulating layer, is magnetized, elastic forces will in general appear as a result in the insulating layer. Now, if the insulating layer exhibits elastic after-effect, this will again be manifested in the occurence of a phase difference in the magnetic phenomcnon.

---

[3]) Thcrc exists, however, still another magnetic after-effect, that of Jordan, for which this does not hold. We merely mention this fact, without going into any details.

It is understandable that the detection of the causes of losses in often made very difficult by such couplings. In the meantime the examples mentioned may suffice and we shall not discuss any other examples.

Our final conclusion is, that the material may exhibit after-effect under the influence of mechanical as well as of electrical or magnetic forces, while the cause must often be sought in a diffusion process. In the case of periodical electrical or magnetic fields the most typical manifestation of after-effect is the phase difference exhibited by the dielectric displacement D or the magnetic induction B with respect to the field, and the losses connected with this phase difference. The diminution of this phase difference is therefore of great importance in electrotechnology, and this has already been succesfully accomplished in various cases.

Contents of *Philips Research Reports 1*, No. 2. January 1946

VOL. 8 No. 3, p. 65-96 MARCH 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

BLOCKING-LAYER PHOTOCELLS

by W. CH. VAN GEEL.

621.383.5

Certain semi-conductors, such as copper oxide and selenium, in combination with a metal and separated from it by a very thin layer of insulating substance (blocking layer), exhibit upon illumination the so-called blocking-layer photoeffect. Upon illumination a photocurrent occurs without an auxiliary voltage being applied. In this article the action of such blocking-layer photocells and the mechanism of the phenomena occurring therein are discussed. In particular the selenium photocell is described in detail.

When light is allowed to fall on the alkali metal electrode of an evacuated tube or one filled with a rare gas, electrons are liberated. These electrons move in the surrounding gas or vacuum of such an alkali-metal photocell. By applying a suitable auxiliary voltage with respect to a metal anode, these electrons liberated by the so-called external photoelectric effect [1]) can be made to move toward the anode of the cell and thus give rise to a photoelectric current. In this way, however, only little electrical energy is obtained from the incident light, and it will usually be necessary to amplify the photocurrent generated before it is used.

In the case of the blocking-layer photocells to be discussed in this paper, however, the incident light is converted into electrical energy without it being necessary to apply an auxiliary voltage. The so-called blocking-layer photoeffect consists in the fact that the light causes electrons to move from a semi-conductor through the blocking-layer to the contiguous metal, from which they may then return to the semi-conductor through an external resistance (for example an ammeter). In this case, without further assistance a photocurrent of measurable size is immediately obtained, a fact which has of course very much promoted the employment of such blocking-layer photocells.

Although the blocking-layer photoeffect was discovered by Adam and Day in 1876, while it was again found by Fritts in 1884, blocking-layer photocells have actually only been developed within the last ten years to the state of technically useful instruments. Since then they have found extensive use in the measurement of intensities of illuminations as applied in photographic exposure meters, luxmeters, photometers, colorimeters and the like.

Structure of the blocking-layer photocell

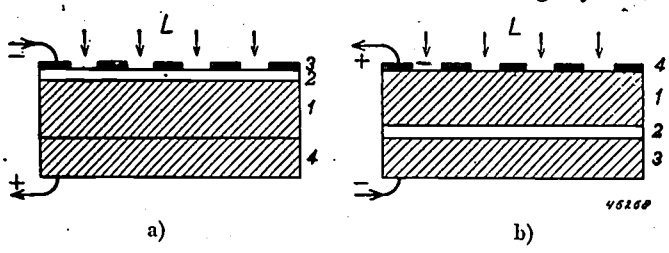
The construction of a blocking-layer photocell is in principle similar to that of a blocking-layer rectifier [2]). In both cases we are concerned with a metal layer and a semi-conductor which are separated by a thin insulating layer, the so-called blocking layer. In order to obtain a blocking-layer photocell, it is now necessary to make the counter electrode or the metal layer which makes contact with the semi-conductor so thin that it is transparent. The light can then penetrate into the semi-conductor and there free electrons.

Two kinds of blocking-layer cells can be distinguished according as the blocking-layer lies in front of or behind the semi-conducting layer and the electrons leave the side of the semi-conductor upon which the light is incident (front-wall cell) or the other side of the semi-conductor (back-wall cell). In *fig. 1* the principles of these two possibilities are shown diagrammatically. In this figure *1* is the semi-conductor and *2* the blocking-layer, while *3* is the metal counter electrode toward which the photoelectrons from the semi-conductor move through the blocking-layer. The layer of metal in contact with the semi-conductor, which serves the purpose of causing the return of the electrons to the semi-conductor after traversing the circuit, is indicated by *4*. The light is incident

[1]) Philips techn. Rev. 2, 13, 1937.

[2]) Philips techn. Rev. 4, 100, 1939.

in the direction of the arrows, so that in the case of the front and back-wall cells, respectively, the metal counter electrode *3* and the blocking-layer *2*,



a) b)

Fig. 1. Diagrammatic representation of the structure of a frontwall cell (*a*) and a backwall cell (*b*). *L* incident light, *1* semi-conductor, *2* blocking layer. The metal layer *3*, serving as a counter electrode is in *a* very thin and transparent, while in *b* it functions as base plate. In case *a*, however, the metal layer *4* making contact with the semi-conductor is the base plate and in case *b*, on the contrary, it is transparent.

or the contact metal layer *4* should be transparent enough to ensure a satisfactory functioning of the cells. As may be seen in fig. 1, the electrons move in the same direction as, and in the opposite direction to, that of the incident light in the front and back-wall cells, respectively.

At present two different main types of blocking-layer cells are being made, namely cuprous oxide and selenium cells; the latter are also made by Philips. The cuprous oxide cells are constructed either as front-wall cell or as back-well cell, selenium cells, on the other hand, only as front-wall cell. In order to make a front-wall cell of cuprous oxide (Cu_2O) one begins with a plate of cuprous oxide containing an excess of oxygen which lends it conductivity. This plate is then both *1* and *4* of fig. 1*a*. By reduction or ion bombardment the surface of this plate is treated in such a way that a blocking-layer *2* is formed on it of pure Cu_2O, which is an insulator. This latter is transparent and must finally be covered with a transparent metal layer *3*, towards which the electrons freed from the semi-conductor by the light will pass. The metal layer then becomes electrically negative and functions as counter electrode.

If on the other hand one wishes to make a back-wall cuprous oxide cell, one begins with a copper base plate whose upper surface is covered with a layer of cuprous oxide (Cu_2O) containing an excess of oxygen by being heated to a high temperature (1040 °C). This combination is then (acc. to fig. 1*b*) the counter electrode *3* and the semi-conductor *1* which are separated by a layer of pure Cu_2O which functions as blocking-layer *2*. Upon this is finally deposited a transparent layer of metal *4* which provides for the electrical contact with the semi-conductor *1*. The electrons freed by the light in the semi-conductor *1* now pass through the blocking-layer *2* to the metal base plate *3* and then return *via* the external circuit to the transparent metal layer *4*, which is in contact with the semi-conductor *1*.

As already mentioned, the selenium cell which is manufactured by Philips is a front-wall cell and the order of the layers is therefore as indicated in fig. 1*a*. Upon a base plate *4* of aluminium provided with a prepared surface a semi-conducting layer *1* of selenium is deposited. For this purpose the selenium has been prepared in the so-called grey modification by means of a heat treatment. The upper surface of this selenium is then provided with a blocking-layer *2* upon which a very thin transparent layer of metal *3* is finally deposited by cathode sputtering. This last layer functions as counter electrode. In most selenium cells the blocking-layer *2* is obtained from the semi-conductor *1* itself by causing a chemical reaction on its surface, which produces a non-conducting surface layer. Since, however, this blocking-layer does not completely cover the selenium, in the Philips selenium cell an additional separate blocking-layer is deposited, which layer has no chemical relation with the semi-conductor. As a result not only is a better blocking-layer obtained, but at the same time there is the advantage that the metal of the thin layer *3* which is deposited by evaporation on the blocking-layer *2* cannot react with the selenium, since the two layers are now entirely different chemically. The selenium cells with a separate blocking-layer can therefore be used much longer than those which do not have such a layer [3]).

Moreover, it is also an advantage of the separate blocking-layer that with it it is much easier to alter the characteristics ot the blocking-layer according to necessity. If, for example, it is made thicker, the photocell can be given a smaller capacity and at the same time the internal resistance can be increased, which may be desirable for certain applications.

Voltage and current of the blocking-layer photocell

When light passes through the transparent metal layer *3* and the blocking-layer *2* to the semi-conductor *1*, two different phenomena take place. Not only do electrons move from the semi-conductor through the blocking-layer to the metal counter electrode (blocking-layer photo-effect, as has already been explained), but at the same time the electrons freed in the selenium by the light produce a higher

[3]) In the case of the blocking-layer rectifiers which are manufactured by Philips, both of these blocking-layers are therefore also present.

conductivity [4]) of this semi-conductor, which is called the internal photoeffect. Everyone suspects that there is a relation between these two effects, but it is still certainly a question whether or not the electrons which in the blocking-layer effect pass from the selenium to the blocking-layer are first freed in the selenium by the internal photoeffect. While the blocking-layer photoeffect occurs immediately upon illumination, the internal photoeffect shows inertia.

When the two electrodes (*3* and *4*) are connected externally, for instance by an ammeter, upon illumination the electrons pass from the counter electrode *3* through the measuring instrument, which usually possesses only a low resistance to the metal layer *4* which makes contact with the semi-conductor *1*. The direction of the electric current is there just the opposite, as is indeed indicated by arrows in fig. 1. In all types of blocking-layer cells, therefore, upon illumination the counter electrode *3* becomes negative with respect to the semi-conductor *1*. Due to this difference of potential over the blocking-layer *2*, the electrons will now have the tendency to return from the metal counter electrode to the semi-conductor. This is indeed the direction of good transmission for a rectifier. If there were no question at all of a returning current, one would be concerned only with a pure photocurrent, which would be directly proportional to the intensity of illumination on the photocell, as is represented by the straight line *a* in *fig. 2*. In addition to this, however, there is the oppositely directed conduction current, which may for example be represented by curve *b* in fig. 2 and which is found to vary approximately as the square of the voltage between the electrodes *3* and *4*. The total current through the external circuit thus finally takes on the shape represented by curve *c* in fig. 2.

Upon closer consideration, however, the concept just given is found to be still too much simplified. The conduction current *b* which flows in the opposite direction and is due to the voltage over the blocking-layer is not determined exclusively by this voltage, but is found to depend also upon the degree of illumination to which the photocell is exposed. The work necessary to take one electron from the metal counter electrode into the blocking-layer (the so-called work function) is somewhat smaller due to the illumination, so that for a given voltage the opposing conduction current *b* will be somewhat higher for the illuminated cell than for the non-illuminated cell.

In the foregoing we have continually spoken of a current which the photocell can send through an external circuit, but it may be of importance to know also what voltage it can deliver under

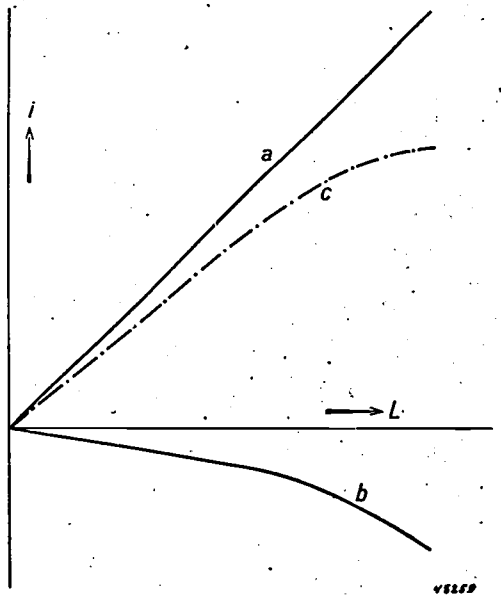


Fig. 2. The current generated i as a function of the illumination L of the photocell: a is the photocurrent which flows with no external resistance, b is the oppositely directed conduction current which is a result of the voltage over the blocking layer and c is the total current which flows through the external resistance.

different circumstances. If the voltage is measured with a voltmeter which consumes no current, or, more simply, if the voltage is calculated from the current and a sufficiently high external resistance, one obtains a voltage, when the illumination is not too intense, which is practically proportional to the intensity of illumination, because both the photocurrent and the counter voltage excited are then practically linear with that intensity. With more intense illumination, however, the counter voltage increases about quadratically with the illumination, so that the total voltage then increases less rapidly than proportional to the intensity of illumination.

Mechanism of the electron movement in blocking-layer photocells

We shall now attempt to explain somewhat more clearly the mechanism of the conduction by electrons in blocking-layer cells.

When free electrons are introduced from outside

[4]) Use is also made of this in technology. Thus for example a selenium cell, which in principle consists of a piece of selenium with two electrodes, can be put under voltage in a closed circuit. When this cell is illuminated the internal resistance decreases and the current consequently increases. (*Cf.* for instance Philips techn. Rev. 2, 13, 1937). By means of this change in current a relay may be closed. This type of selenium cell is often called a resistance photocell.

into a substance which is itself an insulator, the electrons can move in that substance, so that conduction is then possible in such an insulator. According to modern conceptions about the electrical structure of solid matter, however, it is not possible to introduce electrons with any arbitrary amount of energy from the outside. In a solid substance there are alternating regions of energy values (so-called energy bands) which are permitted or not permitted for the electrons which attempt to enter the substance. In *fig. 3* it is indicated diagrammatically that electrons of the energy bands A and B may occur in the solid, but

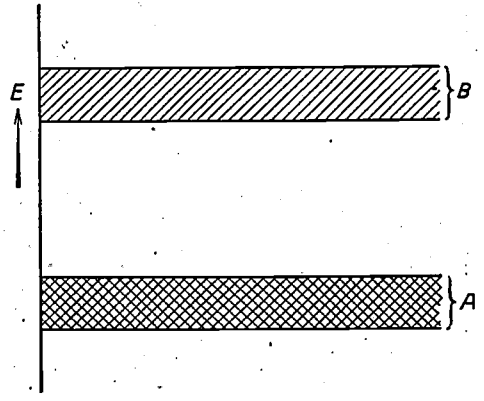


Fig. 3. Diagram of the electron states in an insulator. The energy bands A and B of the electron levels are permitted while between them lies a region which is forbidden for the electron energy. The cross-hatched band A is completely occupied and the single hatched band B is unoccupied.

the energy levels lying between them form a forbidden region.

In one definite energy state [5]), according to modern quantum theoretical conceptions, there may exist only one electron in every elementary cell of the crystal. An energy state may therefore be "occupied" or "unoccupied". If for example in the energy band A all the states are occupied, which has been indicated in fig. 3 by cross hatching, while the singly hatched band B is entirely unoccupied, it is not possible with the help of an electric field to give a small increase of energy to an electron, since all other states with energy values differing only slightly from the original value are already occupied and in order to attain an unoccupied energy state in band B a much too high amount of energy would be necessary! This is thus the case of an insulator in which no conduction of electricity can take place. However, by irradiation with light of a sufficiently high frequency ν an energy quantum $h\nu$ can, according to Einstein, be given to an electron existing in an energy state belonging to the fully occupied band A, by which that electron is brought into a higher state lying in the originally unoccupied energy band B. For electrons existing in energy states of the only partially occupied band B it is now possible, as a result of an external electric field, to pass over to other energy states which lie only slightly higher in the same band B. In this way the insulator has been made conductive by illumination, a phenomenon which was discussed in the foregoing as internal photoeffect.

With the help of the concept of energy bands

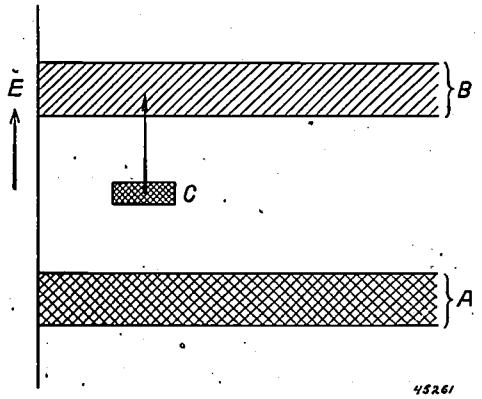


Fig. 4. The cross-hatched energy band A is occupied and the singly hatched band B is unoccupied. The likewise occupied energy band C is produced for instance by an impurity which makes the non-conducting substance a semi-conductor.

here introduced we can now classify solid substances as to their electrical conductivity. If we are concerned with a full and an empty band, whose difference in energy values is large with respect to the average energy kT of thermal agitation, then the latter is not capable of causing the electrons to pass from the full to the empty band where they might be able to contribute to the conductivity. This is thus the picture, which we must imagine, of a perfect insulator. Now by adding a foreign substance to such an insulator it is possible also to create electron states whose energy values lie in the region C between A and B, as indicated in *fig. 4*. If the energy band C caused by the contamination now lies close below the unoccupied band B it is possible, that the average thermal agitation of the electrons will be capable of taking electrons from band C to states in band B, but this cannot take place on a large scale since there are only few energy states in band C available, due to the fact that the "contaminations" usually occur as a small percentage of the substance. In this case we are dealing with a semi-conductor.

When, however, the occupied and the unoccupied bands pass from one to the other without interruption we are con-

[5]) Such a state is characterized not only by the value of the energy but also in general by various other quantities, such for example as the characteristic impulse moment of the electrons, the so-called "electron spin".

cerned actually with a partially occupied band B. In this case a continuous change in the electronic energy is possible and thus also electrical conduction. We have in this way arrived at a concept of a conducting metal. At the same time it is clear from this representation of the state of affairs that upon increase of temperature (larger kT) the average thermal agitation of the electrons will be better able to bring an electron from a lower occupied state into a higher unoccupied one, so that as a result of the partial occupation of the higher energy band B thus produced the electronic conductivity of insulators and semi-conductors will increase at higher temperatures, although it may in general be disregarded compared with the conductivity of metals.

For the combination of a semi-conductor and a metal which are separated from each other by a thin blocking-layer of an insulating material, we arrive, for the state of equilibrium, at approximately the circumstances shown diagrammatically

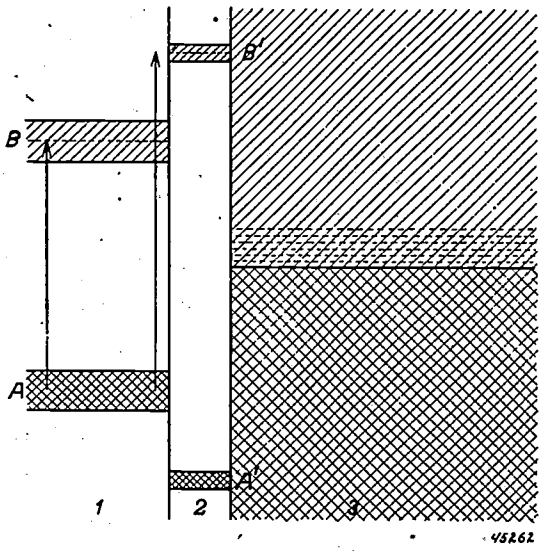


Fig. 5. Diagram of the position of the energy levels in the combination of a semi-conductor *1* with a metal *3* separated by a blocking layer *2*. A' and B' occupied and unoccupied energy bands, respectively, in the blocking layer. In the metal we are concerned with a continuum of permitted energy levels, which are completely occupied (cross-hatched) up to a certain limiting energy and above that only very sparingly occupied (single hatched).

in *fig. 5* for the position of the energy levels. In the case of a selenium cell, which is a front-wall cell as shown in fig. 1*a*, the light (incident from the right in fig. 5) penetrates through the thin layer of metal and blocking-layer into the semi-conductor and will there be able to take the electrons out of the occupied band A to the empty band B, so that those electrons will now increase the conductivity of the selenium (internal photoeffect). The electrons from the full band A of the semi-conducting selenium may, however, also be brought by the light into the still higher unoccupied band B' of the blocking-layer and in that way make that insulator slightly conducting. They then move through the blocking-layer to the transparent metal counter electrode and in that way produce the blocking-layer photoeffect. The amount of energy $B'—A$ which an electron must receive from the incident light for this purpose is called the work function for the electron in the case of the blocking-layer cells and it is quite analogous to the work function of the electrons in the case of alkali-metal photocells, which in that case is necessary to take the electrons out of the alkali metal into the vacuum. The insulator in the blocking-layer photocell thus plays exactly the same part as the vacuum in the case of the alkali-metal photocell. When the photoelectrons once move in the blocking-layer into an energy state of the unoccupied band B', they will pass from there into the continuum of slightly occupied (singly hatched) energy states of the metal, so that the circuit can be closed if we connect the metal by a conductor with the selenium [6]).

For the emission of electrons by the light-sensitive semi-conductor to the blocking-layer the same equation of Einstein is valid as for the emission of electrons in alkali-metal photocells:

$$h\nu - h\nu_0 = {}^1/_2\, m\, v^2 = eV_m \quad . \quad . \quad . \quad (1)$$

In this equation ν represents the frequency of the light which is incident on the blocking-layer cell and ν_0 the minimum frequency necessary to bring the electrons out of the occupied band A of the semi-conductor into the unoccupied band B' of the blocking-layer. The remaining energy is thus possessed by the conduction electrons in the blocking-layer as kinetic energy: ${}^1/_2\, mv^2$. Finally V_m represents the minimum counter voltage which must be applied in order to prevent the electrons leaving the semi-conductor and passing to the blocking-layer; this is thus the so-called blocking voltage.

Since selenium absorbs visible light very strongly the latter can penetrate to only a small depth into selenium, so that the electrons are freed in the semi-conductor in a region which lies close to the blocking-layer. There is thus reason to suppose that the electrons which produce the blocking-layer photoeffect are freed from the semi-conductor by a

[6]) If in this connection the question should be asked why electrons are taken by the light only out of the semi-conductor and not out of the metal into the unoccupied band B' of the blocking-layer, the answer is that in the metal itself such a large number of unoccupied states are present for the electrons which might take on such an energy from the light, that they will have practically no chance of just entering that band B of the blocking-layer, while on the other hand for an electron with that energy no permissible energy state is available in the semi-conductor, so that that electron must indeed be taken up in the unoccupied band B' of the blocking-layer.

photoeffect in the way just described, and that the mechanism described, at least in the main, satisfactorily accounts for the phenomena of conduction by electrons in blocking-layer photocells. It is, however, difficult to give a satisfactory explanation of the way in which these phenomena depend upon the temperature, and we shall not go into it here. We only call attention to the fact that, upon the application from the outside of a negative voltage on the selenium, a very strong photoeffect occurs in which phenomena of inertia appear, so that in this case there is obviously a close relation with the internal photoeffect which also shows a time lag.

Data of the selenium cell

In *fig. 6*, current characteristics are given of the selenium cells of 2 by 4 cm manufactured by Philips. It may be seen that, at least upon illumination with white light, with an external resistance of 10 ohms, there is a linear relation between the electric current and the intensity of illumination, which even at 2500 lux still shows no tendency toward saturation. With 100 ohms a slight saturation appears from 2000 lux onwards, due to the already discussed conduction current in the opposite direction. With still higher external resistance this saturation effect makes itself more and more clearly felt upon stronger illumination, so that with an external resistance of at least 1000 ohms one can indeed scarcely speak any longer of a linear initial section of the current characteristic. With an external resistance of for instance 11 000 ohms, even with an illumination of 2500 lux, the current becomes no higher than 15 μA, which fact can no longer clearly be shown in fig. 6.

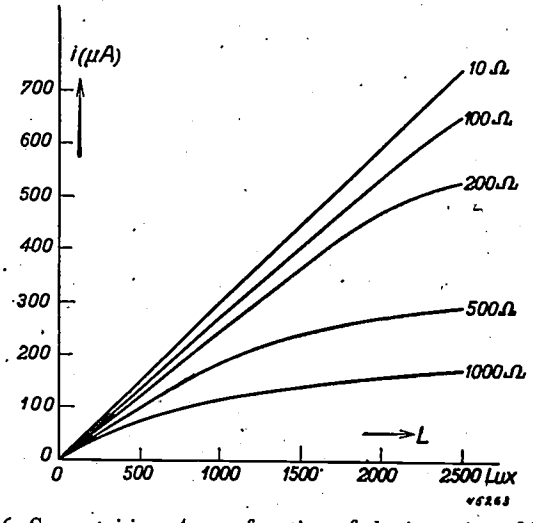


Fig. 6. Current i in μA as a function of the intensity of illumination L with white light at different values of the external resistance.

The behaviour of the voltage over external resistances of 500, 1000, 2000 and 11 000 ohms is shown separately in *fig. 7*.

In *fig. 8* the spectral sensitivity of the selenium cell is shown, *i.e.* the relation between the current which is produced upon illumination with a given energy of an arbitrary wave length and the current which is obtained with the same energy of illumination with a wave length of about 5500 Å. The latter is the wave length for which the selenium cell is most sensitive; this happens to coincide

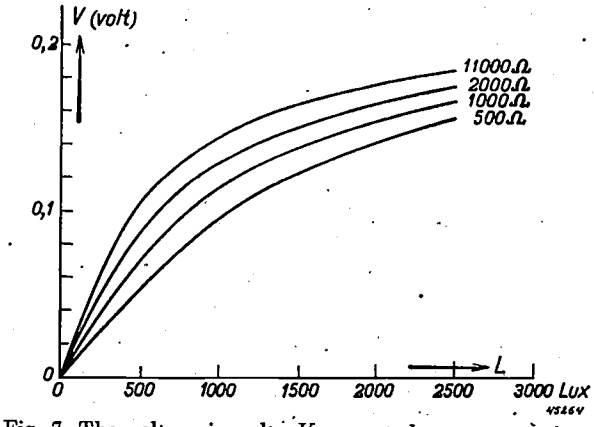


Fig. 7. The voltage in volts V generated over external resistances as a function of the intensity of illumination L in lux.

practically with the wave length for which the human eye is most sensitive. For the sake of comparison the eye-sensitivity curve is drawn as a dotted line in fig. 8. The shape of these two curves is the same on the long-wave side. On the short-wave side, however, the sensitivity of the photocell does not decrease nearly so rapidly as the sensitivity of the eye. This too great sensitivity of the cell for the shorter waves can easily be compensated by using suitable filters. In this way it is possible to imitate very well with the blocking layer photocell the spectral variation of the sensitivity of the eye. The selenium cell with filters

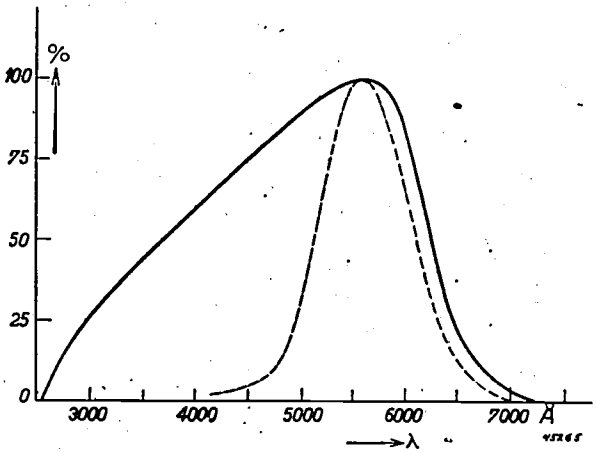


Fig. 8. The continuous curve represents the spectral sensitivity of the selenium cell in comparison with the curve of eye sensitivity shown as a dotted curve.

is therefore very suitable for use in the measurement of illumination intensities.

With a very low external resistance the current is found to be independent of the temperature, but at higher external resistances such a dependence does occur, since the resistance for the counter current depends upon the temperature. The size of the temperature coefficient of the current depends very closely upon the way in which the cell is constructed, but also upon the intensity of illumination. The current may vary by 1 to 10 per thousand per degree of change in temperature. The temperature coefficient is found to increase with the illumination intensity and with the external resistance.

In conclusion we shall briefly examine how great the power is which a blocking-layer photocell can deliver under different conditions. In *fig. 9* the power delivered at different illumination intensities is represented in μW as a function of the resistance in ohms. For every illumination intensity there is an external resistance for which the power delivered is a maximum. This value of the external resistance at which the maximum power is delivered is found to be smaller, the higher the illumination intensity is chosen.

In order to generate a higher power a number of photocells can be connected in parallel, so that with a given voltage one obtains the sum of the separate currents. As already stated, the photocells furnish in the first instance a photocurrent, while the photovoltage is a phenomenon which occurs over the external resistance through which the current flows. The cells must not therefore be connected in series, because in that case one only obtains a photocurrent equal to that of the least sensitive cell.

If one converts radiant energy of the sun into electrical energy by means of a blocking-layer cell, the efficiency proves to amount to only about 2 percent. The power of a large cell with a surface of about 150 cm² in the full sunlight is found to be only just enough to make a small incandescent lamp burn or to drive a small motor; for the generation of higher electrical powers the blocking-layer photocell is not suitable.

Applications of the selenium cell

One of the most familiar applications of the blocking-layer cell is its use as photographic exposure meter. With a given photographic material the blackening attained by that material upon exposure during a certain time is proportional to the logarithm of the incident intensity. Now in the case of the blocking-layer cell, when an external resistance of about 500 ohms is used, the current generated is also practically proportional to the logarithm of the intensity of the incident illumination. We have already observed this in the character of the curves of fig. 6, while in *fig. 10* the curve for 500 ohms is plotted with a logarithmic scale of the illumination intensity and proves to be almost straight. We may therefore say that the electric current generated for a given illumination intensity is a suitable measure of the blackening of the photographic material, and that a good photographic exposure meter may be based upon this principle.

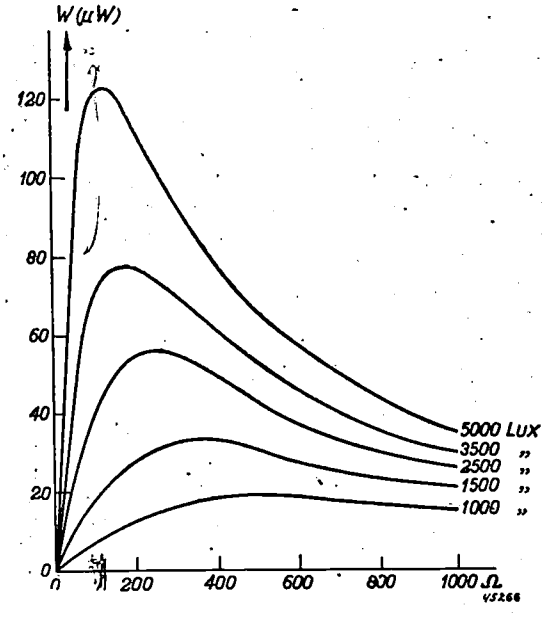


Fig. 9. The power in μW delivered by a Philips selenium cell as a function of the external resistance in ohms at different light intensities in lux.

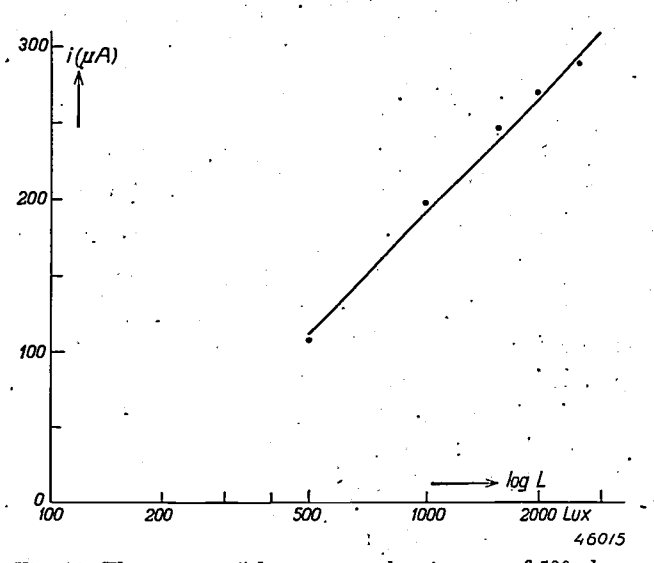


Fig. 10. The current with an external resistance of 500 ohms proves to be approximately a linear function of the logarithm of the illumination intensity in lux.

INCANDESCENT LAMPS FOR FILM PROJECTION

by Th. J. J. A. MANDERS.

621.326.73

Because of the characteristics of the optical system in which film projection lamps are used, the following requirements must be made of such lamps: The filament must have great brightness and must fill a given small area as completely as possible, the diameter of the bulb must be small and the lamp must be able to be adjusted automatically. On the basis of examples from among the numerous types of cinema and 8 and 16 mm film lamps manufactured by Philips, it is explained how these and other requirements are satisfied. Particularly the construction of the filament and the part played by the gas filling of the lamp are discussed from different points of view.

Cinema lamps and 16 mm film lamps

When in addition to the normal film 35 mm wide, used in cinemas, the narrower film (originally $9^1/_2$ mm, later 8 and 16 mm also) began to be used, it seemed at first as if not only the field of application but also the projection apparatus for the two types would differ in principle. The main problem in projection technique is that of passing the large light flux, which must be thrown on the screen, through the small film window, whose area in the case of 16 mm film is about one fifth, and in that of 8 mm film about one twentieth of that for normal film. Narrow film was therefore predestined for use in the home or in small auditoria with a small projection screen, where less light was needed. And while normal film was projected with the help of arc lamps and mirror condenser, the narrow film projector for the small audience worked exclusively with incandescent lamps and lens condensers. Only in that way could they be made sufficiently small and cheap, safe in use, easily transportable and fairly foolproof.

At the present time, however, the former sharp boundary between normal and narrow film projection is becoming vaguer. This is due mainly to the fact, that the advantages of the narrow film (cheaper installation, lower film costs, possibility of the use of non-inflammable film material) have gained more territory for it which belonged originally only to normal film: the halls and auditoria in which 8 and 16 mm film is shown are growing larger and larger and approaching the size of ordinary cinema theaters, and non-portable apparatus is now being made for that size film. Another reason for this development was that the noiseless and quietly burning incandescent lamp, which was being manufactured for the growing 8 and 16 mm film projectors in types of steadily increasing power, was succeeding in displacing the unsteady sputtering carbon arc, even in normal film projectors. This was possible to a certain extent in smaller cinema, but not in large theaters, where, because of its greater brightness, the arc lamp continued in use. During the last few years before the war the super high-pressure mercury lamp appeared as a new serious rival of the arc lamp [1]), and the intrusive narrow film has in the meantime also taken possession of this aid in order to win new territory. The field of projection in auditoria, which are not too large, remains in the undisputed possession of the incandescent lamp, either with normal or with 8 or 16 mm film.

In discussing incandescent lamps for film projection, we shall not, according to the above, be able to draw any sharp boundary between lamps for normal film — called cinema lamps — and lamps for narrower film. There is indeed a difference in lifetime: according to universally accepted standards the lifetime of cinema lamps must be 100 hours, while in the case of narrow film lamps 50, 25 or even only 10 hours is considered sufficient. This difference is, however, based more on tradition than upon logical grounds. In principle the requirements for cinema and for narrow film lamps are the same, and the fact, that in practice each type of projection lamp is stated to belong to one or the other category is mainly due to the fact that in general each type of projection lamp can only be used in a certain type of projector, and that projector will naturally be intended either for normal film or for narrower film.

The optical system

In the process of projection the film picture occupying the film window is thrown on the screen as an enlarged image by the objective; see the right-hand part of *fig. 1*.

In the case of normal film the area of the film window is 3.17 cm^2, the area of the projection screen in a medium large cinema may be for instance 20 m^2. Thus even if all the light which falls on the film reaches the screen, the intensity of illumination of the screen will be more than 60 000

[1]) See Philips techn. Rev. 4, 2, 1939.

times as small as that on the film window. If for example it is desired to have 50 lux on the screen (75 or 100 lux is usually demanded), the intensity of illumination on the film, taking into account the losses by reflection and absorption in the objective (about 50 percent), must amount to about 6 million lux.

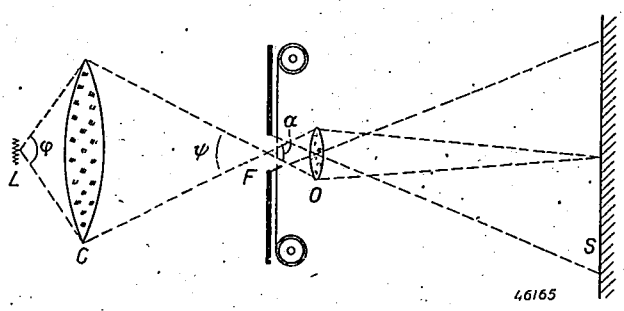


Fig. 1. The optical system used for film projection. The light from the filament of the lamp L is collected over a wide angle by the condenserlens C and focussed on or near the film window F. The film picture passing F is projected by the objective O on the screen S.

This enormously intense illumination is obtained by projecting the image of a very bright light source, in our case the filament of the projection lamp, on or near the film window by means of a condenser system; see the left-hand part of fig. 1. As may be seen in the figure the condenser has the function of collecting the light emitted by the source over a very wide angle φ and concentrating it on the film within a smaller angle ψ. The angle ψ should be equal to the angle α within which the objective can take up light from the film window. If $\psi < \alpha$, the objective is not entirely filled and if $\psi > \alpha$, unused light passes the objective.

Ordinarily the performance of a projector is characterized by the total light flux Φ which the objective throws upon the screen. The following formula can be derived [2]) in a simple way:

$$\Phi = k \cdot F \cdot B \cdot \pi \sin^2 a.$$

F is the area of the film window, B the (average) brightness of the light source and k a factor which takes all the losses into account, the chief of which are the losses in the condenser and the objective, the losses at the film window and the losses by the rotating sector which interrupts the light in the periods when the film is moved.

Requirements made of the light source

Formula (1) shows that in order to obtain a maximum light flux on the screen it is desirable to raise the brightness of the light source as high as possible. In addition the loss factor k must be kept as small as possible. In our case the losses at the film window constitute a large part of these losses. Due to the fact that the filament of an incandescent lamp is not a uniformly luminous plane, it may not be focussed exactly on the film window. If it were, the structure of that filament image would also be projected by the objective on the screen. The filament is therefore focussed farther away, in the objective or at a spot between objective and film window. In other words, the film window is not situated at the narrowest part of the beam of light produced by the condenser, and a larger or smaller part (70 percent for example) of this beam is cut off by the edges of the film window. The more uniform the brightness of the filament, the closer its image may be focussed to the film window and the smaller the loss at that point. The greatest possible uniformity of brightness of the filament is thus a second requirement.

Further requirements follow from a consideration of the condenser part of the projector. The light from the projection lamp is obviously used more economically the larger the angle φ compared to $\psi = \alpha$. If it is desired to limit the distance between source of light and film window to a certain length, in order to prevent the whole projector from becoming too large, a large angle φ means that the filament of the lamp must be able to be placed very close to the condenser, thus that the diameter of the envelope of the lamp must be made as small as possible. Furthermore with given values of φ and ψ a certain enlargement of the filament is obtained. Since the size of the image of the filament is fixed by the size of the objective to be filled, the minimum size of the filament is thus also prescribed. In general it is found that the filament must occupy a rectangle of between 5 and 10 mm width. The filament may be made larger, but there is in the first approximation no profit in that.

With the dimensions the position of the filament is also exactly prescribed. It is therefore desirable in constructing the lamp to take care that in inserting the lamp in the projector the filament automatically occupies the correct position in the optical system.

Finally the requirement about lifetime. The various accepted times were given above. Since the lifetime is ordinarily determined by the velocity of evaporation of the filament, the prescription of the lifetime means for the manufacturer in the first instance that he must choose a suitable working temperature. The brightness is also mainly given thereby. We shall see how the manufacturer, with

[2]) See the article referred to in footnote [1]).

the limited number of available degrees of freedom (nature and configuration of the incandescent body, gas filling), tries to increase the brightness as much as possible. In addition he must take care that the lamp also lives out its natural lifetime and does not succumb too soon to the mechanical vibrations to which it is exposed in the projector due to the jerking motion of the film transport. He must also take care that the light yield of the lamp does not decrease due to blackening of the envelope or deformations in the small, very rigidly fixed filament.

Construction of the filament

The use of spiralized wire

For the construction of the filament, according to the above, there are three primary requirements: it must cover a certain rectangle of from 5 to 10 mm width, its brightness must be as uniform as possible and the average brightness as high as possible.

In connection with the first and second requirements it might be thought that a tungsten ribbon would be the ideal solution, such as is used in the ribbon lamps developed for photometry. A very uniform luminous surface could then be obtained in a simple way. This would, however, be at the expense of the third requirement: compared with the spiralized wire used in electric lamps for the same lifetime the working temperature and consequently the brightness of the ribbon must be chosen considerably lower. This is because of the fact that the ribbon has a comparatively larger surface for evaporation, while, moreover, the tolerances in thickness in the rolling of ribbon are larger than in the drawing of wire. The permissible temperature which must be calculated, so that the thinnest spots will not succumb before the desired lifetime is reached, is thus further depressed in the case of the ribbon.

Also in brightness obtained at a given temperature of incandescence the ribbon is inferior to the spiralized wire, due to the fact that with the latter a certain "black body effect" occurs [3]).

Thus for the sake of the third requirement, great brightness, a spiralized wire is used, and with it the prescribed area is filled as well as possible. For this purpose the spiral is usually divided into a number of sections which are assembled parallel to each other in a vertical plane, for instance as shown in *fig. 2*. The sections of the spiral will preferably be placed as close together as possible. In the first place the uniformity of brightness of the image of the filament is thereby improved or, speaking more accurately, its structure becomes finer,

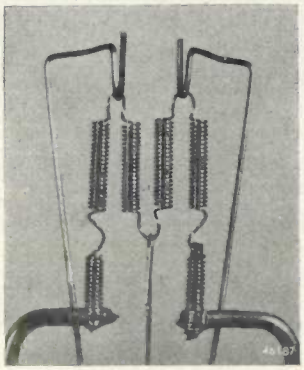


Fig. 2. Filament of a spiralized wire composed of four parallel sections of spiral.

so that a slight lack of definition in the image is sufficient to render the structure invisible. The image of the filament can then be focussed at a spot closer to the film window. In the second place of course the average brightness of the image of the filament becomes larger according as the area covered by the non-luminous intermediate spaces between the wires is smaller.

Filling the filament area

Projection lamps, like most electric lamps for lighting purposes, are filled with gas which helps to prevent the evaporation of the filament and allows the filament to burn at a higher temperature and attain the same lifetime. Due to the gas filling, however, it is impossible to decrease the distance between the parts of the spiral indefinitely because of the necessity of avoiding breakdown between adjacent parts.

In spite of this limitation various means are employed to arrive as far as possible at a complete filling of the prescribed area.

In the first place it is obvious that the lamp should be made for a low voltage. The voltage between the parts of the filament and thus the necessary distance between those parts then automatically become small. In addition there are other, more important arguments for the choice of a low voltage. At a given power the current is inversely proportional to the voltage. A larger current corresponds to a shorter and thicker wire. A thicker wire may, for a given lifetime, burn at a higher temperature, since it offers a relatively smaller surface for evaporation of the tungsten. The lamps

[3]) Clean bare tungsten has an emission coefficient of about 0,5. Due to the fact that, in the case of the spiral, part of the radiation is reflected one or more times back and forth between the windings before it reaches the outside, the radiation takes on more the character of that from a black body, and the emission coefficient correspondingly approaches more closely the value of unity.

for low voltage therefore have a greater brightness for the same lifetime than those for higher voltage. In order to obtain an impression of the difference, let us compare a 110 V, 100 W lamp and a 30 V, 100 W lamp, both for a lifetime of 100 hours. The first has a filament 65 μ thick, the temperature of incandescence is 3000 °K, the brightness obtained 350 stilb; the second has a filament 155 μ thick, temperature 3100 °K, brightness 950 stilb. Moreover, the shorter, thicker wire has the advantage that a shorter spiral is sufficient, thus fewer sections of spiral and consequently less intermediate space between them, while in addition the thicker wire, due to its greater strength, can be spiralized around a thicker mandrel, so that the width of the sections of the spiral is more advantageous compared with the spaces between. It is even possible here, using a thick mandrel, to fill up the whole filament area very satisfactorily with a single spiral without subdivision.

Apart from the heat losses through the leads, which are more important in the case of a short wire and high current (thick lead wires) and which prevent the voltage from ever being lower than about 15 V, a low voltage has only the disadvantage that for connection with the mains an intermediate apparatus is necessary, either a transformer, or, if the purchase price is more important than the efficiency of the projector, a resistance.

Since, because of this disadvantage, many manufacturers of projectors preferred to continue using high voltages, still other means of filling the filament area more uniformly are employed. A method which is used in all kinds of projection lamps (including those for low voltage) is the introduction of a spherical mirror behind the filament. This auxiliary mirror casts a real image of the filament in such a way that the images of the spiral sections fall exactly in the spaces between the actual spiral sections. The advantage in average brightness on the film window to be gained in this way may amount to 50 percent. The best uniformity will obviously be obtained when the intermediate spaces between the sections of spiral are not wider than the sections themselves, as is the case with the filament of *fig. 2*. If, however, the filament is one for a high voltage, the spiral must, on the one hand, be made thin in order for it to be strong enough with the thin weak wire which must be used, while on the other hand the spaces between the spiral sections must be large in order to avoid breakdown; compare such a filament for 220 V with one for 30 V (*fig. 3*). With such a rarefied filling of the area, as the figure shows, even the auxiliary mirror can only ensure a moderate uniformity.

A greater effect is obtained by doubly spiralizing the filament. This principle (sufficiently familiar from the incandescent lamps for ordinary lighting purpose ("Bi-Arlita")) improves the efficiency by decreasing the transfer of heat to the gas, permits the employment of a somewhat higher temperature for the same lifetime, due to a slight decrease in the evaporation, reinforces the already mentioned black-body effect and in our case also offers the possibility of filling the filament area about as well as if we were dealing with a low-voltage wire of the now spiralized single spiral. In this way it is possible, for example, to obtain the same average brightness with a 110 V-double spiral

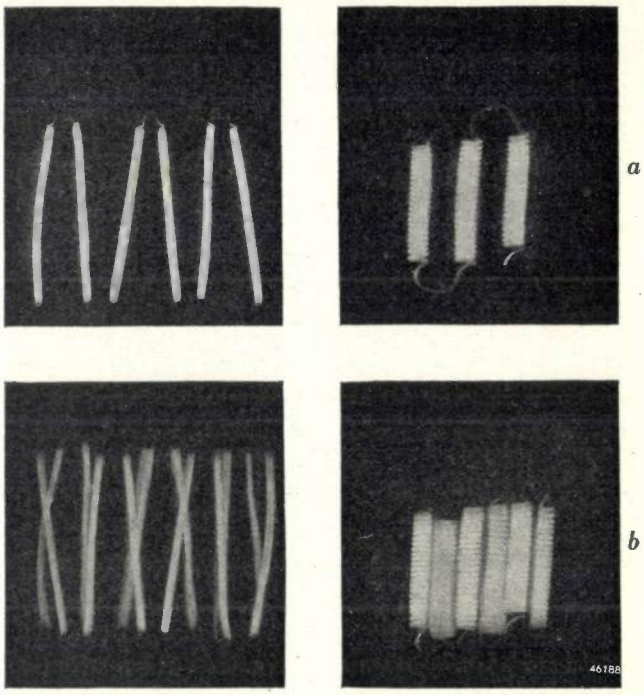


Fig. 3. Filaments of a 220 V narrow film lamp (left) and a 30 V narrow film lamp (right). *a*) The image of the filament without auxiliary mirror, *b*) with auxiliary mirror. In the case of the filament for the high voltage the filling of the area is unsatisfactory, even with the employment of the auxiliary mirror.

Table I.

Average brightness, without auxiliary mirror, of four different projection lamps for 100 W and a lifetime of 30 hours.

Voltage volts	Spiral	Dimensions of filament (mm²)	Average brightness (stilb)
30	single	5,2×4,5	1150
60	single	6,5×5	700
110	double	5,5×5	700
220	double	10 ×8	225

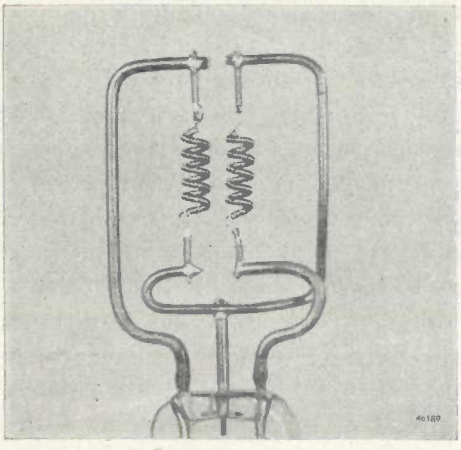


Fig. 4. Filament of doubly spiralized wire (110 V, 250 W). Note the connections of the two spiral sections: the current flows through both in the same direction, so that there is only half the lamp voltage between each corresponding pair of points in the two sections. Because of this the two sections may be situated closer together than when the two extremities at which the current is applied and taken off are situated side by side at the same height.

as with a 60 V single spiral; see *table I*. In *fig. 4* such a doubly spiralized filament is shown.

Finally the two-plane assembly is also much employed: the sections of spiral are not assembled in a single plane, but in two parallel planes in such a way that the spirals of one plane lie behind the intermediate spaces of the other, see *fig. 5*. It is clear that in this way very good uniformity and a very high average brightness can be obtained, especially when an auxiliary mirror is used in addition. As may be seen in *Table II*, these two-plane filaments are far superior to the others. In the case of one of the lamps here mentioned, by means of a combination of methods — two-plane assembly, low voltage and auxiliary mirror — a brightness of 4500 stilb has been obtained. This is certainly the highest value that has yet been attained with incandescent lamps. For the sake of comparison it may be mentioned that the spiral of an ordinary 100 Dlm 220 V lamp for the home has a brightness of 680 stilb, a fluorescent tubular lamp 0.3, the carbon arc 20 000, the high-intensity arc £0 000, the water-cooled super high-pressure mercury lamp (SP 1000) in the axis of the discharge 45 000 stilb.

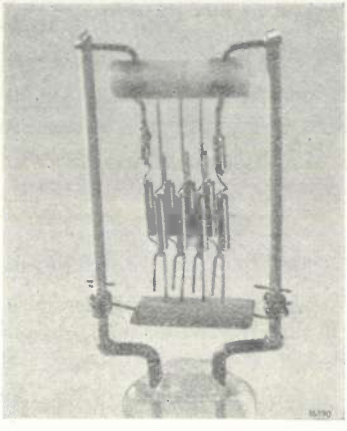


Fig. 5. Two-plane filament, seen obliquely from above. The system of hooks which protect the spiral sections on the upper side from lateral deviations is fastened rigidly to the leads, the corresponding hooks on the lower side, on the contrary, can move up and down in a vertical direction thanks to an arrangement of two tubes which slide over the leads (so-called sag arrangement, see below). The insulators should be noted between the hooks and the leads: these are made of a special ceramic material prepared and shaped in the factory and then fired with the hooks inserted.

Table II.

Data of several Philips cinema and narrow film lamps

Power watt	Voltage volts	Lifetime*) (without auxiliary mirror) hours	Form of filament mm²	Dimensions of filament	Total light fluxlumen	Average brightness*) (with auxiliary mirror) stilb
100	110	100	one plane	9 × 6,2	1 700	480
100	30	50	one plane	5,2× 4,5	2 300	1650
250	110	100	one plane	10 × 8	5 950	1200
250	50	50	one plane (double sp.)	7 × 6	6 000	2300
500	110	100	one plane	12 ×10,5	12 700	1650
500	110	25	two plane	7,8× 7,6	12 500	3500
750	15	100	one plane	8,5× 9,5	20 000	4100
750	110	25	two plane	9,5× 9	19 800	3800
900	30	100	one plane	11 ×12,5	25 000	3000
900	30	100	two plane	11,5× 8	25 000	4500
1200	110	10	two plane	12 ×12	36 500	4200

*) The life time without, the average brightness with auxiliary mirror.

The measures described all have the effect of promoting a greater concentration of the filament and a higher average brightness. Their effect is, however, partly cancelled by a more irregular distribution of temperature, of which they themselves are the cause. Since the temperature of the hottest part of the filament determines the lifetime, the more irregular the distribution of the temperature the lower the average temperature must be chosen for a given lifetime. Because of this, part of the gain in brightness must be sacrificed. This effect — the unfavourable ratio between highest and average temperature — is already appreciable upon passing over from single to double spiral; it is stronger in the case of the transition from single-plane to two-plane filament; and it is further accentuated when an auxiliary mirror is added because of the fact that part of the reflected radiation falls upon the filament again and raises its temperature locally. If the decreased lifetime is accepted, it would actually have been possible to make some gain in brightness without the mirror by raising the temperature of the filament. Independent of this controversy, however, remains the gain obtained in the projector due to the fact that with more uniform brightness the image of the filament can be focussed closer to the film window.

In speaking of the non-uniformity of the brightness of the filament we have had in mind until now the more or less periodic variations which are encountered as one passes across the spiral sections and the intermediate spaces of the filament or even along the length of a spiral over the windings. Superposed on this is a gradual decrease in brightness from the middle to the edges of the filament, caused by the fact that the edges cool more rapidly because of the unhindered lateral radiation, see *fig. 6*. Of itself such a not too rapid decrease toward the edges, which is reproduced in the illumination of the film picture and consequently also in the illumination on the screen, is not a serious difficulty. It necessitates, however, a certain correction in the statement made above about the required size of the filament: it is indeed an advantage to make the filament somewhat larger than the minimum size. When this is done only the middle, hottest part of the filament is used, whose average brightness may be considerably higher than the rest. And in practice this advantage is still further increased by the fact that at a given voltage a larger filament corresponds to a higher power and consequently a higher current, a thicker wire and thus a higher permissible temperature for the same lifetime.

Finally a few words about the gasfilling. The voltage at which, for a given distance between sections of the spiral, breakdown may happen depends upon the gas. For normal lamps a mixture of nitrogen and argon is used. The argon causes less loss of energy by the conduction of heat and therefore improves the light yield per watt, but it decreases the voltage for breakdown. As for film projection the specific light yield is less important than a maximum concentration, these lamps are filled with pure nitrogen, which allows of the sections of the spiral being placed closer together without risk of breakdown. It must, however, be kept in mind that even without that danger a certain distance would have to be maintained between the spiral sections in order to avoid any chance of short circuit between the sections due to possible deformations from mechanical vibrations, heat expansion or recrystallization. It should be mentioned here, that it has only become possible due to the progress in the manufacture of tungsten during the last ten years to diminish the deformations due to recrystallization upon ageing of the lamp to such an extent that it became possible to construct two-plane filaments. In the case of single-plane filaments also, it is of course desirable that the spiral sections shall remain straight during the life of the lamp. But the requirement is not so forcible in this case, since only upon deformation of the spirals in one direction is there chance of short circuit, while with two-plane filaments deformation in almost every direction may result in short circuit.

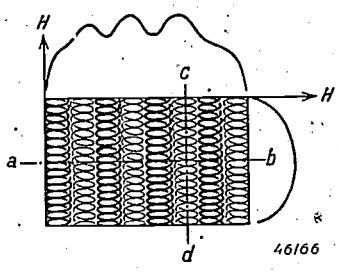


Fig. 6. Distribution of brightness in the filament image obtained with a two-plane filament with auxiliary mirror. Above, the distribution measured along the line *a-b*, to the right, that along *c-d*. The brightness is measured with a photocell of about half the width of the spiral image. It is plotted in arbitrary units. (See: G. Mili, J. Motion Pict. Eng. 28, 168, 1937).

Position of the filament

In order to satisfy the above-mentioned requirement that upon inserting the lamp in the projector the filament will automatically be situated at the correct spot, each lamp is provided with an inner socket and an outer socket, the latter of which can only be placed in the lamp holder of the projector in a single definite way (so-called centering socket). In a model of the given projector the lamp with inner socket is first placed in exactly the correct position with respect to the outer socket, which is checked by projection of the incandescent filament on a small screen with indicating marks, see *fig. 7*. The inner socket is then soldered into the outer one.

It is essential that the adjustment described should be carried out while the lamp is burning normally, because in the cold state the filament is in general in a different position than in the hot state, due to the fact that the leads and hooks and the filament itself expand when hot. A similar method is used for the locating of the auxiliary mirror on the wall of the envelope. The adjustment of the filament with respect to the mirror takes place during the fusing in of the pinch, simply by observing the mirror image of the filament. In the cold state filament and mirror image must be shifted mutually a certain distance in order to fit together satisfactorily when the filament is hot (see *fig. 10*).

It is of course necessary that all parts of the filament remain sufficiently firmly in place during the life of the lamp and are not deformed or

warped by possible tensions in the material. How such tensions may occur is shown in *fig. 8*. The spiral in two sections here shown as an example must be secured by a hook at the lower side to prevent lateral deviation. Due to the fact that the

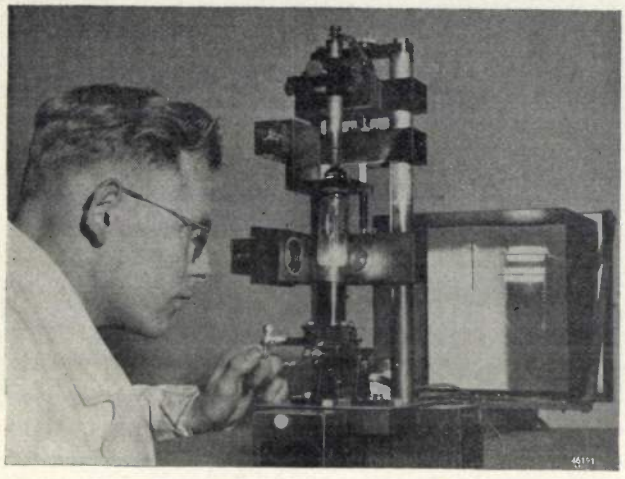


Fig. 7. The instrument for adjusting the filament with respect to the centering socket. To the right on the screen may be seen the filament images, one of which is projected directly on the screen and the other, seen perpendicular to it, is projected *via* a prism. By turning and shifting the lamp the images must be made to coincide with a drawing on the screen. The correct position of this drawing with respect to the lamp holder is checked from time to time with a normal lamp.

thin filament becomes hot and cold much more rapidly than the thick leads, upon switching on the spiral has a slight play for several seconds, upon switching off on the other hand the spiral (and the hook) would be under a high tension for several moments. In order to prevent this in the case of the more complicated filaments (two-plane systems) a so-called sag arrangement is employed; *cf. fig. 5*. The hooks are not rigidly fixed, but are able to move slightly up and down vertically, while still giving the spiral sections sufficient support against lateral deviations.

Thanks to these measures and to the previously

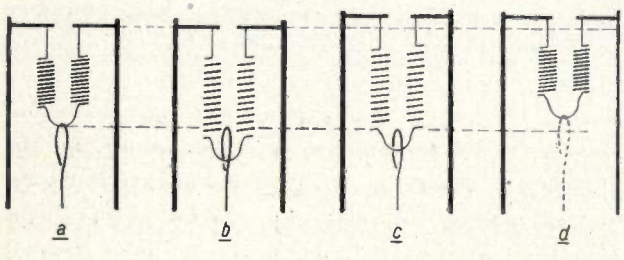


Fig. 8. The expansion and contraction of a two-section filament and the leads upon switching on and off. *a*) Cold state; *b*) immediately after switching on: spiral hot, leads still cold; *c*) in operation; *d*) directly after switching off: spiral cold, leads still hot.

mentioned absence of deformation in the spirals due to recrystallization, all parts of the filament retain their correct positions during the whole lifetime, and no decrease in light occurs in the projector due to gradual loss of adjustment.

The envelope

The requirement mentioned at the beginning, that it must be possible to place the filament as close as possible to the condenser, leads to the fact that the envelope, usually of a tubular form, must be given as small a diameter as possible. A limit is set by the fact that the wall of the envelope would become too hot if it were too close to the incandescent filament. This can also be enhanced by the blackening which normally appears after the course of time on the wall due to the evaporated tungsten, which is deposited on the wall and which absorbs part of the radiation and thus makes the envelope still hotter. Three methods are available of arriving at a small diameter of the envelope. In the first place the envelope is made of a special hard glass which has a very high softening point, namely 750 °K. In the second place, especially in projectors for 8 and 16 mm film, a forced cooling is employed by means of a fan. In the third place attempts are made to prevent the deposition of the evaporated tungsten on the wall in the neighbourhood of the filament. This is indeed desirable in any case because the blackening is accompanied by a gradually increasing reduction of the light during the lifetime of the lamp.

The gas with which the lamp is filled plays an important part in these phenomena. While on the one hand it has the disadvantage that it becomes warm itself by heat conduction from the filament and therefore heats the envelope more strongly, on the other hand it has a favourable effect as far as the blackening is concerned, because of the fact that currents occur in the heated gas which carry the evaporated tungsten along, so that most of it is not deposited close to the filament but at some other spot. This effect can be reinforced by introducing nickel gauzes or plates above the filament in the rising gas current, upon which the tungsten particles are deposited (see *fig. 9*). In this way it has been possible to limit the diameter of the envelope of lamps for 8 and 16 mm film for 1000 W and 1200 W to the same as that for a 750 W lamp, namely only 38 mm. Still more effective is the employment of the so-called hanging construction; *fig. 10*. The lead wires, in this case above the filament, which become very hot close to the filament, function as a sort of chimney, which causes

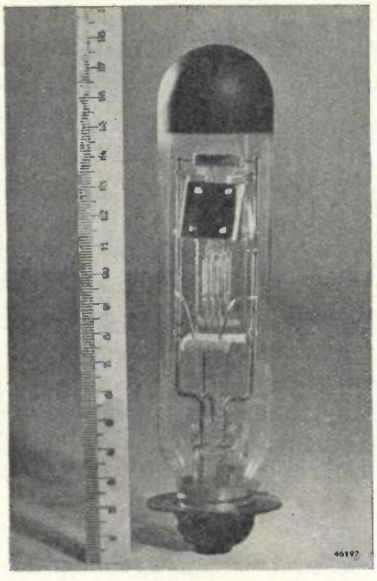


Fig. 9. Philips 16 mm film lamp, 110 V, 1000 W. Above the filament are two nickel plates along which the rising gas flows and upon which the particles of evaporated tungsten have an opportunity to be deposited. The upper part of the envelope is sprayed black in order to prevent light escaping through the ventilation holes of the projector into the auditorium.

the rising current of gas to pass first along the axis of the lamp. In the spherical enlargement at the top of the envelope the gas takes on a whirling motion in which practically all of the tungsten carried with it has the opportunity of being deposited on the wall at that spot. The descending, cooled gas, in which there are only few tungsten particles left, deviates from the wall again slightly

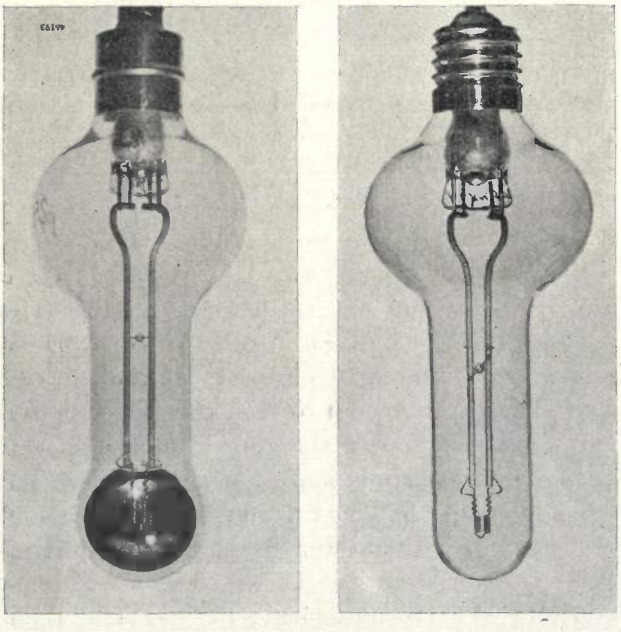


Fig. 10. Two Philips cinema lamps 15 V, 750 W with the so-called hanging construction. The left-hand lamp has an auxiliary mirror of evaporated aluminium on the inside of the envelope.

above the filament and flows toward the filament (*fig. 11*). Therefore at the height of the filament the wall remains almost entirely free of blackening, while also there is no heating of the wall here by passing hot gases. In that way in the case of the 750 W lamp shown in fig. 10, which had to operate without artificial cooling, the diameter could be reduced to almost the same value (44 mm) as in the case of the above mentioned lamps with forced cooling.

The gas currents described can be made visible in a very interesting way by constructing such a lamp with a tube fused on, through which, while the lamp is burning, a small amount of air can be admitted to the lamp. The tungsten particles coming from the filament are then immediately

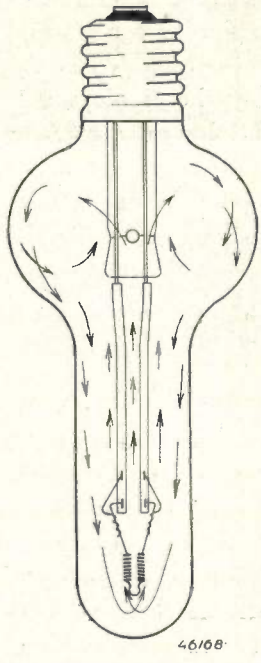


Fig. 11. Gas currents in the lamps of fig. 10.

oxidized to a white powder whose motion can be followed with the eye. The well-known method of Toepler can also be followed, in which the current phenomena are made visible by means of the accompanying very slight local variations in the index of refraction. These methods are very useful in the investigation of the shape of envelope which is most suitable for obtaining the desired course of the currents of gas.

In still another way the gas helps in protecting the envelope. If, due to an incipient blackening or possible insufficient cooling, the envelope should become soft at a certain spot, then at that spot it would be slightly compressed by the external pressure if there was a vacuum or diminished pressure inside the lamp. It would thus come nearer the filament and thus become still hotter and still softer, and this accelerated process would quickly

lead to the destruction of the lamp. Now by filling the lamp with a gas at such a pressure that when hot there is a slight excess pressure in the lamp, the compression of the envelope is made impossible. The excess pressure can be employed without any danger because of the small size and thus the strength of the envelope. The relatively high pressure gives the additional advantage of a still greater increase in the permissible working temperature and an increase in the breakdown voltage in the gas.

Finally it may be mentioned that for some 8 or 16 mm film projectors, with a very compact construction, lamps have been made in which the filament is brought especially close to the wall by excentric placing in the envelope. In the case of the 400 W lamp shown in *fig. 12*, for example, the distance between filament and wall is only 7 mm. This is made possible by the employment of the hanging construction and by an arrangement for cooling in the projector housing in such a way that the part of the envelope lying opposite the filament is the most strongly cooled. For the rest it is striking that in this lamp the spiral sections are horizontal. The vertical position is usually chosen in order to prevent short circuit between the sections in case of any slight sagging of the spiral during use. The horizontal position has the advantage that the light emitted in the axes of the envelope can easily be used for the scanning of the sound track of the film, so that a separate lamp is not needed for that purpose. In the case of the lamp in fig. 12 use is made of this possibility, which exists indeed only with the hanging construction, since in the ordinary, standing construction the light along the axis direction would have to pass through the upper parts of the envelope, which are most subject to blackening.

Although in projectors for normal film the dimensions of the apparatus are much less critical than in the case of narrower film, here also for other reasons the diameter of the envelope must be kept as small as possible. Since in general these projectors work with a mirror condenser, the auxiliary mirror is placed in front of the lamp and thus cuts off a part of the beam of light which the condenser mirror throws on the film. The auxiliary mirror should therefore be small, and since the mirror is placed on or outside the envelope, the envelope itself must be small.

Attempts have indeed been made to avoid this necessity by placing a very small mirror inside the envelope very close to the filament. This mirror had to be made of molybdenum because of the intense heat. The relatively low coefficient of reflection of molybdenum (60 percent in the cold state, about 75 percent at white heat, compared with 95 percent for a mirror of aluminium deposited on the wall) and the practically unavoidable loss of adjustment upon becoming hot, however, make the advantage of this method more or less illusory, especially since the hanging construction has made it possible to reduce the diameter of the envelope very much at the position of the auxiliary mirror (in the lamp of fig. 10, left, it amounts to only 60 mm.)

The lamps in operation

We have already pointed out that the lamp must be resistant to the mechanical vibrations to which it is exposed when the projector is in use. These vibrations are composed of many components of different frequencies and are different for each type of projector. It is therefore impossible to indicate a lamp construction which could be used in all cases. Attempts are rather made empirically during the development of the lamps to render a lamp insensitive to the special vibrations of the type of projector for which it is intended. For this purpose the lifetime is determined of test models of the lamp and a normally working model of the projector, and the construction of the leads and hooks are altered until the desired lifetime in operation is fully achieved. This work is made more difficult by the requirement that the leads and hooks must not stand in the way of the light beams which the filament emits towards the condenser and auxiliary mirror.

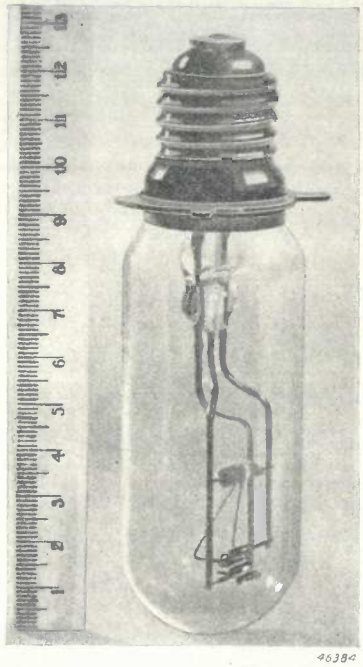


Fig. 12. Philips 16 mm film lamp 30 V, 400 W, with hanging construction and with excentrically placed filament, whereby the distance between filament and wall of the envelope is reduced to 7 mm. The spiral sections are horizontal in this case, the light emitted downward is used for scanning the sound track of the film. The sections of the spiral do not lie in a flat plane but in a cylindrically bent plane, whereby a similar effect is obtained as in the two-plane assembly.

The nominal lifetime can only be attained if the lamp is burned precisely at the prescribed voltage or current. If the lamp is calculated for a certain voltage, an increase in voltage of 1 percent decreases the lifetime by 14 percent, while on the other hand a decrease in voltage of 1 percent causes a decrease in brightness of 4 percent. With a lamp for a certain current the corresponding variations upon change in current are still larger, namely 25 percent in lifetime and 8 percent in brightness. The difference between the two kinds of lamps consists in the fact that, in choosing the dimensions of the former, account is taken of a gradually decreasing current strength during the life of the lamp, due to the fact that the filament is becoming thinner by evaporation, and a consequently increasing resistance; in the case of the second lamp a gradually increasing voltage is discounted in the dimensions for the same reason. Since as a result of this, the brightness will decrease slightly in the first lamp and increase in the second lamp, the use of lamps for constant current is always considered desirable for projectors in order to compensate a decrease in light due to blackening of the walls and deformation of the filament. Since, however, according to the above in the Philips lamps this decrease in light yield has been reduced to practically zero, the burning of the lamp at constant voltage would be more recommendable in connection with the above-mentioned slighter sensitivity to variations in the operating condition, and also because the "gain" in light upon the ageing of the lamps for constant current is of course only obtained at the expense of a lower temperature of incandescence and brightness in the new lamp. In any case the voltage or current stamped on each lamp should be exactly maintained during its use if full profit is to be had with respect to brightness and lifetime. Most projectors are provided with a measuring instrument (although it is unfortunately not always reliable) for controlling voltage or current.

A few words may be said about the efficiency of the lamps. As can be seen from *table II*, it is very high, thanks to the high temperature of incandescence. The lamps with a lifetime of 25 hours have as a rule a specific light yield of 25 lm/W, the 1200 W lamp in the table with a 10 hour lifetime even has more than 30 lm/W, which begins to approach the efficiency obtained with gas-discharge lamps. If one considers the unusually uneconomical way in which the light is used in the projector, the high efficiency mentioned appears to be a waste of effort. A balance of the light flux excited in the lamp for different projectors shows that in the end a maximum of only 1 to 2 percent of it reaches the screen as effective light. Nevertheless, the high light yield of the lamp is of very great practical advantage, not so much because of the lower current costs — these are negligible compared with the cost of the film — but because of the smaller amount of heat developed. This byproduct must be removed from the projector without the occurrence of undesired high temperatures at any spot, and in particular the heat contribution from the radiation which must be concentrated on the film window may not cause at that spot too great overheating. This heating is indeed not so dangerous in the case of film of non-inflammable material, and it is less intense in the case of the lens condenser used for narrow film than with a mirror condenser, since due to the dispersion in the glass the lens has the fortunate property of focussing the infra red rays to an image of the filament considerably farther away. On the other hand, however, in the case of just these narrow film apparatus every limitation of the heat developed is welcome in order to be able to reach a reasonable light production and still retain a small and convenient construction.

THE MEASUREMENT OF REVERBERATION

by W. TAK.

534.844.1

In this article several theoretical considerations about reverberation in rooms and auditoria are discussed. The phenomenon of reverberation is considered as the sum of a number of damped characteristic vibrations. Due to interference, fluctuations are superposed upon the mainly exponential variation. The variation in intensity depends upon the shape and dimensions of the room, upon the distribution and the properties of the absorption material, upon the position of the observer, upon the position of the source of sound and upon the frequency spectrum of the sound produced, but when these factors remain constant it is completely reproducible.

Several principles are indicated according to which apparatus can be constructed which makes it possible to study the variation in intensity during the reverberation and to characterize it by one or more parameters. A more detailed description of the apparatus will be given in a subsequent article.

It is known that in a room where sound is produced the intensity does not immediately decrease to zero after the source of sound has ceased to act. Depending on the circumstances, a certain time is necessary before the sound dies out. This phenomenon of reverberation has great influence, not only on the intensity of the sound, but also on the intelligibility of speech and the quality of music, and is therefore important in judging the acoustics of the room. It is not only a question of the length of time during which the reverberation lasts, but just as much of the variation of the intensity as a function of time, *i.e.* of the character of the reverberation. This character is not, however, a property of the room in question only, but it is also determined by the nature and the position of the source of sound and the position of the observer. The influence of the duration of the reverberation on the intelligibility of speech and the analysis of music have been discussed previously[1]). We shall now go more deeply into the causes which determine the character of reverberation and indicate several principles by which it can be studied experimentally.

Characteristic vibrations of a closed space

The study of the vibrations of the air in closed spaces has shown that such a space possesses a series of characteristic frequencies each of which corresponds to a definite mode of vibration of that space.

We shall begin with the consideration of a simple case. We imagine a source of sound situated between two opposite plane walls of a rectangular room. The source emits a plane sound wave toward the front and rear with a wave front parallel to the walls mentioned above. A system of "stationary" waves will then occur between these walls, which is most strongly developed when the source emits a single frequency such that the distance between the walls is a multiple of a half wave length. These frequencies, which are characterized by the formulae

$$a = p \cdot \frac{\lambda}{2}, \quad \lambda = \frac{2}{p/a}, \quad \nu = \frac{c}{\lambda} = \frac{c}{2} \cdot \frac{p}{a}, \quad (1)$$
$$p = 1, 2, 3 \ldots,$$

where a is the distance between the walls in question, λ the wave length, ν the frequency and c the velocity of sound, form a series of "characteristic frequencies".

In *fig. 1* the modes of vibration of the air (amplitude of the air velocity and the sound pressure as a function of the position) are represented for $p = 1, 2, 3, 4$. The walls are here assumed to be hard, so that they reflect the sound almost completely. In fig. 1 the spectrum belonging to this series of characteristic vibrations is also given. Actually

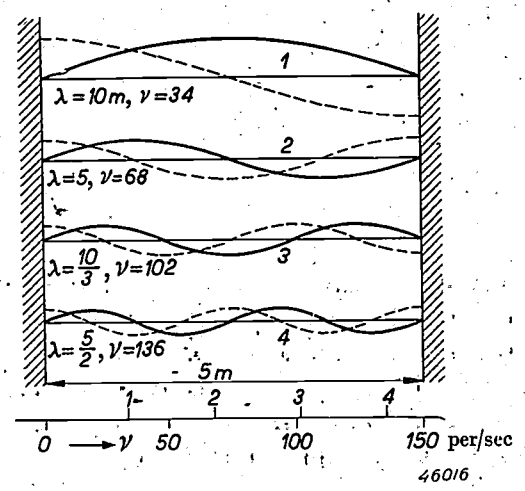


Fig. 1. The first four characteristic vibrations of a column of air between two parallel walls. Beneath the figure the spectrum of the corresponding characteristic frequencies is given.

1) Philips techn. Rev. 3, 65, 1938.
Philips techn. Rev. 3, 143, 1938.
Philips techn. Rev. 3, 368, 1938.

the frequency spectrum in a rectangular room is more complicated; in the first place because there are three pairs of opposite walls, each with its own distance apart, and further because plane waves also occur which are not propagated parallel to one of the walls.

The general formula which must then be used instead of (1) is the following:

$$\nu = \frac{c}{2}\sqrt{\frac{p^2}{a^2}+\frac{q^2}{b^2}+\frac{r^2}{h^2}}, \quad \lambda = \frac{2}{\sqrt{\frac{p^2}{a^2}+\frac{q^2}{b^2}+\frac{r^2}{h^2}}}, \quad (2)$$

$$p, q, r = 1, 2, 3, 4, \ldots,$$

where a, b and h represent the length, width and height of the room and p, q and r are arbitrary numbers.

can simply be set equal to the volume of the spherical shell, multiplied by the average density of the frequency points. The following is then found for the number of characteristic vibrations in the interval ν, $\nu + \Delta\nu$:

$$\tfrac{1}{2}\pi\nu^2\Delta\nu \cdot \frac{8abh}{c^3} = 4\pi\frac{abh}{c^3}\cdot\nu^2\Delta\nu.$$

For the above-mentioned room, in the neighbourhood of $\nu = 1000$ and when $\Delta\nu = 1$ c/s, the number of characteristic vibrations is already about 25.

What happens when the source of sound is now switched off? While previously at every point in the room there was a vibration with the frequency (or frequencies) of the source, we now observe a motion which can be conceived as a superposition of a number of characteristic vibrations, whereby

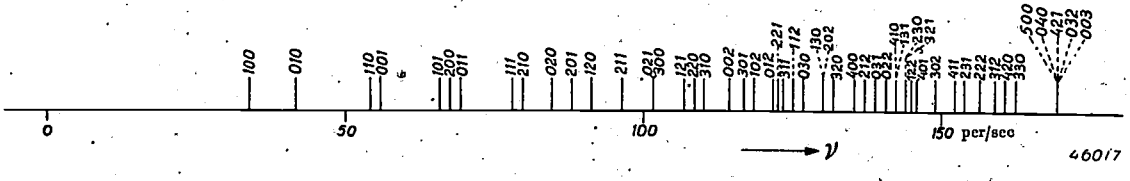


Fig. 2. Part of the frequency spectrum of a rectangular room with the dimensions $a = 5$m, $b = 4$m, $h = 3$m.

In *fig. 2* the frequency spectrum is given, as an example, for a room for which $a = 5$m, $b = 4$m and $h = 3$m. It may be seen from this figure that the number of vibrations in a given frequency interval increases rapidly with increasing frequency. This can also be proved as follows. If in a three-dimensional system of coordinates one constructs the points for which

$$x = \frac{cp}{2a}, \quad y = \frac{cq}{2b}, \quad z = \frac{cr}{2h},$$

those points form a regular lattice in space with the average density:

$$\frac{8abh}{c^3}.$$

Furthermore:

$$\nu = \sqrt{x^2+y^2+z^2},$$

so that the distance from the point to the origin represents the frequency. The number of frequencies in the interval ν, $\nu + \Delta\nu$ is equal to the number of points lying in an octant of a spherical shell included between the spheres with radii ν and $\nu + \Delta\nu$, whose volume amounts to

$$\tfrac{1}{2}\pi\nu^2\Delta\nu.$$

With a sufficiently large value of ν this number

especially those characteristic vibrations contribute to the sound image whose frequency differs little from the frequencies present in the source. Due to the absorption always present, these vibrations will be damped, so that the reverberation may therefore be conceived as *the sum of a number of damped characteristic vibrations.*

Intensity of the reverberation as a function of the time

Let us assume, by way of hypothesis, that in the closed space considered, only one of the characteristic vibrations, with the frequency ν_e, is active. At every point in the space pressure and velocity are sinusoidal functions of time. The sound energy, consisting of a potential part which is proportional to the square of the pressure, and a kinetic part proportional to the square of the velocity, is a periodic function of the time, which also depends upon the position of the observer. If there is no damping, the energy per unit volume ($\bar{E}$), averaged over the space, is constant. If there is slight damping the total energy $\bar{E}\cdot V$ decreases with the time according to the formula: [2])

$$\frac{dE}{dt} = \frac{d}{dt}(\bar{E}\,V) = -\frac{c\,\bar{E}\,A}{4},$$

[2]) For the deduction of this formula see Philips techn. Rev. 3, 65, 1938.

so that:

$$\bar{E} = \bar{E}_0\, e^{-kt}, \text{ where } k = \frac{cA}{4V}.$$

V is here the total volume, $\bar{E}_0$ the average energy per unit of volume at the moment $t = 0$, c the velocity of sound, $A = \int a\, \mathrm{d}O$, where $\mathrm{d}O$ represents an element of the wall and a the absorption at that sp t. The sound energy thus decreases according to an exponential law. In the presence of damping the sound *pressure* at every point in the room will be an exponential function of the time, so that:

$$p = p_0\, e^{-k't} \sin 2\pi\nu_e t, \text{ where } k' = \frac{k}{2} = \frac{cA}{8V}.$$

This is represented in *fig. 3a*. The absolute value of the sound pressure may also be considered and averaged over a small number of periods. The average sound pressure $\bar{p}$ thus obtained also varies according to an exponential function of the time:

$$\bar{p} = \bar{p}_0\, e^{-k't}.$$

In fig. 3*b* and *c*, $\bar{p}$ and $\ln \bar{p}$ are represented as functions of t.

If there is more than one characteristic vibration active, each characteristic vibration individually will die out exponentially. Since, however, the frequencies differ, these vibrations will cause beats by interference, so that the pressures will exhibit fluctuations. When a single tone is emitted in a room by means of a loudspeaker, upon interruption of the sound the amplitudes of the excited characteristic vibrations will be smaller the larger the differences between the frequencies, and thus only those characteristic frequencies will be noticeably active whose frequencies differ only slightly from the given frequency. In a room in which the absorbent materials are distributed uniformly over the surface, the damping constants of these characteristic vibrations can in general be set equal to each other, because while the absorption coefficient depends upon the frequency, in a narrow frequency region it varies only slightly. The interference picture will now be quite simple and the amplitude modulation which occurs as a result of the interference will be clearly observable (*fig. 4a* and *b*). If the tone emitted by the loudspeaker is replaced by a more composite sound, the interference picture will exhibit a different character because a larger number of characteristic vibrations are then excited whose beats partially compensate each other, whereby the fluctuations, while more numerous, are less pronounced (*fig. 5a* and *b*). The damping constants can now vary more widely; this will in general have little effect on the variation in the intensity as long as the sound emitted does not cover a very extensive part of the spectrum (for instance several octaves). If the latter is, however, the case, and the wall absorbs for example

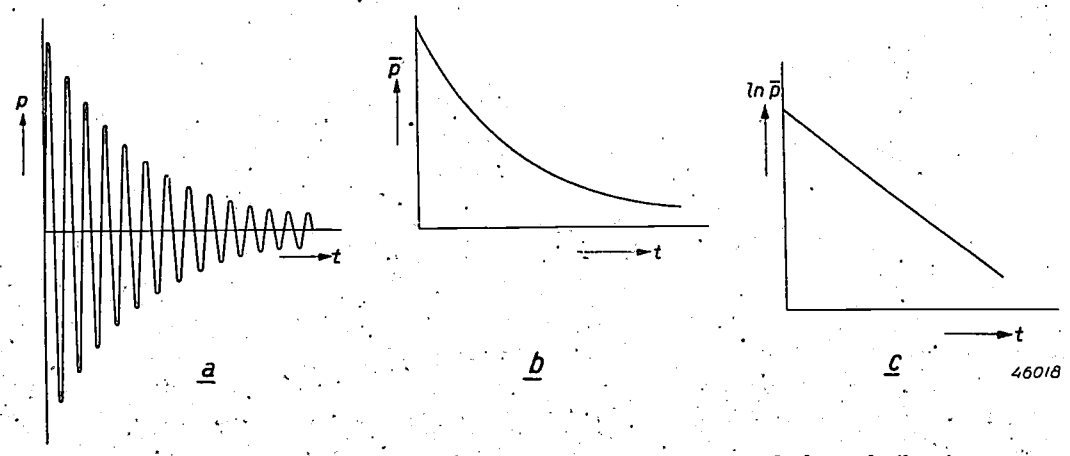


Fig. 3. Plotted as a function of t: *a*) the sound pressure p for a single damped vibration, *b*) the average sound pressure $\bar{p}$, *c*) $\ln \bar{p}$.

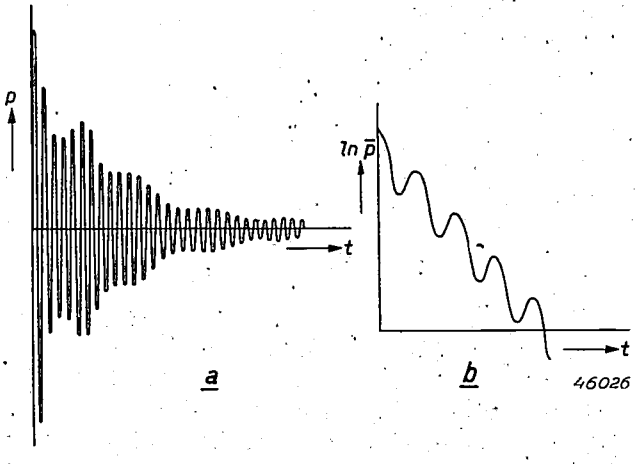


Fig. 4. Plotted as a function of t: *a*) the sound pressure p upon superposition of two damped characteristic vibrations with different frequencies, but with the same damping, *b*) $\ln \bar{p}$.

only the low frequencies, then due to the rapid dying out of the low excited frequencies the room will have a shrill sound. If on the other hand absorption in the high frequencies is predominant, the room has a hollow sound.

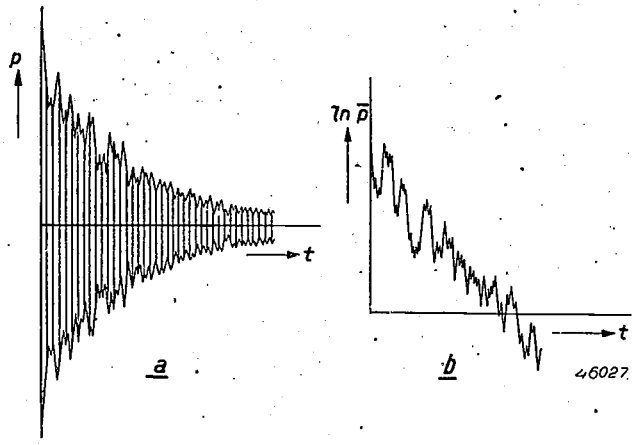


Fig. 5. Plotted as a function of t: a) the sound pressure p upon superposition of a large number of characteristic vibrations, b) $\ln \bar{p}$.

The situation is different when the absorbent materials are distributed very irregularly over the surface, for instance in the case of a room with hard walls, whose floor is covered with a soft carpet. Here the vibrations whose waves planes are parallel to the walls will be much less damped than those with horizontal wave surfaces. In such a room characteristic frequencies with very different damping will occur at the same time. In the variation of the pressure with the time, apart from the fluctuations, a superposition of different exponential functions can be recognized, because the strongly damped characteristic vibrations first die out, and only later the weakly damped ones. A similar situation is encountered when a strongly absorbent room is connected by an opening with a room with slight damping, for example a "soft" room which is connected by an open door with a "hard" corridor. In that case it is advisable to consider the characteristic vibrations of each room separately. The characteristic frequencies will be slightly affected by the opening, which acts as a "coupling". The characteristic vibrations will only be altered in as much as those of the hard room penetrate slightly into the soft room, and *vice versa*. In this case also in the combined room, characteristic vibrations with practically the same frequency but with very different damping will exist at one and the same time. If the source of sound is situated in the room, the corridor will resonate faintly. An observer in the room hears chiefly the strongly damped characteristic frequencies of the room, followed by a faint, more lasting reverberation of the corridor (*fig. 6a*). An observer in the hard room, however, will notice chiefly the slowly dying sound in that room, while the quickly damped sound of the soft room, which initially reached his ear, will quickly escape his attention (fig. 6b).

Reverberation time

We have seen that in the simplest case the average pressure varies as an exponential function of the time:

$$\bar{p} = \bar{p}_0 e^{-k't}.$$

This variation can now obviously be characterized in a simple manner. If we now define the reverberation time as the time t_{60} during which the intensity decreases by a factor 10^6 (60 db) (and thus the pressure by a factor 10^3), we find for this time t_{60}:

$$10^{-3} = e^{-k't_{60}},$$

$$t_{60} = 2.3 \frac{3}{k'} = \frac{13.8}{k},$$

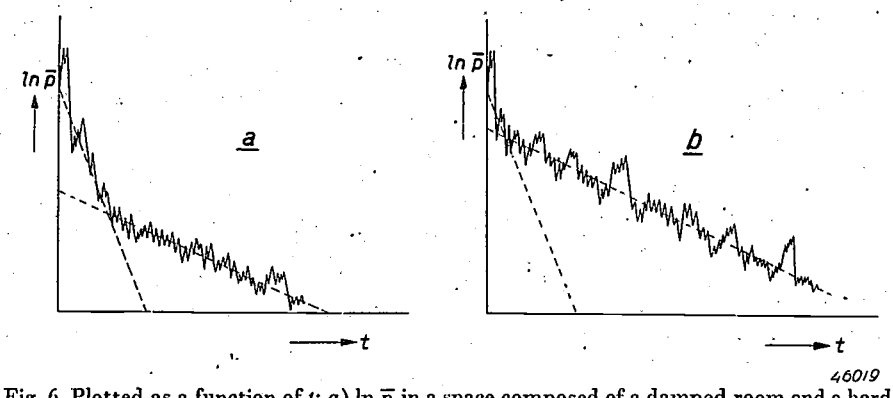


Fig. 6. Plotted as a function of t: a) $\ln \bar{p}$ in a space composed of a damped room and a hard corridor for a point of observation in the room, b) the same for a point of observation in the corridor.

or, since $k = cA/4V$,

$$t_{60} = 0.16 \frac{V}{A},$$

where V is expressed in m^3 and A in m^2 of open window. In the case of an *exponential* drop the t_{60} measured is a parameter which completely determines the character of the reverberation. We shall therefore call it the reverberation time of the exponential reverberation.

If one has a diagram in which the variation of $ln\,\bar{p}$ (or $2\,ln\,\bar{p}$) with the time is given, the reverberation time can be determined therein by a simple construction. In *fig. 7* this construction has been performed.

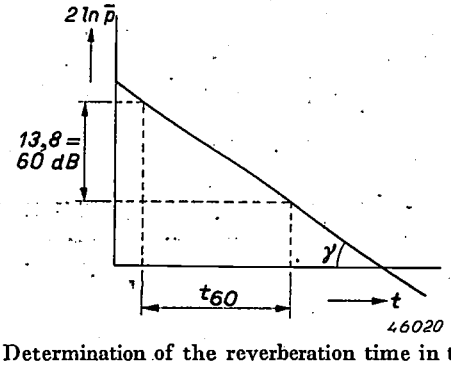


Fig. 7. Determination of the reverberation time in the case of an exponential drop of the intensity. The situation of the right line has been rendered by the measurements. It follows from the graph that $\tan t_{60} = 13.8 \cot \gamma$.

In more complicated cases, for example when strong fluctuations are superposed on the exponential variation, there is of course no simple relation between the reverberation time and the character of the curve. In that case it is obvious that an average exponential curve can be drawn, disregarding the fluctuatons, and a value of t_{60} for that curve can be found. We shall call this quantity the average reverberation time. It is true that the phenomenon is not thereby completely characterized, because the fluctuations have been disregarded, but it gives, nevertheless, the general character of the reverberation.

If the reverberation consists of the superposition of different exponential functions, each with its own constant k', perhaps combined with fluctuations, the obvious method is first to replace the curve by an average curve without fluctuations and to analyze that curve into its purely exponential components. If one then determines t_{60} for the separate components, one obtains a number of reverberation times which together to some extent fix the general character of the reverberation.

Quantitative investigation of reverberation

There are various experimental methods of investigating reverberation. The oldest is that of Sabine, by which only the reverberation time is determined. For that purpose the time was measured which elapsed between the moment at which the source of sound ceased to function (this might only take place when a stationary sound intensity had been reached) and the moment at which the sound became inaudible. This method was later improved by the use of an entirely automatically working apparatus which determines electrically the time in which the intensity of the signal decreases by 60 db [3]).

We have seen, however, that the reverberation depends very much upon different factors and that it may have a very capricious variation, so that the sole determination of the reverberation time cannot in this way give a picture of the true variation of intensity. At the present time use is made of recording meters, which have the advantage that the exact behaviour of the reverberation is recorded, making it possible to draw an average curve and to determine from it one or more average reverberation times, while, moreover, from the curve obtained conclusions may be drawn about the acoustic properties of the room being studied [4]). The principle of these recording meters is simple. After being amplified the voltage of the microphone is connected to the poles of a potentiometer, which is then automatically regulated so that the voltage taken off remains constant. The mechanical regulator at the same time drives a style which records the position of the potentiometer contact on a roll of paper. The necessary high velocity of recording, in connection with the unavoidable mechanical inertia, makes the whole mechanical part of this apparatus extremely delicate and difficult to transport. We have therefore developed two methods which are not subject to this objection and we shall discuss them in the following.

The first method we have called the method of "exponential amplification", the second that of the "exponential time base".

Method of exponential amplification

For the sake of simplicity we shall consider the case where the sound pressure varies purely exponentially, so that:

$$p = p_0\, e^{-k't} \sin 2\pi\nu t.$$

When we observe this quantity by means of a microphone, an amplifier and an electron-ray oscillograph, and arrange the apparatus in such a way that after

[3]) M. J. O. Strutt. De Ingenieur 4 E, 20, 1932.
[4]) *C.f.* Philips techn. Rev. 3, 65, 1938.

the interruption of the source of sound the amplification increases according to an exponential function of the time $e^{+\alpha t}$, then, when α is chosen equal to k', the influence of the damping will just be cancelled by that of the amplification. The result is a simple sine curve with a constant amplitude, while for $\alpha > k'$ and $\alpha < k'$ the amplitude respectively increases and decreases with the time

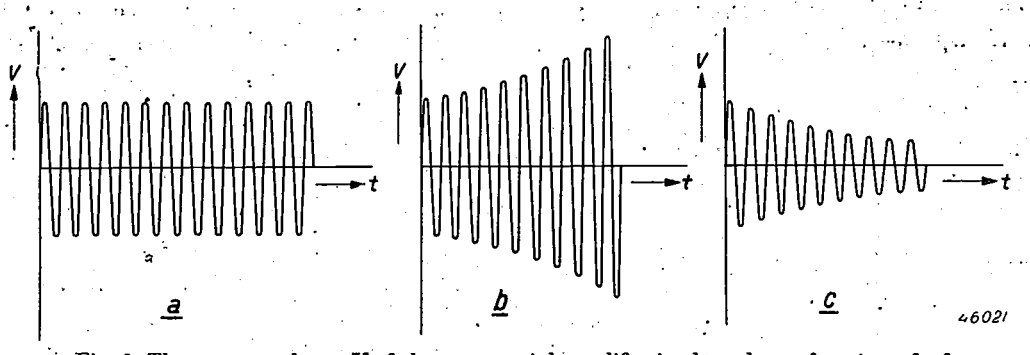


Fig. 8. The output voltage V of the exponential amplifier is plotted as a function of t for the case of a single damped characteristic vibration, when $a)\ \alpha = k'$, $b)\ \alpha > k'$, $c)\ \alpha < k'$.

(*fig. 8a, b* and *c*). It is clear that we can also apply this to those cases where fluctuations are superposed on the general exponential behaviour.

The fluctuations are, to be sure, not then eliminated from the result, but with a suitable choice of α the average amplitude can be kept constant (*fig. 9*). The whole can be made visible by connecting the output voltage of the "exponential amplifier" to the vertical diffraction plates of an electron-ray oscillograph. By synchronisation of the time base with the apparatus which regulates the switching on and off of the source of sound, it is possible repeatedly to project the image of the reverberation on the screen. Since the variation of the intensity in a given case, with the same arrangement of microphone and source of sound, is entirely reproducible, a constant image appears on the screen. By varying α, that arrangement can be found with which the image on the average has a constant amplitude. In that case $\alpha = k'$, and from that the average reverberation time can be deduced, while at the same time the character of the reverberation can be studied in the image observed. In some cases (for example the above-mentioned combination of soft room + hard corridor) the reverberation consists of several terms with different values of k'. From the oscillogram it may then be seen that at several values of α the amplitude can be made constant for a part of the image (*fig. 10a* and *b*), from which the different

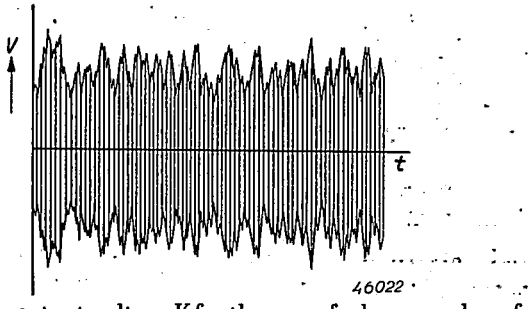


Fig. 9. The output voltage V for the case of a large number of damped characteristic vibrations.

reverberation times present simultaneously can be deduced. Here also the method described shows clearly the general character of the reverberation together with the details.

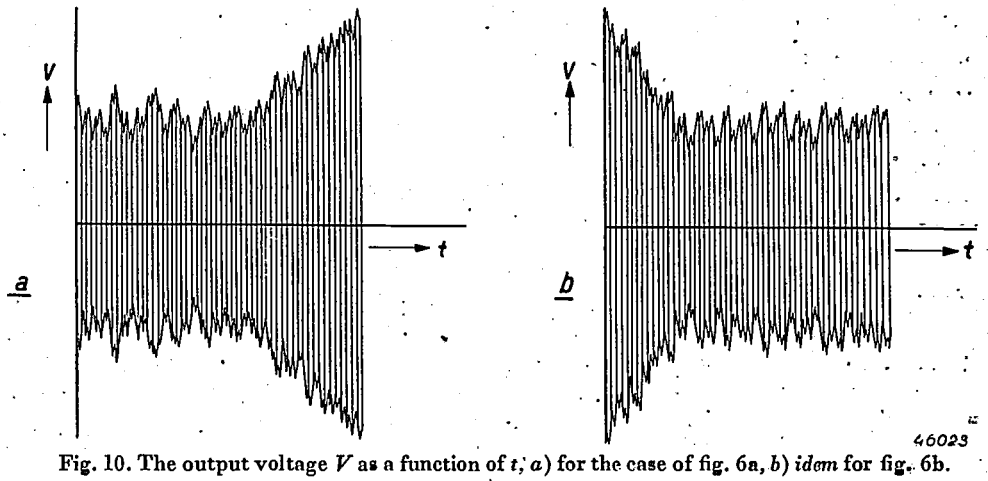


Fig. 10. The output voltage V as a function of t, $a)$ for the case of fig. 6a, $b)$ *idem* for fig. 6b.

When the intensity of the reverberation decreases, there comes a moment at which the deviation on the screen becomes just as large as the deviations which are the result of unavoidable extra noises and of the noise of the amplifier. The intensity will then no longer decrease and at the end the image on the screen will always show an increase in the amplitude. In order to avoid false conclusions the apparatus is so adjusted that upon reaching the noise level the curve is interrupted and the signal switched on again.

The method of the exponential time base

In this method the AC voltage from the microphone is connected *via* a rectifier with a smoothing arrangement to the plates for vertical diffraction of an electron-ray oscillograph. To the plates for horizontal diffraction a voltage is applied which does not increase linearly with time, as is customary (linear time base), but which varies according to an exponential function $e^{-\beta t}$. The movement of the light spot on the screen is then given by the functions

$$y = a\, e^{-k't},$$
$$x = b\, e^{-\beta t}.$$

If we eliminate t from these, we obtain:

$$\left(\frac{y}{a}\right)^{\beta} = \left(\frac{x}{b}\right)^{k'},$$

$$y = C\, x^n, \qquad C = a\, b^{-n}, \qquad n = k'/\beta.$$

The curves represented by this equation (*fig. 11*) are in general sorts of parabolas. There is, however,

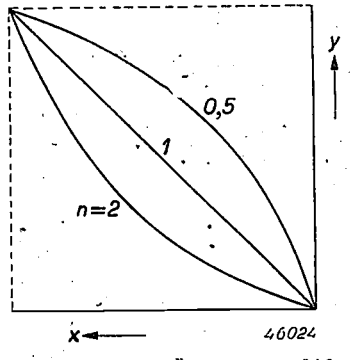


Fig. 11. The curves $y = Cx^n$ with $n = k'\beta$ as paramcter.

one of them which has degenerated into a straight line ($n = 1$), which is the case when $k' = \beta$. By the adjustment of β, therefore, an image can be obtained in which the fluctuations take place about a straight line. Depending on the character

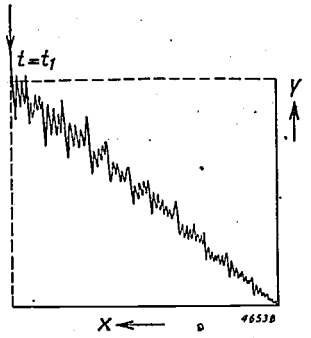


Fig. 12. The image of that part of the reverberation for which $t > t_1$, obtained by the method of the exponential time base.

of the reverberation, this adjustment is more or less sharply defined. In any case it is also possible to study the character of the reverberation precisely. One objection to this method is that only a relatively small part of the whole phenomenon of reverberation can be observed at one time. After a decrease in intensity of about 20 db the voltage applied to the vertical plates becomes so small that the change in it can no longer be clearly observed, due to the size of the light spot. This difficulty can be met by increasing the amplification, so that for $t = 0$, y falls outside the screen. Only at the moment $t = t_1$ at which the light spot, after having fallen vertically, again becomes visible on the screeen, is the horizontal time base switched on (*fig. 12*); the curve appearing on the screen then relates to the part of the reverberation for which $t > t_1$. The moment t_1 is determined by the amplification, so that in this way the behaviour of the reverberation can be studied during different parts of the reverberation time, especially also the case where the reverberation consists of the sum of a number of exponential functions. It is our intention to discuss these apparatus in a subsequent article and to give details of their construction and of the results obtained with them.

COMPARISON OF FREQUENCY MODULATION AND AMPLITUDE MODULATION

by Th. J. WEIJERS.

621.396.619

As a continuation of a previous article on frequency modulation [1]) the circumstances are here discussed under which frequency modulation offers advantages over amplitude modulation for radio broadcasting. It is possible as a consequence of this to extent the frequency spectrum to be transmitted to the highest audible frequencies and there is no need in reproducing music to strain the dynamics. An amplitude limiter must be introduced in the receiver, while, moreover, it is desirable to use a push-pull detector. Further, the frequency sweep must be chosen much larger than the low-frequencies to be transmitted. In order to obtain good quality of reception also the deformation in transmitter and receiver should be kept extremely small. Only on short waves with wave length of not more than a few meters can a sufficiently large frequency sweep be used, and therefore only for such wave lengths is frequency modulation preferable to amplitude modulation.

In a previous number of this periodical [1]) it was explained what is meant by frequency modulation and amplitude modulation, and how in principle with these methods the modulation takes place in the transmitter and the detection in the receiver. We shall now compare the two methods of modulation, and in particular note their relative advantages and disadvantages.

Although amplitude and frequency modulation both make it possible to obtain an undistorted reception of high quality, provided transmitter and receiver, including microphone and loudspeaker, satisfy high requirements, they behave quite differently with respect to interferences: with frequency modulation under certain conditions a much lower noise level can be obtained. In this article we shall explain under which conditions the advantage can be obtained and how it appears in the quality of the reproduction. According to the manner in which the noises appear, we must distinguish two cases. During the modulation of the desired signal the interferences are to a large extent masked by the desired sound; during a pause in the modulation the interfering sounds are in general most annoying. We shall therefore in the first place ascertain what audible interferences occur when in a frequency-modulation receiver, as the simplest case, a sinusoidal interference is added to the unmodulated sinusoidal carrier. We shall then discuss the results of such an interference when the carrier is modulated. Finally we shall deal with the influence on the reception of the interferences due to noise and atmospherics, electrical apparatus and other transmitters.

The case in which a sinusoidal interference of small amplitude and the desired modulation are present alternately.

We shall now ascertain the influence of a sinusoidal interference added to the carrier at a moment when the latter is not modulated. For the sake of simplicity we shall confine ourselves for the present to the case where the amplitude of the interference is small compared with the amplitude of the carrier, for instance smaller than 1/10.

In *fig. 1* the unmodulated carrier *a* and the sinusoidal interference *b*, as well as their sum *c*,

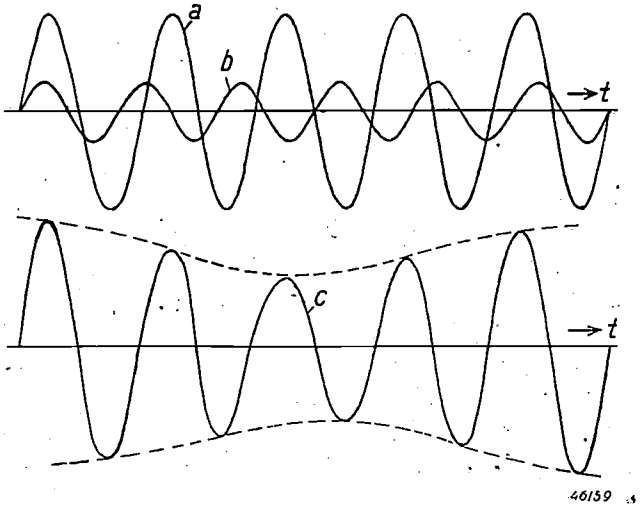


Fig. 1*a*) Unmodulated oscillation.
b) Interfering oscillation with small amplitude and a frequency somewhat higher than that of the oscillation *a*. For the sake of clearness the interference *b* is not drawn < 1/10, but about half as large as the main oscillation *a*.
c) The sum of *a* and *b*; from the form of *c* it can immediately be seen that the disturbance of *a* by *b* amounts not only to an amplitude modulation but also to a frequency modulation.

are plotted as a function of the time. From this figure it is clear that the amplitude of the resulting oscillation *c* is not constant, but that it varies with a frequency equal to the difference between the frequencies of the two oscillations *a* and *b*. The positions where the curves pass the zero axis are also changed, so that the distances between them are no longer equal. Thus, due to the interference, the carrier in question is given an amplitude modulation as well as a frequency modulation. The amplitude modulation has a depth equal to

[1]) Philips techn. Rev. 8, 42, 1946.

the ratio of the amplitude of the interference b to that of the unmodulated carrier a.

In order to determine the sweep of the frequency modulation the various oscillations are plotted in a vector diagram in *fig. 2*. The unmodulated carrier a,

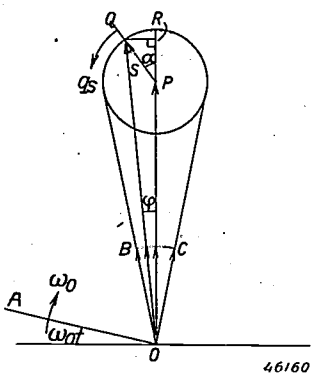


Fig. 2. Vector representation of the oscillations of fig. 1. The stationary vector OP with the length 1 is the unmodulated carrier a with a frequency ω_0 at which the time line OA rotates clockwise; PQ with a length s is the interfering oscillation b with a frequency $\omega_0 + q_s$, so that this vector rotates counter-clockwise around the point P with the frequency difference q_s. The resultant OQ is the modulated oscillation c. The amplitude limiter provides that the vector for the limited oscillation moves back and forth between the positions OB and OC.

whose amplitude we assume to be equal to 1 and which therefore can be represented by $\sin \omega_0 t$, is represented by the stationary vector OP. The interference b which can be represented by $\sin (\omega_0 + q_s) t$, where $s \ll 1$ and q_s is the difference between the angular frequencies of carrier and interference, is indicated by the vector PQ. This latter vector has the length s and rotates about the point P with an angular velocity q_s. The end point of this vector thus describes a circle with the radius s and centre P. The resultant oscillation c is now represented by the vector OQ moving back and forth, whose starting point O remains fixed, but whose end point Q describes the previously mentioned circle. From this it again appears that the amplitude of the resultant oscillation varies between $1-s$ and $1+s$, thus that the interference causes an amplitude modulation with a depth of modulation s. The instantaneous angular frequency of the resultant oscillation is equal to the instantaneous angular velocity of the vector OQ with respect to the time line OA which rotates about O with an angular frequency ω_0. In order to find this angular frequency we first determine the angle φ between OP and OQ. For that purpose we drop the perpendicular QR from Q on OP. The angle α between OP and PQ is $q_s t$, so that $QR = s \sin q_s t$. Since $s \ll 1$, OQ does not differ much from 1, and thus $\sin \varphi$ is approximately equal to $QR/OQ = s \sin q_s t$. Since in our case the angle φ always remains small, we may set $\sin \varphi = \varphi$, so that $\varphi = s \sin q_s t$. The angular velocity of OQ is $\mathrm{d}\varphi/\mathrm{d}t = s q_s \cos q_s t$; with respect to the time line OA, which rotates with an angular velocity ω_0, the instantaneous frequency of the resulting oscillation c is: $\omega_c = \omega_0 + s q_s \cos q_s t$. The frequency sweep caused by this interference is therefore $s q_s$, the product of the ratio s between the amplitudes of the interference and the carrier and the frequency difference q_s between the two oscillations.

We shall now examine how the receiver reacts to an oscillation of the form just discussed. The human ear can only receive sound impressions in a limited frequency region; the auditory limit lies at about 15 000 cycles per second. The angular frequency [2]) corresponding to this auditory limit, which thus amounts to $2 \cdot \pi \cdot 15\,000$, is called q_a.

For the sake of simplicity we shall assume that the frequency spectrum is cut off sharply at this frequency q_a by the low-frequency part of the receiver, the loudspeaker and the ear, so that higher frequencies are not heard and lower frequencies are perfectly heard.

In an amplitude-modulation receiver the detector provides a low-frequency oscillation with a frequency equal to the difference q_s between the frequencies of the two oscillations a and b and an amplitude proportional to the amplitude s of the interfering oscillation b. After amplification this is made audible by the loudspeaker. The frequency modulation which the carrier has obtained from the small sinusoidal interference has no influence on this process of detection; only the amplitude modulation caused by the interference makes itself felt in the same way as a desired amplitude modulation. If the carrier a is sinusoidal amplitude-modulated with a depth of modulation m, the low frequency detected desired signal is proportional to m. Upon maximum modulation $m = 1$. The ratio of the amplitude of the interference during a pause in the desired modulation to the maximum amplitude of the desired signal is therefore at the loudspeaker the same as at the input terminals of the receiver, at least when the difference q_s between the frequencies of the interfering oscillation and the carrier is not larger than q_a. With a larger frequency difference the interfering oscillations are inaudible.

In a frequency modulation receiver the situation is different. There the frequency detector

[2]) In places where it can cause no confusion we shall speak of frequency when angular frequency is meant, as in the article referred to in footnote [1]).

converts a frequency modulated oscillation into an oscillation which also contains amplitude modulation, as was already explained in connection with fig. 9 in the article referred to in footnote [1]). With a frequency sweep sq_s the amplitude modulation depth becomes sq_s/q_1, while the frequency of this modulation amounts to q_s. The oscillation c in question contains, however, also amplitude modulation with a depth

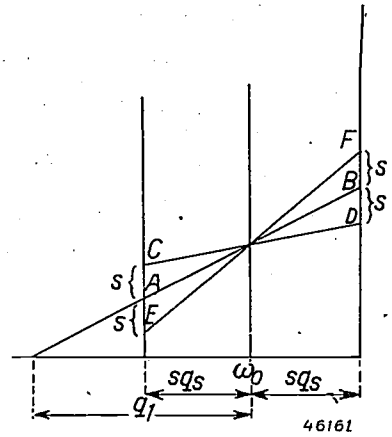


Fig. 3. Variation of the voltage furnished by the frequency detector when the signal is both amplitude and frequency-modulated. The line AB is the approximately straight part of the characteristic which upon extension interscets the frequency axis at a distance q_1 from ω_0.

of modulation s, and the same frequency q_s. The instantaneous amplitude of the oscillation produced by the frequency detector is proportional to the instantaneous amplitude of the supplied oscillation. The amplitude modulation of the oscillation c thus also has an influence on the amplitude modulation of the oscillation which the frequency detector delivers. The voltage which the frequency detector delivers, as a function of the instantaneous frequency, is now not represented by the straight part AB of the characteristic (*fig. 3*), but by the line CD or the line EF, according as the frequency of the interfering oscillation b is lower or higher than the frequency of the carrier a, as may be deduced from the vector diagram of fig. 2. The resultant depth of modulation of the voltage delivered by the frequency detector is $s\ (1 \pm q_s/q_1)$. It therefore seems justified in this simple case simply to add together or to subtract from each other the interferences s and sq_s/q_1, which would be caused by the amplitude and frequency modulation respectively independently of each other. This is in general not permissible, as we shall explain later in this article. In order to obtain the great advantage of frequency modulation, namely the slight influence of interferences, the maximum frequency sweep q should be smaller than q_n; the angular frequency corresponding to the limit of audition (as will appear below) q_s/q_1 is thus always smaller than unity.

The first term in $s\ (1 \pm q_s/q_1)$, caused by the amplitude modulation which the carrier has received from the interference, therefore has the greatest influence. The interference can now in general be diminished by taking care that this term does not occur. This can be achieved by introducing a limiter in front of the frequency detector, which provides that the amplitude of the oscillation which reaches the frequency detector is constant; then only the frequency modulation remains, so that the amplitude modulation depth of the oscillation which the frequency detector furnishes only amounts to sq_s/q. In the vector diagram of fig. 2 the action of the limiter is manifested in the reduction of the length of the vector OQ to a small amount, so that the end point of this limited vector moves along the circular arc BC around O.

Various methods may be followed for the limitation. *Fig. 4* shows a simple, but quite satisfactory circuit. Its action is based upon grid detection. With increasing amplitude of the A.C. voltage on the grid the negative grid bias becomes larger, whereupon the amplification of the valve decreases. With a suitable choice of the elements of the connections, the amplitude of the output voltage depends scarcely at all on the input voltage, within certain limits. *Fig. 5* shows the change taking place in the original signal d by the limiter. Since according to fig. 4 the latter contains resonance circuits, it will be clear that d is not simply cut off at e horizontally, but that the smooth curve f, here drawn with a dotted line, is the result of the limitation. If we assume that the limitation is ideal, thus that it furnishes a frequency-modulated oscillation with a constant amplitude, after detection a low-frequency signal occurs whose amplitude is proportional to sq_s/q_1. If in the absence of interference the signal is sinusoidally frequency-modulated with a frequency sweep $\Delta\omega$, the detector furnishes a desired low-frequency oscillation whose amplitude is proportional to $\Delta\omega/q_1$. The ratio of the

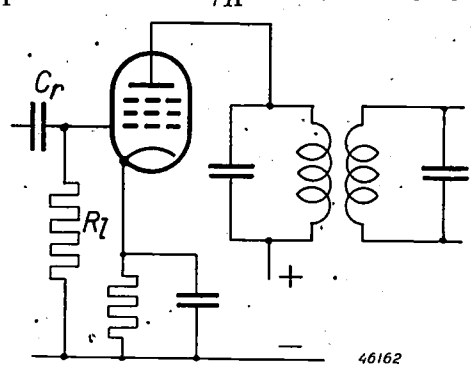


Fig. 4. Diagram showing the principle of an amplitude limiter for a frequency-modulation receiver. C_r grid condenser and R_l leakage resistance.

amplitude of the interference during a pause in the desired modulation to the amplitude of the desired signal upon modulation with a frequency swing $\Delta\omega$ is, therefore, $sq_s/\Delta\omega$. The frequency of this interference is q_s and it can therefore only be heard when $q_s < q_a$. With amplitude modulation this ratio was s. With frequency modulation the influence of the interference may therefore be made much smaller than with amplitude modulation, namely in the ratio $q_s/\Delta\omega$ when the amplitude is considered, *i.e.* in the ratio $(q_s/\Delta\omega)^2$ when the energy is considered. This should be done by taking $\Delta\omega$, the frequency sweep, of the desired modulation much larger than q_a, the highest modulation frequency to be transmitted (this is the modulation frequency corresponding to the limit of audition).

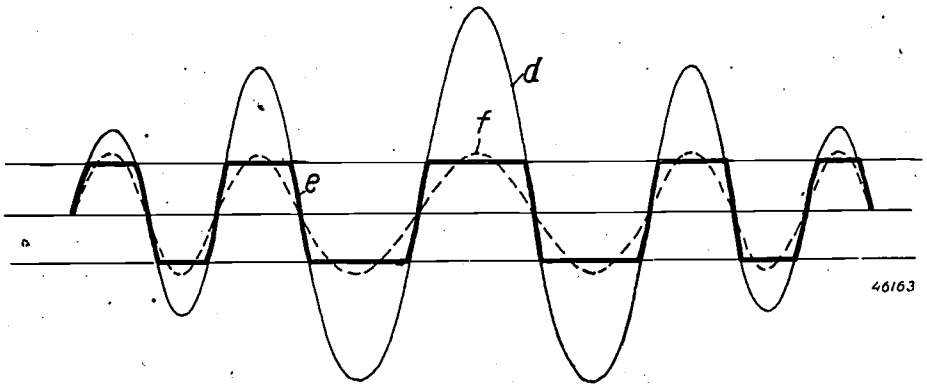


Fig. 5. If the amplitude limiter would only cut off the too high peaks of the amplitude de modulated signal d the heavier drawn line e would remain. In reality the amplitude limiter furnishes the dotted flowing curve f.

From these considerations the conclusion may be drawn that in a receiver for frequency modulation with limiter a simple sinusoidal interference with a small amplitude leads to a less audible interference than is the case in a receiver for amplitude modulation, when the frequency difference between interference and carrier is less than the angular frequency q_a corresponding to the limit of audition. Without a limiter, however, the audible interference may be larger than in the case of an amplitude-modulation receiver. Interferences with larger frequency differences than q_a are inaudible in both methods of modulation, provided they do not occur simultaneously with the desired modulation.

The case where a sinusoidal interference of small amplitude and the desired modulation are present at the same time.

We shall now examine what audible interference is caused by an oscillation with a small amplitude, whose frequency deviates only little from that of the carrier to be received, when the carrier is modulated with an audible frequency. For the sake of comparison we shall first consider the case where the desired signal is amplitude-modulated, while our receiver is intended for amplitude-modulated signals. We shall then consider the case where the desired signal is modulated in frequency, while our receiver is designed for frequency-modulated signals. A sinusoidal amplitude-modulated signal can be represented by $(1 + m \sin pt) \sin \omega_0 t$; the interfering oscillation by $s \sin (\omega_0 + q_s)t$. At small values of s the sum of these oscillations can approximately [3]) be represented by $(1 + m \sin pt + s \cos q_s t) \sin (\omega_0 t + s \sin q_s t)$. The detector in an amplitude-modulation receiver gives as low-frequency oscillation practically: $m \sin pt + s \cos q_s t$. The interference is now heard if $q_s < q_a$; if $q_s > q_a$ no interference can be heard. The amplitude of the interfering sound divided by the amplitude of the desired sound is in this case s/m; at 100 percent depth of modulation of the desired signal this ratio becomes s, just as in the case of the alternate occurrence of the interference and the desired signal.

A sinusoidal frequency-modulated oscillation can be represented by $\sin (\omega_0 t + m \cos pt)$; the interfering oscillation is again $s \sin (\omega_0 + q_s)t$. At small values of s the sum of these oscillations can be approximately represented by:

$$\{1 + s \cos (m \cos pt - q_0 t)\} \cdot \cdot \sin \{\omega_0 t + m \cos pt - s \sin (m \cos pt - q_0 t)\}.$$

We begin with the assumption that our frequency-modulation receiver contains a limiter, which is always the case in practice. This reduces the ampli-

[3]) Stricktly speaking this approximation is only permissible as long as $s < 1-m$, which means that at all times the interference s must be small compared to the momentary amplitude of the desired signal; with great depth of modulation (*i.e.* when m is approximately unity) the formula becomes more complicated. See also T. Ned. Radiogen. **8**, 340, Oct, 1940.

tude of the oscillation to a constant amount, whereby the oscillation which reaches the frequency detector is only proportional to:

$$\sin \{\omega_0 t + m \cos pt - s \sin (m \cos pt - q_s t)\}$$

The instantaneous frequency of this is:

$$\omega_m = \omega_0 - \Delta\omega \sin pt + s (q_s + \Delta\omega \sin pt) \cdot \cos (m \cos pt - q_s t),$$

where $\Delta\omega = mp$. (*cf.* the article referred to in footnote [1]) represents the frequency sweep. The low-frequency signal which is applied to the loudspeaker is proportional to the difference between the carrier frequency and the instantaneous frequency, thus to:

$$\omega_0 - \omega_m = \Delta\omega \sin pt - s (q_s + \Delta\omega \sin pt) \cdot \cos (m \cos pt - q_s t).$$

The first term of this represents the desired signal, the second term the interference. This interference is an oscillation whose amplitude as well as whose frequency changes with the time. The instantaneous frequency of the interference, according to the definition given in the article referred to in footnote [1]), is $q_s + \Delta\omega \sin pt$. Since the preceding selective circuits do not pass the higher frequencies, in the case of the frequency detector only those values of q_s come into consideration which are smaller than q_a, while the maximum value of $\Delta\omega$ is a little smaller than q. The instantaneous frequency of the interference in the detected low-frequency signal thus varies between zero and $q_s + \Delta\omega$. But the velocity at which this instantaneous frequency varies is in general not always small compared with the instantaneous frequency itself; it is therefore of little use here to speak of the instantaneous frequency. The audible interference here has the character of a noise which, in the commonest practical case, namely when $\Delta\omega > q_a$, is only heard for part of the low-frequency cycle. If nevertheless, for the sake of simplicity, but then only in rough approximation, it is desired to continue to speak of instantaneous frequency, it may be said that the interference can only be heard when its instantaneous frequency lies between zero and q_a. We shall here omit a calculation of the ratio between the audible interference and the audible desired signal. In most practical cases, however, the ratio of the energy of the audible interference to the energy of the desired audible signal is smaller with frequency modulation than with amplitude modulation. On the other hand there are cases where with frequency modulation interference is indeed heard, but not with amplitude modulation. With amplitude modulation only those interfering oscillations produce an audible interference whose frequencies lie in the region of $\omega_0 - q_a$ to $\omega_0 + q_a$, while with frequency modulation this region extends from $\omega_0 - q_1$ to $\omega_0 + q_1$, where as a rule $q_1 > q_a$.

Interferences by noise

After having explained in the foregoing the influence of a single sinusoidal interference with a small amplitude, we shall now investigate how the receiver reacts to interferences of less simple nature such as often occur in practice. In the first place we shall consider the interferences due to noises which are caused by the thermal motions of the electrons in resistances and by the corpuscular structure of electricity (electrons) in amplifier valves. This noise can be conceived as an oscillation in which all frequencies occur in equal intensity. As in the foregoing chapter, we shall first study the effect of the noise during a pause in the modulation of the desired signal. For the sake of simplicity we consider the source of noise to be localized at the input terminals of the receiver, which is always permissible because all cases of noise can be reduced to that case. Let us now in the first place consider the case of an amplitude-modulation receiver. Due to the noise at the input terminals uniformly distributed over the whole frequency spectrum we then obtain a noise at the loudspeaker likewise uniformly distributed over all audible frequencies. The energy of the noise at the input terminals of the receiver in every arbitrary frequency band dq_s we call $s^2 dq_s$, when the energy of the desired carrier is set equal to unity. The energy of the noise in the loudspeaker in every frequency band dq_s is also equal to $s^2 dq_s$ when the energy of the detected desired signal is again set equal to unity at 100 percent depth of modulation. In order to find the total effect of the noise this energy must be integrated over the whole audible frequency region, and this is:

$$\int_0^{q_a} s^2 dq_s = s^2 q_a.$$

when the intensity s of the noise is uniformly distributed over the whole audible frequency region. The ratio of the energy of the noise interference to that of the desired signal is therefore $s^2 q_a$.

In the case of a frequency modulation receiver the frequency spectrum of the noise as it reaches the loudspeaker is different from that at the input terminals. In the consideration of a single sinusoidal interference we have already noted that in the presence of a limiter the amplitude-modulation

depth of the oscillation which the frequency detector furnishes during a pause in the modulation amounts to sq_s/q_1. The energy which is represented by this oscillation is then proportional to $s^2(q_s/q_1)^2$. The energy of the noise as it reaches the loudspeaker in the frequency band from q_s to $q_s + dq_s$ is proportional to $s^2(q_s/q_1)^2\,dq_s$. The energy of the noise in the loudspeaker for the whole audible frequency region is then:

$$\int_0^q s^2 \frac{q_s^2}{q_1^2}\,dq_s = \frac{1}{3}\,s^2\,\frac{q_a^3}{q_1^2}.$$

The amplitude of a desired detected signal caused by a sinusoidal frequency-modulated oscillation with a frequency sweep $\Delta\omega$ with the same proportionality factor is $\Delta\omega/q_1$; the energy is thus $(\Delta\omega/q_1)^2$. The ratio of the energy of the noise interference to that of the desired signal is thus $^1/_3\ s^2q_a^3/(\Delta\omega)^2$. This ratio in the case of frequency modulation is $^1/_3\ (q_a^2/\Delta\omega)^2$ times that in the case of amplitude modulation. Frequency modulation is in this respect therefore more advantageous than amplitude modulation if the frequency sweep $\Delta\omega$ is taken much larger than the highest audible frequency q_a to be transmitted, as was the case with a single sinusoidal interference.

We shall explain this result on the basis of a figure. The amplitude of the low-frequency signal after detection of a frequency-modulated oscillation is proportional to the frequency sweep. After detection the amplitude of the single sinusoidal interference in the absence of the modulation of the desired signal, divided by the amplitude of the desired signal with sinusoidal modulation, is equal to $sq_s/\Delta\omega$. In *fig. 6* OA represents the amplitude 1 and OB the maximum frequency sweep $\Delta\omega$ of the desired signal; BC the amplitude s as this was defined in the foregoing for an interference by noise, the line OC the amplitude of the oscillation which the frequency detector furnishes during a pause in the modulation amounts to sq_s/q_1. The energy which is represented by this oscillation is then proportional to $s^2(q_s/q_1)^2$. The energy of the noise as it reaches the loudspeaker in the frequency band from q_s to $q_s + dq_s$ is proportional to $s^2(q_s/q_1)^2\,dq_s$. The energy of the noise in the loudspeaker for the whole audible frequency region is then:

$$\int_0^q s^2 \frac{q_s^2}{q_1^2}\,dq_s = \frac{1}{3}\,s^2\,\frac{q_a^3}{q_1^2}.$$

The amplitude of a desired detected signal caused by a sinusoidal frequency-modulated oscillation with a frequency sweep $\Delta\omega$ with the same proportionality factor is $\Delta\omega/q_1$; the energy is thus $(\Delta\omega/q_1)^2$. The ratio of the energy of the noise interference to that of the desired signal is thus $^1/_3\ s^2q_a^3/(\Delta\omega)^2$. This ratio in the case of frequency modulation is $^1/_3\ (q_a^2/\Delta\omega)^2$ times that in the case of amplitude modulation. Frequency modulation is in this respect therefore more advantageous than amplitude modulation if the frequency sweep $\Delta\omega$ is taken much larger than the highest audible frequency q_a to be transmitted, as was the case with a single sinusoidal interference.

We shall explain this result on the basis of a figure. The amplitude of the low-frequency signal after detection of a frequency-modulated oscillation is proportional to the frequency sweep. After detection the amplitude of the single sinusoidal interference in the absence of the modulation of the desired signal, divided by the amplitude of the desired signal with sinusoidal modulation, is equal to $sq_s/\Delta\omega$. In *fig. 6* OA represents the amplitude 1 and OB the maximum frequency sweep $\Delta\omega$ of the desired signal; BC the amplitude s as this was defined in the foregoing for an interference by noise, the line OC the amplitude s as this was defined in the foregoing for an interference by noise, the line OC the ampli-

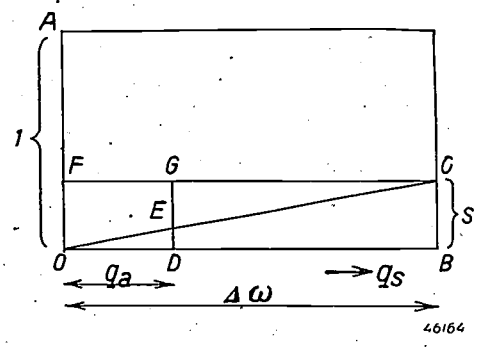


Fig. 6. The amplitude of the detected interference in the case of a frequency-modulation receiver as a function of the frequency difference q_s between the desired carrier and the interfering signal is represented by the straight line OC through the origin. In the case of an amplitude modulation receiver the horizontal line FG gives the corresponding relation.

modulation the amplitude increases linearly with increasing q_s; for $q_s = 0$ this amplitude is zero, for $q_s = q_a$ this amplitude is $sq_a/\Delta\omega = DE$ in fig. 6. Only the noise components whose frequency is smaller than q_a (OD in fig. 6) contribute to the audible interference. The energy of the interfering noise in a given frequency band dq_s is proportional to the square of the amplitude. The total audible noise energy is thus found by integrating the squares of the ordinates in fig. 6 from zero to q_a. When this integration is performed the already mentioned ratio $(^1/_3\ q_a/\Delta\omega)^2$ is again obtained for the energy of the noise with respect to that of the desired signal.

Since the influence of the noise in frequency modulation can be made small by choosing the frequency sweep $\Delta\omega$ much larger than the limiting frequency q_a of the audible region, it might be thought that by choosing $\Delta\omega$ larger and larger the noise could be diminished to an unlimited extent. This, however, is incorrect, since the frequency sweep $\Delta\omega$ cannot be increased directly above a definite limit. We shall, however, return later to this question.

In the foregoing discussion it was always assumed that a limiter was present in the receiver in front of the frequency detector; moreover, the influence of the noise was only studied during a pause in the desired modulation. If, however, the limiter is absent and if the desired and undesired modulations are present simultaneously, one finds for the relation of the quotient of noise energy and the energy of the desired signal with frequency modulation to that with amplitude modulation:

$$\frac{1}{3}\left(\frac{q_a}{\Delta\omega}\right)^2 + \left(\frac{q_1}{\Delta\omega}\right)^2 + \frac{1}{2}.$$

which expression we shall not here derive. The first of these three terms is the same as what we found above for the case where a limiter was present and is therefore due to the frequency modulation caused by the noise. The second term is at least equal to unity and represents the interference contribution caused by the amplitude modulation which the unmodulated carrier receives as a result of the noise. The origin of the third term is not so easy to indicate, since it only occurs when interference and desired frequency modulation are present simultaneously, so that we may call it the interaction term. It may be remarked in conclusion that the interference energy due to noise in the case of frequency modulation without limiter is more than $1^1/_2$ times as large as in the case of amplitude modulation. If, however, there is a limiter present then also during the

tude of the detected interference as a function of the frequency difference q_s between the desired carrier and the interfering signal in the case of a frequency modulation receiver. With amplitude modulation all audible noise components take on the same amplitude, namely OF; with frequency presence of the desired modulation the above mentioned relation is the same as during a pause in the desired modulation, *i.e.* only $^1/_3\ (q_a/\Delta\omega)^2$, which with a sufficiently large value of the frequency sweep $\Delta\omega$ is small compared with unity.

The effect of the limiter thus consists in the second and third terms, which amount respectively to at least 1 and $^1/_2$, being eliminated, so that only the first term is retained. It is therefore advisable to make the first term as small as possible, as has just been explained.

In practice, however, a limiter is never perfect; some influence of the amplitude modulation always remains. If, however, a pushpull detector is used, as discussed in the article referred to in footnote [1]) where the output voltages of the two diodes are opposite, then even when there is no limiter present the influence of the amplitude modulation is eliminated in as far as the second term of the expression just discussed disappears, but the third term, the reciprocal action term, then remains. The best result is therefore obtained by using a limiter and a push-pull detector together. In that case the interference energy found experimentally is indeed found to correspond to the small first term of our expression.

Interferences with large amplitudes

Until now it has been assumed that the amplitude of the interferences due to noise or other causes was small compared with the amplitude of the desired signal. If, however, this is not the case, the decrease of the interference with frequency modulation compared with that with amplitude modulation is not so great. If the amplitude of the interference is larger than the amplitude of the desired signal, then the interference is not drowned out by the signal, but the signal by the interference. As long as the interference is smaller than about half the desired signal, frequency modulation offers advantages over amplitude modulation, although to a smaller degree than would correspond to the expression given for that: $(q_s/\Delta\omega)^2$ for discrete interferences and $^1/_3\ (q_a/\Delta\omega)^2$ for noise. If the average amplitude s of the noise is smaller than about 1/10 of that of the desired signal, the interference is practically not larger than the expression gives. If for example we choose the frequency sweep $\Delta\omega = 5q_a$, the ratio of the interference energy with frequency modulation to that with amplitude modulation is 1/75, thus in our case the average amplitude of the noise: $\sqrt{1/75} \cdot 1/10 = 1/86$ of the amplitude of the desired signal, which still makes an acceptable reception possible. With amplitude modulation, however, with a noise with an average amplitude equal to 1/10 of the amplitude of the desired signal, satisfactory reception of broadcasting is out of the question. It has been found in practice that a voltage of about 10 μV on the input terminals of a well-constructed frequency modulation receiver is sufficient for satisfactory noise-free reception in the absence of other interferences, while with amplitude modulation about 1000 μV are necessary under the same conditions.

The amplitude of the interferences at the limiter is the decisive factor in this respect. This amplitude depends upon the frequency region passed by the band filters; the width of this region depends upon the maximum frequency sweep for which the receiver is built, because the high-frequency and intermediate-frequency band filters must pass a frequency band of at least $2\Delta\omega$. The average amplitude of the noise is proportional to the square root of the band width. The larger the frequency sweep $\Delta\omega$, therefore, the larger the noise amplitude at the limiter. But, as was already noted, the characteristic advantages of frequency modulation disappear if the noise amplitude at the limiter is no longer smaller than about half the amplitude of the desired signal. The frequency sweep cannot therefore without danger be increased indefinitely. If one assumes 15 000 cycles per second to be the highest number of oscillations per second to be observed, which therefore it is desired to receive from the desired modulation and which the low-frequency part of the receiver including the loudspeaker must reproduce, the practically most favourable value for the frequency sweep is about 75 000 cycles per second; thus calculated as angular frequency: $\Delta\omega = 2\pi \cdot 75\,000$.

Interferences from electrical apparatus and atmospherics

The influence of these interferences, which usually consist of several short impulses, cannot be accurately given in a general form because of their irregular character. With equal effective values, the maximum values of these interferences are in general larger than those of interferences due to noise. If the maximum value of the interferences reaches the level of the desired signal, the interference will be considerable during these maxima, *i.e.*, not smaller than with amplitude modulation, while

during the time that the momentary value of these interferences is small, little or no disturbance is caused by them. The character of these interferences is thus quite different with frequency modulation than with amplitude modulation, in that with not too violent interferences frequency modulation with a larger frequency sweep is still more advantageous than amplitude modulation.

Interference by other transmitters

With respect to interferences by other transmitters frequency modulation is in general less subject to disturbances than amplitude modulation. An amplitude-modulated interfering transmitter with a frequency lying in the region that is occupied by the instantaneous frequency of a desired frequency-modulated transmitter behaves in this respect in the same way as a single sinusoidal interference. Thus if the interfering signal is much smaller than the desired signal it will cause little trouble. A frequency-modulated interfering transmitter which gives a smaller voltage in the receiver than the desired transmitter, and which occupies a frequency region coinciding partially or wholly with that of the desired frequency-modulated transmitter, will give little interference, in the first place for the same reason as in the case of other interferences and in the second place because in a first approximation an interference is only heard when the difference between the "instantaneous" frequencies of the two transmitters lies in the audible region, which will only be the case during part of the time, especially if the frequency sweep is large. If, however, the interfering frequency-modulated signal is not much smaller than the desired frequency-modulated signal, it is necessary, for interference-free reception, that the regions of the instantaneous frequencies of the two transmitters should not overlap. The frequency distance between the carriers must therefore in that case amount to at least twice the frequency sweep.

Conclusions

We have explained that the use of frequency modulation in radio broadcasting may under certain conditions offer advantages over amplitude modulation. These advantages consist mainly of less trouble with noise and other interferences, provided the amplitude of these interferences is smaller than the amplitude of the desired signal. In order actually to obtain these advantages it is necessary that an amplitude limiter should be introduced into the receiver, whilst, moreover, a push-pull detector is desirable. The frequency swing must be considerably larger than the highest low frequencies to be transmitted. A favourable practical value for the frequency sweep for broadcasting is 75 000 cycles per second. With this large frequency sweep the region practically occupied in the frequency spectrum is almost independent of the modulation frequency. It is therefore possible that the audio-frequency region can be extended to the auditory limit, for example, 15 000 cycles per second, without increasing the region occupied by the frequency spectrum, whereby the quality of reception can be improved to the utmost. If it is indeed desired to profit by this great advantage it is necessary that the transmitter (from microphone to aerial) as well as the receiver (from aerial to and including loudspeaker) should be designed for that purpose. If, however, this extension of the frequency region is actually to produce an improvement in quality, it is in addition required that the distortion in transmitter and receiver should be unusually small.

In the case of radio broadcasting with amplitude modulation it is customary to diminish the large differences in intensity which are indispensable for enjoyable reproduction of music[4]). This is necessary in that case, since otherwise the soft passages would be drowned out by the noise and the strong passages would be very much distorted. In the case of frequency modulation, however, this is not necessary since in practically the whole region in which the transmitter can be received free of interference no noise can be heard even during the softest passages. With frequency modulation, therefore, the broadcasting company is able to attain a quality of reproduction of music in which no violence has been done to its dynamics. Due to the large frequency sweep frequency modulation in the manner described can only be employed on short waves ($\lambda < 10$ m); for longer waves amplitude modulation is therefore to be preferred. Short waves can only be received in a limited region since they are not reflected by the ionosphere. Because of this the same carrier frequency can be repeated at a certain distance (for instance 300 km) without difficulty. Since the amplitude ratio of interfering and desired transmitters in the case of frequency modulation does not need to be so small for freedom from interference as with amplitude modulation, the aera of the earth's surface where two transmitters with the same carrier frequency interfere with each other is moreover much smaller.

[4]) See for instance: V. Cohen Henriquez, Philips techn. Rev. 3, 204, 1938.

VOL. 8 No. 4, p. 97-128 APRIL 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

A NEW METHOD OF COUNTERACTING NOISE IN SOUND-FILM REPRODUCTION

by W. K. WESTMIJZE.

534.862.4

Reproduction of the sound recorded on sound-film is usually accomplished by means of a narrow beam of light thrown upon the film in a direction perpendicular to that in which the sound track is moving. The fluctuations in the light flux passed through are converted into sound. With this method a noise results which is caused by the fact that part of the light passed through is intercepted by specks of dust, scratches, etc. on the sound track, especially when the film has already been used several times. This articles describes a method of counteracting this noise in cases where the sound is recorded as so-called amplitude writing. The beam of light is replaced by a series of equidistant light spots moving with great velocity perpendicular to the sound track. In addition to the theoretical fundamentals of the method, a practical form of application is also discussed.

The ordinary method of reproduction

The reproduction of sound recorded on sound film is usually reproduced in the following manner. A narrow beam of light is thrown on the film perpendicular to its direction of motion. Confining ourselves to the case where the sound is recorded as so-called amplitude-writing, such as for example with the Philips-Miller film [1]), the quantity of light passing through the film depends upon the width of the sound track (and of course of the beam). The light passes through to a photocell and is converted into an electric current which may be considered as a direct current upon which an alternating current is superposed. The magnitude of this direct current depends upon the width of the so-called zero track, *i.e.* the track which is made when no sound vibrations are being recorded. The zero track is unavoidable, since otherwise modulation would be impossible. It is easy to understand that its width must be equal at least to once or twice the maximum modulation amplitude, according as the modulation takes place on one side or on both sides of the track.

The alternating current depends upon the modulation of the track and thus on the sound vibrations recorded, and if the light beam were infinitesimally narrow the trend of this current would be an exact copy of the sound vibrations. Actually the beam has a finite width Δ, but even so the relation between the sound vibrations recorded and the corresponding vibrations of the light flux can easily be determined. Let us assume that the sound track is modulated by one harmonic vibration. Such a vibration is represented in *fig. 1*. When

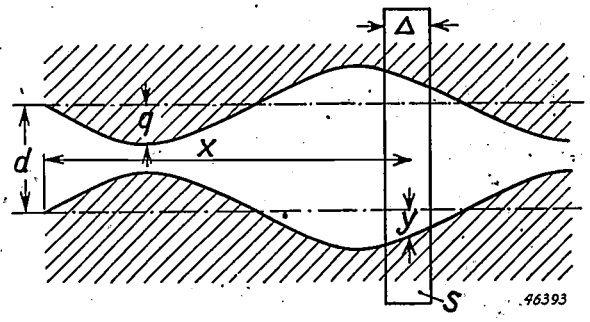


Fig. 1. Diagram of the usual method of scanning. The film with the modulated sound track travels past this beam S. The variations in the light flux passed through are registered by a photocell. In the diagram the track is modulated on both sides by a purely sinusoidal vibration. d width of the unmodulated track, q amplitude of the vibration with which the track is modulated, y depth of modulation at the point with the abscis x, Δ width of the slit.

this vibration corresponds to a tone of ν oscillations per second, and when the velocity at which the film is travelling is v cm/sec, there are ν/v vibrations per cm of film and the vibration can be represented by the equation $y = q \cos 2\pi \cdot \nu/v \cdot x$, where γ is the depth of modulation and x the length of film passed, measured from an arbitrary zero point,

[1]) For a description of the Philips Miller system, see Philips Techn. Rev. 1, 107, 135, 211, 1936.

The amount of light passed through is then proportional to

$$\Delta \cdot d + 2 \int_{x-\frac{1}{2}\Delta}^{x+\frac{1}{2}\Delta} q \cos\left(2\pi \frac{\nu}{v} \xi\right) d\xi = \Delta \cdot d + \Delta q \cdot \frac{\sin \pi \frac{\nu}{v} \Delta}{\pi \frac{\nu}{v} \Delta} \cos 2\pi \frac{\nu}{v} x .$$

d representing the width of the zero track. From the result it is immediately clear that a DC and an AC component are present, while it is also clear that the amplitude of the AC component is multiplied by a factor which depends upon the frequency ν. This factor

$$\frac{\sin \pi \frac{\nu}{v} \Delta}{\pi \frac{\nu}{v} \Delta}$$

is equal to unity when $\nu = 0$, and then decreases. In order that the highest frequencies to be reproduced should not be attenuated by more than about 3.5 db compared with the lowest (such an attenuation is still permissible) it is necessary that

$$\pi \cdot \frac{\nu_{max}}{v} \cdot \Delta < 1.5 .$$

With $\nu_{max} = 8000$ and $v = 32$ cm/sec this results in $\Delta < 0.002$ cm. The light-beam may therefore not be wider than 20 μ. [2]).

When there are specks of dust or dirt on the sound track or when it has been scratched, as is particularly the case with much used sound films, these tiny specks and scratches, irregularly distributed over the surface of the film, cause a noise. They cannot, however, be observed individually, as is the case with larger particles (> 80 μ), which cause an annoying ticking or bubbling sound. It would mean a considerable improvement in reproduction if this noise could be avoided.

For some time already a system has been in use which diminishes this noise. It is based on the following principle. The noise is most annoying during the soft passages, *i.e.* when the depth of modulation is slight. In sound recording it is now arranged, by means of suitable connections, that during these passages the zero track becomes narrower, thus reducing the area upon which the troublesome specks or scratches may occur and thereby also the noise. During the louder passages the zero track again becomes wider, and thus also the noise becomes louder, but this is less troublesome here because for the greater part it is drowned out by the music or speech.

This method, therefore, does not eliminate the noise, but only reduces it during the soft passages.

Principle of high-frequency scanning

We have seen that in the scanning method described above the noise is caused by contaminations on the transparent part of the film between the two edges of the sound track. This phenomenon therefore also occurs when the edges of the track, which actually represent the sound, are ideal. With the method of high-frequency scanning, about to be discussed, only the edges of the track are scanned; the influence of the part between the edges is eliminated and thus also the noise, in so far as it is caused by specks on the transparent part of the film. Of course the noise resulting from imperfections in the edges of the sound track, to which we shall return later, still remains, just as with the method of zero track adaptation discussed above.

With this method of scanning, instead of a narrow slit of light, we have a series of light spots moving at a very high velocity and at regular intervals perpendicularly across the film. Since the sound track is also moving, the light spots actually move in an oblique direction across the film. Here, too, the light passing through falls on a photoelectric cell, which gives a current impulse during the time that the light spot is moving between the edges of the track. The image of this impulse is approximately rectangular. The intensity of the impulse is determined by the intensity of the beam of light employed. The duration of the impulse depends upon the width of the track at the point where the light spot crosses it. Thus in *fig. 2 AB* in the lower half corresponds to *ab* in the upper half, the same being true of *CD* and *cd*, etc. It is essential to note that *the beginning and end points of the blocks* are fixed by the *edges of the sound track.* (For the sake of clearness the obliqueness of the paths of the light spot across the film is exaggerated.) Contaminations on the film are manifested by variations in the beam of light passed through and consequently the image of the current impulses is not actually as shown in fig. 2b, but as in *fig. 3;* between *A* and *D* the current is not constant, variations occurring of an accidental nature. The great advantage achieved lies, however, in the fact that the disturbances are separated from the phenomenon to be reproduced, the former affecting

[2]) *Cf.* J. F. Schouten, Synthetic Sound, Philips Techn. Rev. 4, 167, 1939.

the height of the blocks, while the latter only affects the beginning and end points of the blocks.

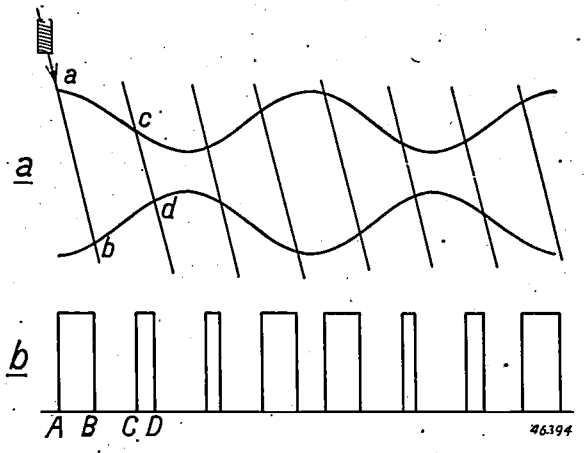


Fig. 2. Diagram of the high-frequency method of scanning. A series of equidistant spots of light travel at a high velocity across the film. Owing to the fact that the film is also travelling at the same time, the light spots describe paths which are oblique with respect to the film and which are given in fig. *a*. The slope of these paths is very much exaggerated for the sake of clearness. As long as a light spot is inside the edges of the track, a current flows in the photocell. The form of the signal leaving the photocell is shown in fig. *b*. The block AB corresponds to the path ab, etc.

Therefore the disturbances can easily be eliminated by sending the whole signal through a limiter which only passes signals up to a certain amplitude. In this way the disturbances are, as it were, cut off. For the current variation shown in fig. 3 a limitation to the level EF would be sufficient to bring about this elimination. If the signal is afterwards so amplified that the amplitude is increased in the ratio BA/EA, a signal is obtained which is absolutely identical with what would have been obtained if the sound track had been everywhere uniformly transparent.

We must now consider the question as to how we can derive the original sound frequencies from the block-signal. The frequency spectrum of this

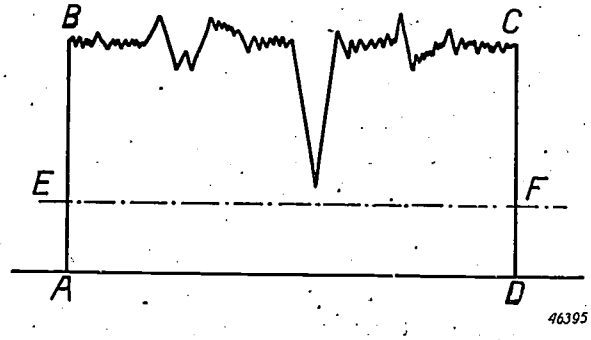


Fig. 3. Image of a current impulse from the photocell. The intensity variations are due to contaminations on the film in the path of the scanning light spot. The essence of the method lies in the fact that the influence of the contaminations can be eliminated by passing the signal through a limiter. Limitation to the level EF would in this case be sufficient.

signal must first be investigated. This involves complicated calculations which will be further dealt with on another occasion. Suffice it here to go into a few qualitative considerations. Let us first examine the unmodulated signal. It consists of congruent blocks having a frequency μ (the scanning frequency). If a Fourier analysis is made of this signal, vibrations with the frequencies μ, 2μ, 3μ, etc. are obtained. If we now modulate the block signal with a frequency ν, secondary frequencies then appear in the spectrum: $\mu \pm \nu$; $\mu \pm 2\nu$; $\ldots 2\mu \pm \nu$; $2\mu \pm 2\nu$; $\ldots 3\mu \pm \nu$; etc. It is, however, quite obvious that also the frequency ν itself will occur. Let us again consider fig. 2. The blocks corresponding to the wide parts of the track are wide and consequently the wide parts give rise to relatively long current impulses with short interruptions. In the case of the narrow parts of the track it is just the reverse. If we now pass this signal through a suitable filter, *i.e.* a low-pass filter, with limiting frequency coinciding with the highest frequency that has to be passed through, the result is that the signal, roughly speaking, is replaced by a progressive average over a certain time interval approximately of the order of 1/4 of the time of vibration of the limiting frequency. Thus in each case a number of successive blocks is averaged and the result is a signal which is strong when the blocks are wide and weak when they are narrow, thus an alternating current with a frequency ν corresponding to the frequency of the vibration originally registered.

For reproduction it is essential that the frequency ν should occur but that 2ν, 3ν etc. should be absent. That this is indeed the case is proved by calculation, though it is not easy to imagine. It is obvious, however, that this is of importance, for, as a rule, with ν also 2ν and possibly 3ν etc. lie in the audible range.

We can now also make it clear that the scanning frequency μ must be much greater than the highest frequency ν to be reproduced, because in addition to μ owing to the modulation also the tones $\mu - \nu$, $\mu - 2\nu$, etc. occur. These tones become weaker as we get farther away from the frequency μ.

Calculation shows that the frequency $\mu - 5\nu$ is already 60 db weaker than the frequency ν. The frequency $\mu - 4\nu$ would still be strong enough to be disturbing. If we are to eliminate this by means of a filter, then it must fall outside the audible range, and this means that:

$$\mu - 4\nu_{max} > \nu_{max}, \text{ or } \mu > 5\nu_{max},$$

where ν_{max} represents the highest frequency of the audible region which is to be reproduced.

Taking $\nu_{max} = 8000$ c/sec, it follows that $\mu = 40\,000$ c/sec.

Limitations of the effect of the method

It must be pointed out that not all disturbances can be eliminated in the manner described. Two cases must be examined separately.

In the first place a contamination may be so large and consequently intercept so much light as to cause the photocurrent to fall below the value that passes through the limiter. The result is a "dent" in the corresponding block which again causes a disturbance. This is especially the case when the light is entirely cut off by the contamination, in which event one light spot produces two current impulses (blocks). However, by giving the light spot an oblong shape it is possible to ensure that this case seldom occurs. Already in the beginning of this article it was observed that the width of the light beam in the ordinary method of scanning may not be more than 20 μ, because otherwise the high tones would be weakened. This applies also for the width of the light spots, but not for their height. (By width we mean here the dimension perpendicular to the motion of the light spots and by height the dimension parallel to that motion.) An increase in the height for instance to 100 μ has by first approximation the same effect on the fluctuations of the transmitted light flux as if the zero track had been taken $100 - 20 = 80$ μ wider and the height left the same. This can easily be explained: Owing to the finite height of the light spots the photocurrent impulses do not have the form of rectangles (apart from the disturbances due to contaminations), but of equilateral trapezia. During the time that the light spot is moving over the edge of the track, the intensity increases from zero to the maximum value and decreases again from the maximum to zero. In *fig. 4* two cases are depicted for different heights of the light spots. It is assumed that they begin to pass over the track at the same moment. The photocurrent impulses then begin at the same moment for both, at the point A. We further assume, of course, that the two light spots move at the same velocity, so that the light intensity increases in the same way and the trend of the photocurrent will be the same in both cases, for instance along AB. A difference occurs only when the lowest light spot is completely over the track, let us say at B. From that moment the corresponding photocurrent (except for disturbances) remains constant. For a short while, however, the current corresponding to the highest spot continues to increase at the same rate, until this spot is also entirely over the track, let us say at B', from which moment the second current, too, is (practically) constant.

As soon as the upper edge of one of the spots has reached the other side of the track, the corresponding current begins to decrease again. Under our assumptions this will take place at the same moment for both currents and the points at which this takes place, namely C and C', lie vertically above each other. The decrease is at the same rate as the increase and thus equal for both spots (the current curves are equilateral trapezia). The

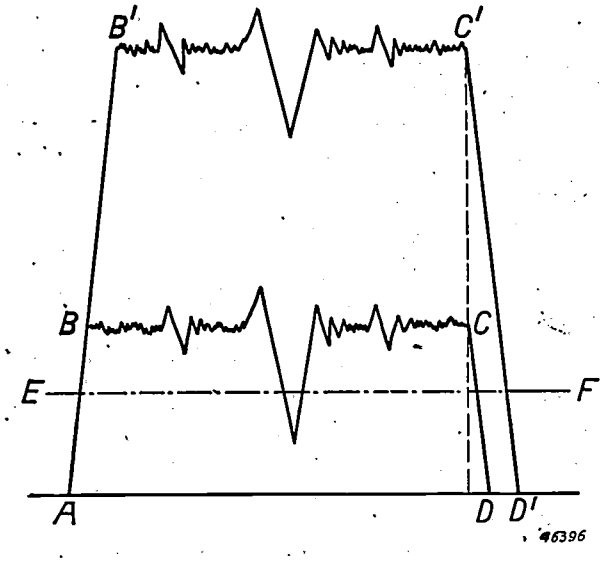


Fig. 4. Influence of the height of the light spot on the form of the photocurrent impulses excited by the light spot. Owing to the fact that the spot has a finite height, some time elapses before the whole spot is over the track. During that time the current increases continually. The impulse $ABCD$ is due to a low spot, the impulse $AB'C'D'$ to a higher one, the top-side of both spots having reached the edge of the track at the same moment. In the second case the average current of the photosignal is larger. Limitation of the signal to the level EF is therefore sufficient to eliminate all disturbances in the second case but not in the first case.

currents thus decrease according to two parallel straight lines, $CD // C'D'$. Therefore they do not end at the same moment. The difference DD', however, is entirely determined by the difference in intensity CC' (and the velocity of the spots, which is however, the same for both), and this in turn depends exclusively on the difference in height of the spots. If we pass the two signals through the same limiter then from our reasoning it follows that the signals finally obtained differ only in length, but that this difference is the same for all blocks and therefore has no effect on the sound to be ultimately reproduced. It only alters the DC component of the photocurrent signal, just as a change in the width of the zero track would do, and this is suppressed by a filter. If the height of

the spot is greater than the width of the track, the situation is somewhat different, but a closer investigation shows that in this case too the length of the blocks of the limited signal, except for a constant, is proportional to the width of the track at the place where the light spot passed.

From the foregoing it will be clear that it is possible to choose such a height of the spot that practically speaking the transmitted light cannot be cut off by contaminations to such a degree that after limitation such disturbances still have any effect. This is in fact demonstrated in fig. 4. The absolute changes in intensity of the transmitted light beams resulting from contaminations are the same for both spots. Therefore the noise assumed to be present in this case is without influence on the limited signal with the higher spot, but with the lower spot it does leave a disturbance in the limited signal.

The second possibility of disturbances occurs when a contamination lies exactly tangent to or across the edge of the track. This alters the form of the limitation. The disturbance caused by such an imperfection in the edge of the track is not eliminated by the method discussed here. The chance of such a disturbance occurring, however, is slight compared with that caused by a speck elsewhere on the track. The modulated track is at least 1 mm wide, so that the chance of contaminations, even of the size of 100 μ, coming to lie at the edge is only 20%; most contaminations, however, are much smaller and there is therefore still less chance of their lying at the edge of the track.

One possible construction of the apparatus

The above-described high-frequency scanning can be realized in different ways. In the first place the sound track can be scanned by a moving light spot. as has been assumed in the foregoing. In principle the same results can be attained by projecting the image of the sound track and causing this image to vibrate with respect to a diaphram. In both cases we may consider the vibration as being brought about with a moving light source and a stationary optical system, but also with a stationary light source and a moving optical system. Finally the vibrations may be construed as being brought about by electrical means as well as by mechanical means. We shall here confine ourselves to the description of a method worked out by us in which the scanning is accomplished with a moving light spot obtained from a mechanically moved optical system.

In *fig. 5* a diagram is given of the arrangement employed. The light from a linear source is projected by a lens several mm from the edge of a disc which can be rapidly rotated. In this disc radial slits have been sawed beginning at the edge. When the disc is rotating rapidly, therefore, each slit allows a fraction of the light from the image to pass through. The image of the illuminated opening is

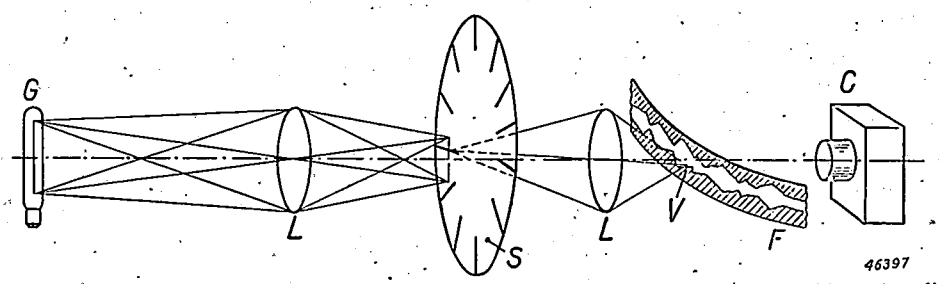


Fig. 5. Diagram of the set-up for high-frequency scanning. *G* source of light with linear filament. *L* lenses. *S* rotating disc with slits. *F* film with modulated sound track. *V* light spot. *C* photocell.

focussed on the sound film by means of a second lens. The light passed through the sound track falls on a photocell and gives rise to the photocurrents already mentioned.

The practical realization of such a set-up involves a number of technical difficulties which we shall now discuss.

The choice of light source

We have already remarked that the width of the light spot on the film may not amount to more than 20 μ. Furthermore it must be very sharp (the transition from light to dark must take place within a distance of not more than a few μ) and not only when the projection is along the axis of the system, but also when the image is about 1 mm above or below it. Finally the light must be of sufficient intensity to excite a reasonably amplifiable photocurrent. These conditions make certain demands on the optical system and the source of light.

Linear light sources whose incandescent body is narrower than 80 μ are difficult to produce.

This implies that the optical system must be a reducing one. The same conclusion is reached from the requirement of sharpness of projection. A five-fold reduction suffices for both requirements. This reduction is mainly effected by the second lens. The first lens gives practically an image of 1 : 1. The requirements for sharpness of the image make it necessary to work with small opening angles.

Finally from the minimum required light intensity of the beam that falls upon the photocell and from the dimensions of the optical system it is to be deduced that the brightness of the light source employed must be at least one thousand candle power per cm^2. In order to satisfy these requirements a special lamp was constructed.

Construction of the rotating disc

The greatest difficulty lay in the construction of the disc. As already mentioned, the required frequency of the light spots is 40 000. The width of the track for Philips-Miller film can be set at a maximum of 1.6 mm, hence a velocity of the light spots of 6400 cm/sec. Since, as mentioned above, the second lens reduces by a factor 5, this leads to a peripheral velocity of the disc of 32 000 cm/sec. Now the peripheral velocity determines the stresses occurring in the disc. Similar discs of different diameters but with equal peripheral velocities exhibit exactly the same stresses at corresponding points. At a velocity of 32 000 cm/sec. these stresses are enormous and approach the yield point. It is clear that this sets an upper limit for the velocity. In fact if this limit is reached the disc flies to pieces.

Since for different materials under otherwise similar conditions the stresses are proportional to the specific weights, a material had to be found with the most favourable ratio of yield point to specific weight. Moreover, having regard to the motive power for the disc, the material had to be electrically conductive, so that practically only duraluminium and electron could be considered. Furthermore, since the highest stresses occur where the hole is drilled for the spindle, the disc was given a very slightly conical profile.

It can then be calculated that both for dualuminium and for electron the maximum stresses occurring, even at a velocity of 40 000 cm/sec, still remain below half the yield point value. This was in fact confirmed experimentally by investigating at what peripheral velocity a test disc flew to pieces. This was found to be at 60 000 cm/sec (the stresses are proportional to the square of the velocity). From *fig. 6*, which is a photograph of the fragments of the disc, it is apparent that the break began at the spindle, as was to be expected.

Furthermore, as it was desirable not to make the apparatus too cumbersome, the disc could not be made too large. Its radius was therefore fixed at 5 cm. This means that a speed of rotation of $32\,000/10\pi = 1000$ rev. per sec. is required. Since the slits have to be about $5 \times 1.6 = 8$ mm apart, $10\pi/0.8 =$ approx. 40 slits can be made on such a disc. They are 0.6 mm wide and 3.5 mm long (from this it follows that the length of the light spots on the film is 120 μ). The cutting of the slits requires much care.

In the first place they have to be spaced at exactly equal distances and must be exactly alike, as otherwise the frequency of revolution of the disc appears in the frequency spectrum, and since this lies in the audible region there will be a whistling tone in the sound reproduced. The scanning frequency, which is 40 times as high, lies, as we know, outside this region.

In the second place very careful finishing is essential because otherwise at the high speeds of rotation the disc might crack at the slits. For that reason before the slits are cut small holes are drilled at the spots where the slits end.

Bearings and motive power of the disc.

With the above mentioned very high number of

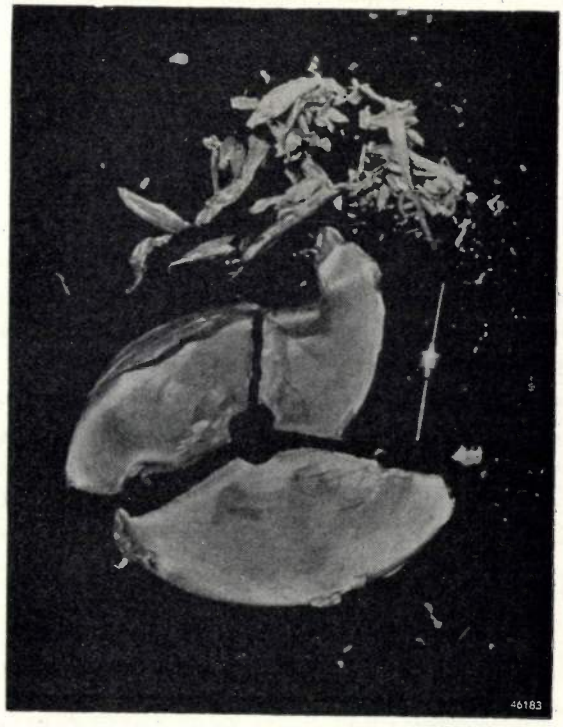


Fig. 6. Fragments of a rotating disc which flew to pieces at a peripheral velocity of about 60 000 cm/sec. The photo shows that the fracture began where the spindle passed through.

revolutions of 1000 per sec special demands are of course made of the bearings. Even a slight excentricity of the centre of gravity of the disc with respect to the centre of the bearings gives rise to enormous centrifugal forces as the speed increases, resulting in high pressures on the bearings, vibration of the motor, high friction and heavy wear. In order to avoid this the principle of the de Laval shaft was employed, with a thin flexible spindle instead of the usual rigid shaft. Due to the centrifugal force the spindle will sag already at a low number of revolutions, and this sag becomes greater as the speed of rotation increases. When a certain speed is reached, the so-called critical speed, the sag will theoretically even be infinite. Above that speed the sag decreases rapidly and at the limit for infinitely high speed the disc will rotate about its centre of gravity. When this state is reached the sag of the spindle and consequently the pressure on the bearings is very small. The bearing pressure is then mainly determined by the disc's own weight.

A difficulty in working with a de Laval spindle lies in the passing of the region of the critical frequency when starting up. It is possible to do so without breaking the spindle if that region is passed so quickly as to leave no time for the disc to assume large deflections. In our case, however, the driving couple was not large enough for this and we therefore decided to suppress the dangerously large deviations by applying a suitable damping arrangement to the spindle. For that purpose the spindle is passed through eyelets at a short distance from the disc on either side. These eyelets are connected by rods to small pistons moving up and down with a little play in small cylinders containing oil. By this means the lateral movements of the disc are damped, and by choosing suitable dimensions for this device the vibrations in the critical region can be kept suffficiently low. Once the critical region is passed, the disc runs very quietly and speeds of 1000 and 2000 revs/sec are easily attainable.

As already remarked in passing, the disc is driven electrically. It is placed in the field of two mutually perpendicular magnetic circuits activated by alternating currents with a frequency of 1500 c/sec and shifted 90° in phase with respect to each other. Each circuit consists of two pole shoes, between which air gaps of about $^1/_2$ cm have been cut. The disc is placed in these air gaps. The combination of the two alternating magnetic fields produces a rotating field which turns the disc — made of a conducting material especially for this purpose — and is able to give it sufficient velocity. In order to minimize friction the disc with the complete driving mechanism is placed in an air-tight housing, so that it can function in a vacuum.

The limitation of the signal and its conversion into sound

The current impulses from the photocell, which are of the order of 10^{-7} A, are first very strongly amplified. For this purpose a wide-band amplifier is used which gives amplification constant within 6 db in a region from 30 to 500 000 c/sec. These voltage impulses are modulated, in the first place by fluctuations resulting from contaminations on the sound film, but in addition a noise connected with the powerful amplification is superposed on the whole signal.

As has already been mentioned in discussing the principle of the method, these disturbances are eliminated by limiting the signal. For this purpose a pentode with high anode resistance is used. As is known, by introducing a sufficiently high resistance in the anode circuit of such a valve the I_a--V_g characteristic can be made to assume the shape of the curve k in *fig.* 7 [3]). If, then, we apply

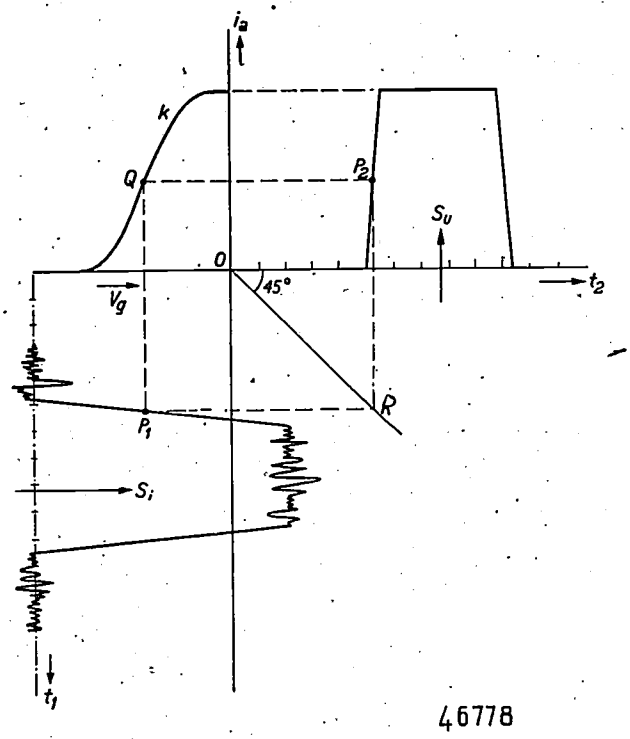


Fig. 7. Diagram of the double limitation of the photocurrent signal by means of a pentode. S_i incoming signal showing disturbances caused by contaminations on the film and disturbances due to the powerful amplification. S_u outgoing signal. k I_a-V_g characteristic of the pentode. Starting from an arbitrary point P_1 of S_u the corresponding point P_2 of S_u can be constructed with the aid of the auxiliary points Q and R. Since the time units on the t_1 and t_2 axes are similar, OR cuts the angle between the t axes through the centre.

[3]) *Cf.* also Philips techn. Rev. 5, 61, 1940.

to the valve a negative grid voltage so high that $I_a = 0$ even for the most powerful disturbances occurring, in the absence of a signal, and make provision for the signal, on the other hand, to be so powerful as always to generate the maximum anode current, likewise for the most powerful disturbances, then the object has been attained (see fig. 7).

Finally the signal prepared in this manner needs only to be sent through a filter that allows the frequencies of the audible region to pass through and eliminates all the others. It may then be fed to the loudspeaker *via* a power pentode.

Conclusion

By means of the method of counteracting noise described here it is possible to obtain a perceptible improvement in quality of the sound reproduced. At the present stage of development the improvement in the case of new films, which are therefore practically free of contamination, is of no significance. In the case of films which have been used several times, however, the improvement is considerable. The method described thus makes it possible to use a film much longer than was previously possible, with retention of the original quality.

A 400 KILOVOLT INSTALLATION FOR X-RAY THERAPY

by W. HONDIUS BOLDINGH and W. J. OOSTERKAMP. 621.386.1 : 615.849

A description is given of a specially constructed therapy installation for 400 kV, 10 mA, built a few years ago for the Academic Hospital in Groningen. In contrast to the construction usually employed for this voltage, in this case the anode of the X-ray tube is earthed. The focus, which is the source of the X-rays, is situated at the end of a long, earthed metal tube which projects from the high-voltage chamber through a partition into the irradiation chamber. In this way absolute protection against the high voltage is ensured for doctor and patient, while the necessary adjustability of the beam of X-rays is provided for by making the X-ray tube movable and rotatable. In this exceptional arrangement the cathode and the focus of the tube are about 1.80 m apart. The measures necessary for the focussing of the electron beam at such a great distance, as well as several other particulars of the installation, are discussed in detail.

Installations for X-ray depth therapy in hospitals are usually for a voltage of about 200 kV, but higher voltages are now no longer exceptions. In fact Philips have already constructed several installations for 400 kV and higher, such as for example the 1000 kV one installed in the Antoni van Leeuwenhoekhuis in Amsterdam, which was described several years ago in this periodical [1]). The medical value of such an installation can by no means be measured simply by the value of the tube voltage applied. The increase in the so-called depth quotient (ratio of the dose at a certain depth under the skin to that on the skin), which was the original reason for raising the voltage, is of course important [1]). But, in addition to a large depth quotient, there are also other properties which are of importance for therapeutic uses, such, for example, as the adjustability of the X-ray tube and thus of the beam of X-rays. It is the wish of the doctor to be able to determine quite freely the direction from which the X-rays fall upon the patient. This is especially important in the application of the methods of cross or rotation irradiation, where the X-ray dose necessary for destroying a tumour in the body is divided into several portions, which are applied from different directions in order to spare the healthy tissue and the skin above the tumour as much as possible, thus with the same object as that which holds in the attempt to attain a large depth quotient. The higher the voltage is chosen, the larger the apparatus and the more difficult it is to provide for the adjustability of the X-ray tube. In the case of the 1000 kV installation in Amsterdam, where the object was not only to attain a large depth quotient, but also to investigate possible biological effects of extremely hard X-rays, the possibility of adjustment of the X-ray tube had to be entirely abandoned and consequently the doctor has at his disposal only a fixed X-ray beam, which emerges from the floor of the irradiation chamber and only the diameter of which can be varied by means of diaphragms. He must place his patient as well as he can with respect to this beam. Cross irradiation is only possible by turning the patient, which meets with objections sometimes of a medical nature (the dropping of internal organs) and sometimes of a practical nature, as it will be difficult to fix the patient in some positions, quite apart from the fact that the necessary positions may be very uncomfortable for the patient.

While in the case of the 1000 kV installation referred to above adjustability of the X-ray tube had to be abandoned entirely [2]), it is still quite difficult to satisfy this requirement also with 400 kV, especially because the patient must in any case be absolutely protected from the high voltage. In the usual tube construction with symmetrical voltage distribution, thus where cathode and anode are each at half voltage with respect to earth, adequate protection can only be realized by surrounding the entire tube with an earthed metal housing. At a tube voltage of 400 kV, however, this is already found to be quite a problem because of the necessary large insulation distances, at least when it is desired to retain easy mobility of the tube [3]). Therefore in the case of tubes for 400 kV it had to suffice, in general, to set up light screens of insu-

[1]) J. H. van der Tuuk, Philips Techn. Rev. 4, 153, 1939.

[2]) In the meantime new constructions have been developed which ensure a certain degree of mobility and adjustability also with a 1000 kV X-ray tube.

[3]) In this and other respects the insulated flexible cables connecting the tube with the high voltage generator constitute a particular difficulty. In some constructions the necessity of these cables is avoided by arranging the tube and generator in one assembly unit, but then the heavier weight of such a unit does not allow of ease of mobility.

lating material to prevent the doctor and the patient from coming dangerously close to the voltage-bearing components.

In the construction of the special 400 kV installation supplied by Philips in 1941 for the Academic Hospital in Groningen, a different method has been followed. The X-ray tube is not symmetrical with respect to voltage distribution, but has an earthed anode. The cathode must then, of course, be brought to a potential of —400 kV with respect to earth, and this, *inter alia*, makes the insulation less simple than when there are two poles at 200 kV with respect to earth, but on the other hand there are the following advantages: 1) the cooling of the anode is very much simplified (this will be discussed later) and 2) the operator, who is only concerned with the anode end of the tube (the actual source of the X-rays) can now be adequately protected against the high voltages while the necessary manoeuvrability of the tube is retained. At the anode end the X-ray tube is extended in the form of an earthed metal tube at the end of which is situated the tungsten lozenge that emits the X-rays. The tube is passed through a partition in the wall of the irradiation chamber in such a way that only the earthed anode tube projects into this chamber, while the cathode end is in the neighbouring room, in which the high-voltage generator is set up. Thus in the irradiation chamber there are no components under high voltage.

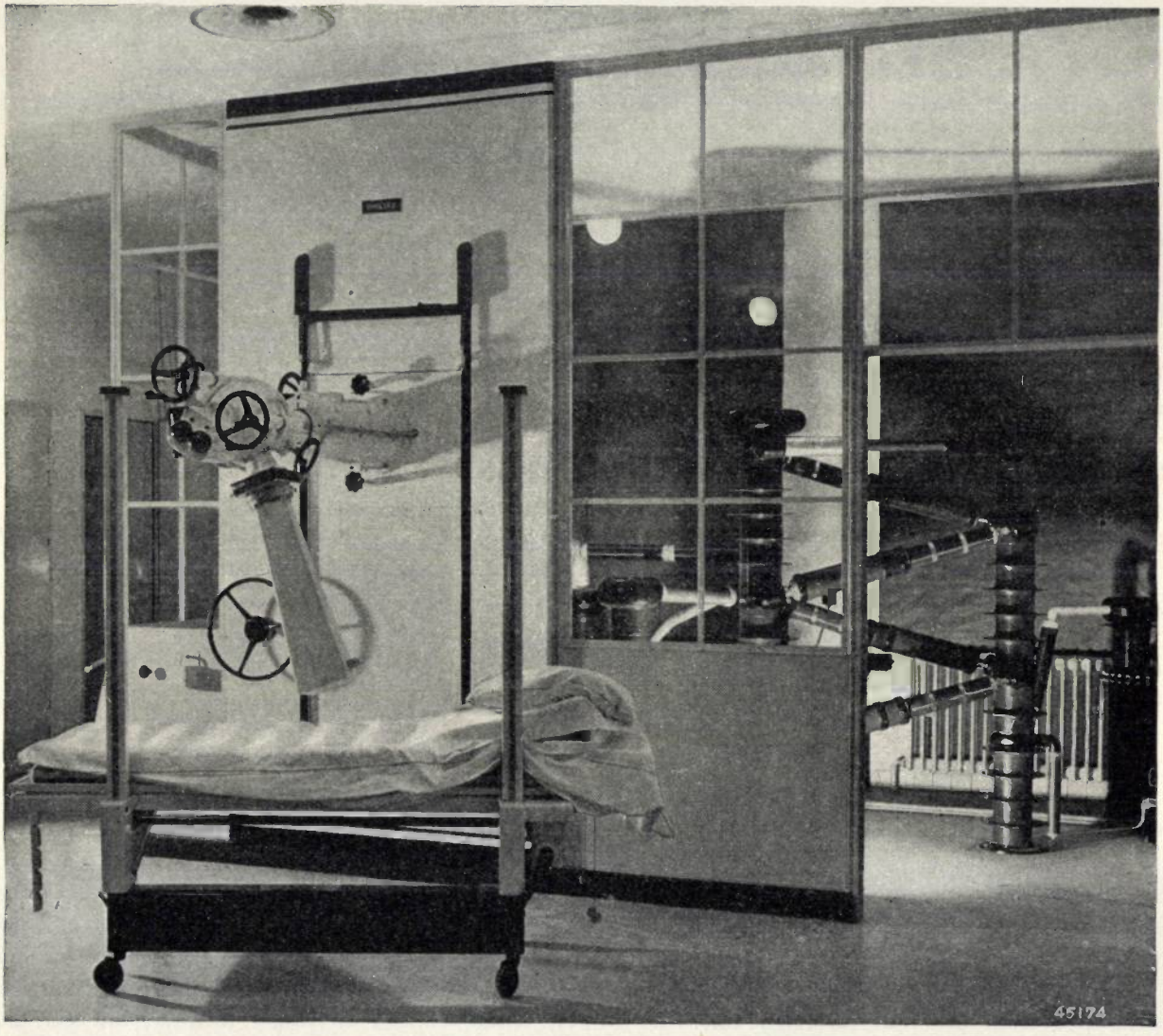


Fig. 1. General view of the therapy installation. On the left may be seen the portion of the X-ray tube projecting into the irradiation chamber, with an irradiation tube affixed to it. In the chamber behind the irradiation chamber is the cascade generator for the excitation of the voltage of 400 kV.

This arrangement is shown in *fig. 1*.

By making the metal partition in the wall in which the tube is placed movable vertically and at the same time providing that the tube can be

rotated around its longitudinal axis, sufficient possibilities of adjustment are obtained to be able to approach a patient on a wheeled bed from all directions. It is only necessary that the focus from which the X-rays are emitted should be sufficiently far away from the wall, at least half the length of the bed.

The remarkable consequence of this is that the cathode and the tungsten lozenge, which in an ordinary 400 kV tube are about 8 cm apart, are here about 1.80 m apart. Special means are therefore necessary to focus the electron beam emitted by the cathode on the lozenge. The focussing takes place in three steps, by means of three "lenses", which can be seen in *fig. 2*. An electrostratic lens (the cathode cap) concentrates the electrons on the entrance to the earthed anode tube. A magnetic lens (magnetic coil) at the beginning of the tube concentrates and directs the beam in such a way that on its passage through it does not touch the walls of the tube. A second magnetic coil 30 cm in front of the lozenge finally serves to give the beam the necessary convergence to form a focus of the desired size.

For the first magnetic coil the direction of the magnetic field is very critical. This coil is therefore mounted on the anode tube by means of a ring and four supporting points adjustable by means of screws, so that the position of the coil may be corrected during the assembly of the X-ray tube.

The significance of the second magnetic coil requires some explanation. The size of the focus is determined by two requirements: it may not be so small that at a given power of the tube the specific loading of the tungsten lozenge and consequently its temperature becomes too high; on the other hand the focus should not be made larger than necessary. The edges of the diaphragm through which the cone of X-rays passes out of the tube cast a penumbra which may make a correct dosage at the edges of the field of irradiation difficult, and the width of this penumbra is proportional to the dimensions of the focus. The desired size of focus is thus fixed fairly accurately, and with it the desired diameter of the electron beam on the lozenge. For this diameter the absolute tolerances are even smaller than for the (apparent) width of focus, since the lozenge is at an unusually small angle (30°) to the axis of the electron beam. It is therefore clear that with the first magnetic coil alone, at a distance of 1.5 m from the lozenge, it would be practically impossible to regulate the beam diameter with the required precision.

For depth therapy a very high X-ray intensity is desired, since for the sake of a large depth quotient the tube is generally placed at a fairly great distance from the patient. Now in order to use a high power in the tube without it being necessary to enlarge the surface of the focus too much, it is important that the specific loading capacity of the lozenge should be high. Owing to the fact that the anode is earthed in the installation in question it was

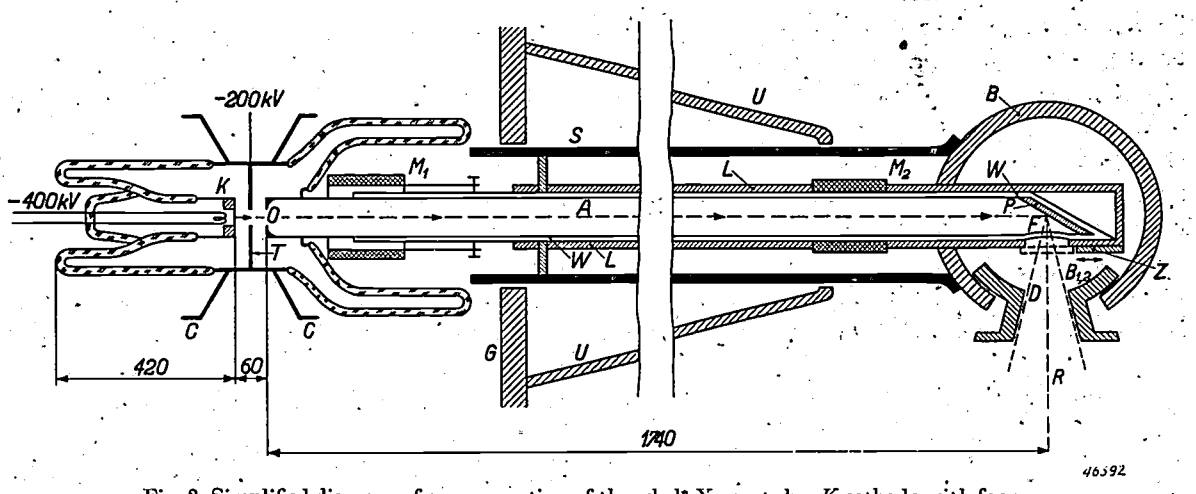


Fig. 2. Simplified diagram of a cross-section of the whole X-ray tube. *K* cathode with focussing cap, *T* partition, *O* opening of the earthed anode tube *A*, M_1 first magnetic coil, M_2 second magnetic coil. *P* tungsten lozenge on which is the focus *F* which emits the X-rays *R*, *W* cooling jacket with running water from the mains, *L* lead jacket around the anode tube, *B* lead sphere, $B_{1,2}$ lead spherical caps with diaphragm openings *D*; *Z* lead slide by means of which the beam can be cut off, *S* steel tube supporting the anode tube and movable in the wall of the room C_1 and the collar *U*; *C* corona gaps. For the sake of clearness the case of one irradiation beam only is drawn. Actually there are two exits for the rays in the two spherical caps, one of which is directed forwards and the other backwards with respect to the plane of the drawing. *Cf.* the photograph.

relatively easy to attain this high loading capacity, because the anode could be provided with water cooling connected to the water mains direct. With a focus of 12 × 6 mm it was thus possible to increase the power converted into heat in the anode to 4 kW (10 mA at 400 kV DC voltage [4]).

The cooling water is conducted not only along the tungsten lozenge, but also along the whole anode tube. In spite of the focussing, 10-20 percent of the primary electrons, as well as the secondary electrons from the lozenge strike the walls of the tube, so that a power of about 1 kW converted into heat has to be dissipated from the anode tube. At the same time the cooling water must keep the magnetic coils cool, since heating of these coils may change the coil resistance and with it the current through the coil and thus the intensity of the magnetic field.

The fact, noted above, that the lozenge is placed at a very small angle to the electron beam is due to the desire to be able to treat two patients simultaneously. In order to be able to irradiate each of the two patients from any direction in spite of the mutual hindrance of the two beds, the axes of the two X-ray cones to be used must be far enough apart and also capable of being directed otherwise than perpendicular to the axis of the tube, while at the same time the X-rays may not be emitted at too small an angle to the surface of the lozenge. The sketches in *fig. 3* show that it is favourable in this repsect to choose the angle of the lozenge to the beam as small as possible. The focus

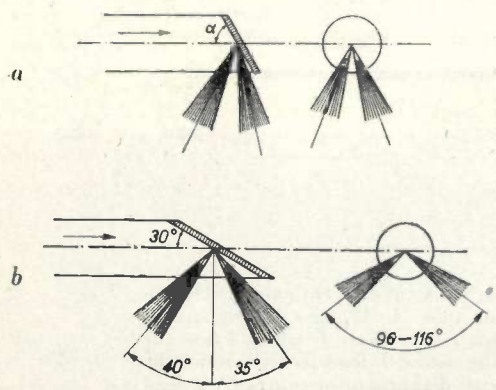


Fig. 3. If the lozenge is at a large angle α to the horizontal electron beam (*a*), the axis of the effective cone of X-rays has only little freedom of motion perpendicular to the beam (right) and still less in forward directions from the perpendicular position (left). When the lozenge is placed at a small angle to the electron beam (*b*) there is much more freedom of movement and two X-ray cones with their axes approximately perpendicular to each other can easily be used at the same time.

[4]) In the construction hitherto customary the maximum load at 400 kV and with the same focus dimensions amounted to only 4 mA. At 400 kV and 10 mA tube current the X-ray intensity with the self-filter of 0.7 mm Cu amounts to 33 r/min at a distance of 1 m from the focus.

on the lozenge emits X-rays in all directions. The two narrow cones of X-rays (maximum angle of divergence 30°) are obtained by providing the end of the anode tube with two spherical lead caps each with a diaphragm opening. In order to direct the two cones of rays, each of the spherical caps can be turned through a certain angle about two mutually perpendicular axes by means of two hand-wheels. A lead sphere fixed rigidly to the tube and enveloping the spherical caps is provided with two openings sufficiently large to permit the passage of the effective cones of rays in all desired directions and prevents outward radiation in undesired directions at any position of the spherical caps. The total thickness of lead in all directions is at least 13 mm, which is sufficient to ensure that everywhere outside the irradiation chamber the so-called tolerance intensity (10^{-5} röntgen per second) is not exceeded. For the adjustment of the X-ray beams with respect to the patients and for the accurate limitation of the field irradiated, an irradiation tube of the desired length, for exemple 1 m, can be fastened over each of the diaphragm apertures. This can be seen in the photograph (fig. 1).

In spite of the adjustability of the two cones of X-rays, the degrees of freedom of the tube mentioned in the beginning of this article (rotation and vertical displacement) are not yet sufficient for the simultaneous treatment of two patients; it must at least be possible to place the two patients at different levels relative to the tube, thus the beds must be adjustable in height. In order to facilitate the placing of the two patients in the desired positions, and particularly to avoid having to move the beds about too much several additional, not strictly necessary degrees of freedom are given to the tube: it can be moved horizontally another 20-30 cm not only perpendicular to but also in the direction of its axis. Vertically the tube can be moved 1.70 m, and the possible angle of rotation abouts its axis is 330°, so that, for example, patients can be treated with ray beams directed upwards from underneath the bed. All these movements, as well as the adjustments of the diaphragms, are effected by turning the various hand-wheels seen in fig. 1.

Besides the end of the anode where the tungsten lozenge is situated the anode tube itself also had to be enveloped in lead along most of its length, for the 15 percent (approx.) of the primary electrons reaching the wall of this tube also excite an intense X-radiation. Here, however, the emitting material is not tungsten, as is that of the lozenge,

but chromium, since the anode tube is made of internally chromium-plated copper. Since the X-ray-emitting property of this metal is much less than that of tungsten (about 30 percent of the emission of tungsten) the lead jacket around the anode tube need only be 6 mm thick. The anode tube with its cooling jacket and lead jacket is supported by a wide steel tube, which also supports the lead sphere and spherical caps on the end of the tube and lends the necessary rigidity to the whole. This steel tube is mounted in the wall partition in such a way as to be rotatable and movable axially.

For the rest the construction of the X-ray tube does not differ very much from what Philips have already supplied for a voltage of 400 kV. Only a few details remain to be mentioned.

In order to obtain the necessary protection against voltage the principle of subdivision of the voltage has been applied [5]): between the cathode and the beginning of the anode tube a metal partition is introduced which has an aperture for the passage of the primary electron beam and is brought to a potential halfway between that of the cathode and the anode. The high voltage for the tube is furnished by a cascade generator with two stages, which can be seen in fig. 1 in the room at the rear. The necessary voltage of 200 kV for the intermediate electrode of the tube can be drawn off directly between the two stages. The cathode, which is at a potential of 400 kV with respect to earth, is heated by two transformers in series, each with insulation for 200 kV between the primary and secondary windings.

In X-ray tubes for such voltages the various high voltage lead-in points are generally enveloped in large, well rounded metal caps, in order to keep the electric field strength in the space around the tube sufficiently low: the field is "homogenized", *i.e.* an attempt is made to approach as nearly as possible the state of a homogeneous field between two plane parallel electrodes, where the field is smaller than in any other configuration with the same distances. In our case, however, the introduction of large metal caps was undesirable because it would have increased the weight of the already very heavy tube. Therefore the homogenization was effected with the help of a corona gap, *i.e.* toothed metal discs. The high electrical field strength at the teeth of such a disc causes a corona discharge to take place there, which is accompanied by ionization of the air. As a result the air in the vicinity of the disc edge becomes to a certain extent conductive, and this conducting layer of air has the same action as if a metal conductor were provided there.

In *fig. 4* the corona gap on the cathode supply connection of the tube can clearly be seen. The connection of the cathode with the 400 kV terminal of the high-voltage generator must allow for the necessary freedom of motion of the tube in the vertical and in the two horizontal directions, as well as for rotation. A flexible metal tube is used for this purpose. A cap in the form of a long upright open cylinder, which is visible in the centre of fig. 4, provides that the field at the 400 kV terminal remains sufficiently homogenized at all positions of the tube.

The X-ray tube can, if desired, also be used at a lower voltage than 400 kV. For this purpose the initial voltage of the cascade generator has been made adjustable in a large number of steps with the help of an auto-transformer. When the tube voltage is varied, however, the currents of the two magnetic coils serving for the focussing of the electron beam in the anode tube must also be varied. The diffractive action of the magnetic field on the electrons is inversely proportional to their velocity and thus inversely proportional to the square root of the tube voltage (when relativistic corrections are disregarded). The corresponding regulation of the coil currents takes place automatically, since the supply rectifier is coupled with the voltage regulator. The voltage regulator is placed on the operation desk of the installation, where the tube voltage, measured by the current in a resistance in parallel with the tube, can be read off directly on a meter. At the lower voltages the current through the tube can be increased to a maximum of 20 mA, so that the tube can also be fully loaded at a voltage of 200 kV.

On the operation desk there are various auxiliary instruments, such as the indicator of the dosagemeter, with which every irradiation is checked. The irradiation of each of the two patients can be begun and ended independently by opening or closing the corresponding "door" of the lead spheres around the lozenge. For this purpose both doors are fitted with two 13 mm thick lead slides which can be operated from the operation desk by two small servo motors mounted at the end of the tube. The position of these slides, open or closed, is indicated by signal lamps. Further there are two series of six signal lamps corresponding to six different ray filters which can be slid in front of each of the two doors. In this way the doctor or

[5]) For a detailed discussion of this now generally accepted principle see the article referred to in footnote [1]).

his assistant who is in charge of the irradiation treatment, and who may under no circumstances enter the irradiation chamber during a treatment because of the danger from rays scattered by the patient, can nevertheless always see whether the correct filter has been inserted and whether the slides have been opened or closed, as the case may be.

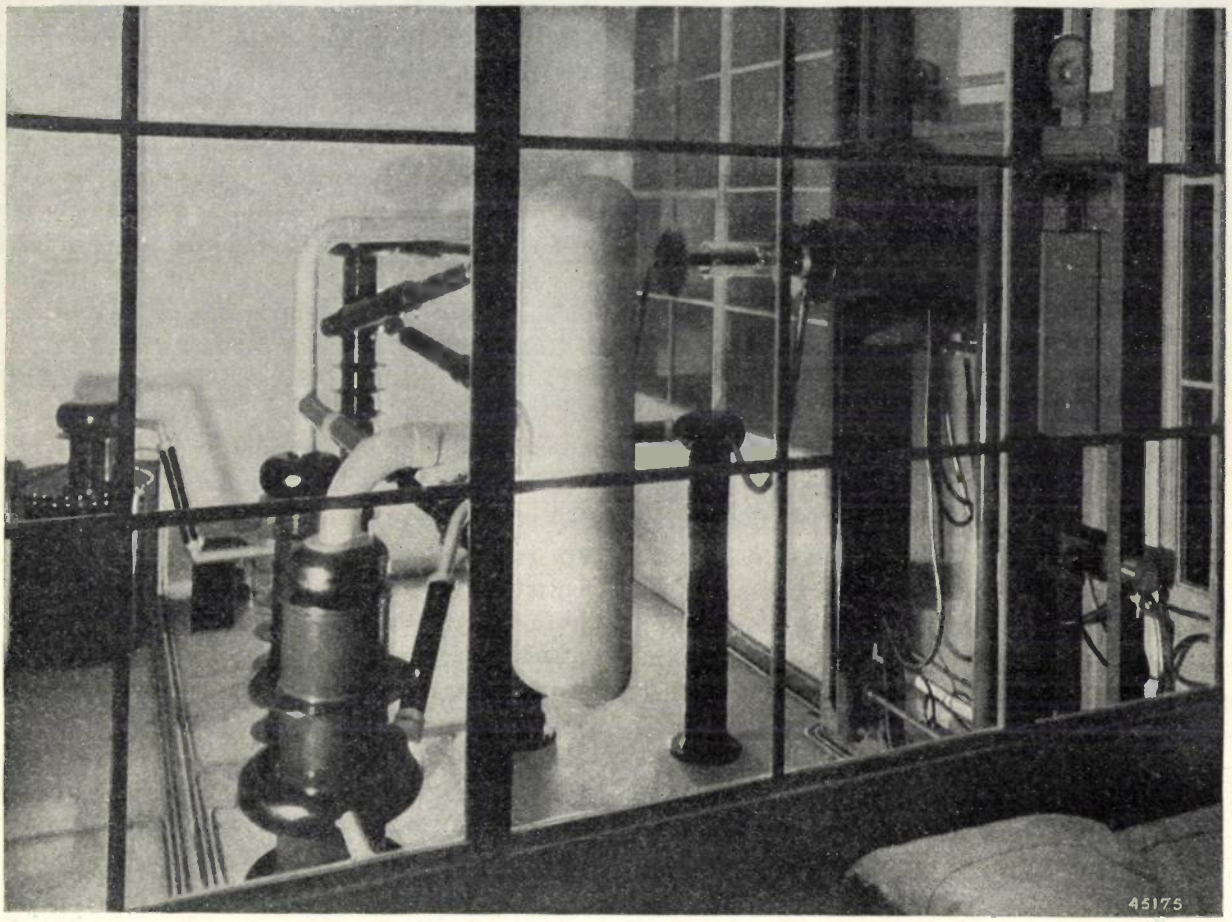


Fig. 4. View of the high-voltage chamber. On the right the end of the X-ray tube projecting through the wall. The voltage supply lead for the cathode can be seen with a corona gap, and in the middle the large hollow electrode cylinder cap for keeping the field homogeneous when the tube is moved horizontally or vertically. To the right of this may also be seen the voltage lead for the intermediate electrode of the tube. To the left in the foreground the transformers for heating the cathode. In the background the cascade generator.

RADIO INVESTIGATION OF THE IONOSPHERE

by C. J. BAKKER.

621.391.11 : 551.510.535

This article gives a survey of investigations made concerning the layers of ionized air (ionosphere) present in the atmosphere high above the earth which are able to reflect radio waves back to the earth, thus making radio reception possible at a great distance beyond the horizon of the transmitter. Various phenomena in shortwave reception, such as fading and the skip distance, are due to the presence of the ionosphere. The density of ions and the so-called "effective" altitude of the ionosphere can be derived from measurements that have been taken. It has been found that particularly the density of ions, and also the reception of short waves, depends to a large extent upon the sun, namely upon its height above the horizon and its eruptive activity. The main, but not the only cause of the ionization is in fact the irradiation of the atmosphere by sunlight, in particular by the part of the spectrum in the far ultra-violet. In the process of ionization this part of the sun's spectrum is absorbed and cannot, therefore, be observed on the earth. Interesting conclusions as to the intensity of the sun's radiation in the far ultra-violet can be drawn from data concerning the ionosphere obtained experimentally by means of reflected radio waves.

After the achievement of the first transatlantic wireless connection by Marconi in 1901, a discussion quickly arose as to how it was possible that the eletromagnetic waves sent out could reach the receiving station around the curved surface of the earth. If the propagation of the waves were rectilinear, signals from a transmitter situated beyond the horizon could never be received, so that it had to be assumed that the path of the waves was not a straight line but a curved or a broken line; both hypotheses found supporters.

One year after Marconi's experiments Kennelly and Heaviside, independently of each other, proposed the hypothesis that the high, very rarefied layers of air behave as a mirror for electrical waves and reflect the radio signals back to the earth. The propagation of the waves could then take place along broken lines. The reflective power of these higher layers of air might be caused by an ionization of the air, *i.e.* a splitting of the molecules into electrons and positive ions. Kennelly estimated the lower boundary of the ionization region, the so-called ionosphere, to be at an altitude of 80 km, which, as we now know, is fairly correct, at least as far as the order of magnitude is concerned.

In the meantime attempts had also been made to explain the possibility of reception below the horizon without the hypothesis of a reflecting layer. Various theorists (among whom Rayleigh, Poincaré, later Watson, Sommerfeld, Bremmer and van der Pol[1])) showed that the electromagnetic waves are bent along the conducting earth's surface, so that, especially in the case of long waves, reception far below the horizon is possible.

In the course of further investigations it has been found that this theory of refraction must not be considered as an argument for making the hypothesis of the ionosphere reflections superfluous, but that the two theories supplement each other excellently. In radio communications with waves shorter than 10 m the ionosphere plays no rôle. In the very important wave-length region from 10 to 50 m, on the other hand, the bridging of long distances with relatively slight energy would be impossible without the reflection of the waves against the ionosphere.

Until about 1920 only wave lengths above 200 m were used. After the possibilities of the wave-length region below 200 m had been proved by the bridging of the Atlantic with relatively low power transmitters at about that time, shorter and shorter waves began to be used. Several new phenomena were then encountered which did not occur on long waves or only sporadically, and the cause of these was thought to lie in the presence of the ionosphere. It was found, for example, that reception was generally stronger at night than in the daytime. Another phenomenon was that radio *H*-location on short waves is subject to larger errors at night than in the daytime. To these phenomena could be added that of fading and the fact that with increasing distance reception from an ultra-short wave transmitter does not decrease in intensity continuously, but after having passed a "dead zone" (skip distance) may abruptly attain a greater intensity again. Another interesting phenomenon, and one which is of technical importance, is that the northern lights, the magnetic storms

[1]) See: Philips techn. Rev. 4, 245, 1939.

(*i.e.* great changes in the magnitude and direction of the magnetic field of the earth) and certain kinds of fading are found to be intimately related to each other. Thus, for example, in 1928 it was at first impossible to make wireless contact with the Nobile expedition which had come to grief near Spitsbergen, since the attempts to do so happened to coincide with magnetic storms in the northern lights zone.

The phenomena mentioned have been the subject of profound investigations carried out in many countries, and these have now reached a stage where not only has a fairly complete technical insight been obtained into the causes and the mechanism of many of the phenomena referred to, but also various new fields of geophysical, meteorological and astrophysical investigation have been opened.

Refraction and reflection of radio waves by the ionosphere

A radio wave is an electromagnetic wave which differs from a light wave only in its wave length. It may therefore be expected that the refraction and reflection of radio waves by the ionosphere will in many respects follow the known laws of optics. We shall therefore try to characterize the properties of the inosphere in the manner customary in optics, by assigning to every point in space a definite index of refraction. What do we mean by this index of refraction? For the case when dissipation can be disregarded it can be derived in a simple way.

According to a familiar relation of Maxwell the index of refraction of a substance n is related to the dielectric constant E by the equation:

$$n^2 = \varepsilon \quad \ldots \ldots \ldots \quad (1)$$

We can therefore determine the index of refraction by studying how a gas of the nature of the ionosphere behaves as a dielectric. In doing so we shall consider the ionosphere for the present as a homogeneous gas which in addition to the molecules contains N electrons and N positive ions per cm^3.

Suppose that in such a gas electrical vibrations are excited with an angular frequency ω and an amplitude E_0. This electrical alternating field in a dielectric gives rise to electric currents whose density we shall call j. The dielectric constant is then defined by

$$\varepsilon = j/j_1,$$

where j_1 represents the displacement current for a vacuum.

The electric current j consists in general of two components, the first of which is the above-mentioned displacement current for a vacuum. This is given by:

$$j_1 = \frac{1}{4\pi} \cdot \frac{\mathrm{d}E}{\mathrm{d}t} \quad \ldots \ldots \quad (2)$$

The second component j_2 is a convection current, to be ascribed to an actual displacement of charges. If N is the number of charged particles per cm^3, e the charge and v the velocity, then

$$j_2 = N e v$$

and the dielectric constant is thus:

$$\varepsilon = (j_1 + j_2)/j_1 = 1 + \frac{4\pi Nev}{\mathrm{d}E/\mathrm{d}t} \quad \ldots \quad (3)$$

For free electrons such as are present in the ionosphere the velocity v can easily be calculated [2]). The acceleration of particles with charge e and mass m is determined by the equation

$$m \frac{\mathrm{d}v}{\mathrm{d}t} = eE. \quad \ldots \ldots \quad (4)$$

When E changes sinusoidally with an angular frequency ω, then the same is true for v, *i.e.*

$$\frac{\mathrm{d}^2v}{\mathrm{d}t^2} = -\omega^2 v.$$

From this and equation (4), after differentiating t, it then follows that:

$$v = -\frac{e\,\mathrm{d}E/\mathrm{d}t}{m\omega^2};$$

and by substituting this in (3) one obtains:

$$n^2 = \varepsilon = 1 - \frac{4\pi Ne^2}{m\omega^2} \quad \ldots \ldots \quad (5)$$

The apparent dielectric constant is thus less than unity, because the convection current is in opposite phase to the displacement current, and the same is true of the index of refraction. Thus, speaking in optical language, the ionosphere is a less dense medium than empty space.

Now it is known that a wave which passes from an optically dense medium into an optically less dense medium is not only refracted but may also be totally reflected. If n is the index of refraction of the less dense medium, n' that of the denser medium, the condition for total reflection is $n \leqslant n' \sin a$, where a is the angle of incidence on the

[2]) In addition to free electrons the ionosphere contains an equally large number of positive ions. Since, however, these possess a mass about 50 000 times as large, they are accelerated to a much smaller extent by electric fields, so that in practice we need only consider as mobile particles the free electrons.

surface of reflection. In our case $n' = 1$. If we now first consider the case where $a = 0$ (signal directed vertically upward), the condition is then $n < 0$, or, according to (5),

$$\frac{4\pi\ Ne^2}{m\omega^2} \geqslant 1. \quad \ldots\ldots\ldots \quad (6)$$

For sufficiently low values of ω this condition is always satisfied; low-frequency signals are thus totally reflected by the ionosphere. Above a certain frequency, however, this is no longer the case; for the critical frequency, according to equation (6), the following holds:

$$\frac{4\pi\ Ne^2}{m\omega^2_{\text{crit}}} = 1, \quad \text{or} \quad f^2_{\text{crit}} = \frac{Ne^2}{\pi m}. \quad \ldots \quad (7)$$

The critical frequency f_{crit} is thus a measure of the electron density N in the ionosphere.

The signals directed vertically upwards for which we have calculated the critical frequency are of great importance in the study of the ionosphere, as will be discussed below. In practical radio transmission, however, one is usually concerned with waves directed obliquely upwards. The chance of total reflection is thereby increased. If a is the angle of the wave to the vertical then $n < 0 \sin a$ holds for total reflection, or according to equation (5)

$$1 - \frac{4\pi\ Ne^2}{m\omega^2} \leqslant \sin^2 a = 1 - \cos^2 a. \quad \ldots \quad (8)$$

If we set $a = 0$ we again obtain equation (6) for the critical frequency. If, conversely, we choose any arbitrary frequency f greater than f_{crit}, equation (8) then gives us a critical angle of reflection:

$$\cos^2 a_{\text{crit}} = \frac{4\pi\ Ne^2}{m\omega^2} = \frac{Ne^2}{\pi m f^2} = \frac{f^2_{\text{crit}}}{f^2}. \quad \ldots \quad (9)$$

Reflection only occurs for $a > a_{\text{crit}}$, while waves with an angle of incidence $a < a_{\text{crit}}$ pass through the ionosphere. Practically this amounts to the fact that only at a certain minimum distance from the transmitter can the reflected signals be observed.

A simplified representation of the situation is illustrated by *fig. 1a*, where it is assumed that the altitude of the ionosphere is so small in relation to the radius of the earth that the earth may be considered flat, while, moreover, the reflection from the ionosphere takes place in the form of a sharp angle.

The fact that the reflected radiation only becomes observable at a definite distance r_1 is familiar in practice in the existence of a "dead zone" [3]) (skip distance). If one is concerned with case a in fig. 1 the radius of the skip distance can easily be calculated, and one finds:

$$r^2_1 = 4h^2 \left| \frac{\pi f^2 m}{Ne^2} - 1 \right|,$$

where h is the altitude of that layer of the ionosphere from which the radiation is reflected. Since this height may amount to several hundred kilometers, however, in many cases the assumption of a flat plane as the earth's surface is not sufficiently accurate, and consequently the curvature of the earth has to be taken into account. Furthermore, the angle of reflection is not so sharp, since the boundaries of the ionosphere are more or less vague and the deflection is therefore more gradual. Taking both these factors into account, the picture represented in fig. 1*b* is obtained. The region commanded by the radiation reflected at the ionosphere then has not only an inner boundary r_1 (skip distance) but also an outer boundary r_2, namely where the reflected ray still just touches the earth. Under favourable conditions this distance amounts to about 1700 km for signals reflected by a layer at an altitude of 100 km and about 3600 km for a layer at 300 km. Transmission of short-wave signals over greater distances by single reflection at the ionosphere is therefore impossible. With multiple re-

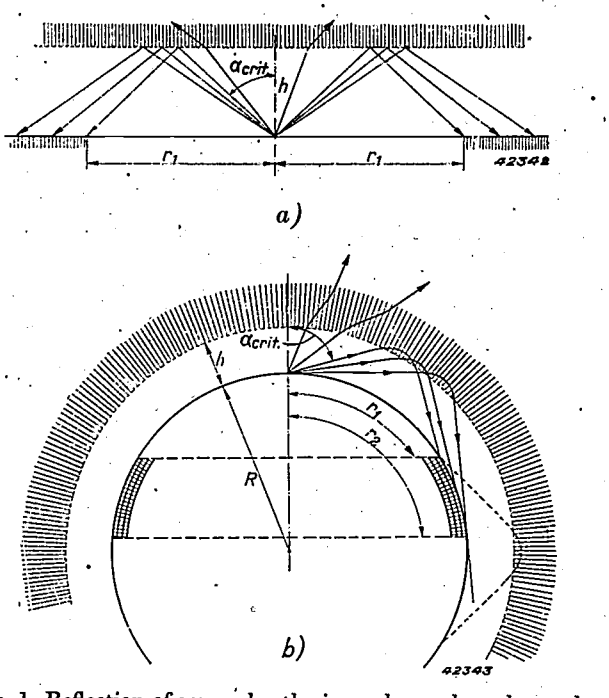


Fig. 1. Reflection of waves by the ionosphere when the earth is considered flat and reflection sharp (*a*) and when the earth is curved and reflection more gradual (*b*). Rays with an angle smaller than a_{crit} are not reflected but pass through the ionosphere. The region commanded by the once-reflected waves is limited by a minimum radius r_1 (skip distance) and in the case of the curved surface of the earth also by a maximum radius r_2.

[3]) Within the dead zone signals can sometimes be received at irregular times; these are, however, unsuitable for regular communication. The reflection in these cases probably takes place from irregularly occurring electron clouds in or between the normal layers of the ionosphere.

flection at the earth and the ionosphere greater distances can of course be bridged.

The higher the frequency is chosen, the narrower the annular region commanded by the reflected waves, since the radius r_1 of the skip distance zone becomes steadily larger. For the limiting frequency at which it shrinks to nothing one finds about 30 megacycles/sec ($\lambda = 10$ m), depending on the activity of the sun, while f_{crit} lies at about 3 megacycles/sec ($\lambda = 100$ m). In the following we shall go more deeply into the numerical data and also discuss what these empirical values are able to show us about the altitude and condition of the ionosphere.

Altitude and electron density of the ionosphere

Experimental investigation of the ionosphere is being carried out in very many countries, so that in spite of the fact that it has been in progress only some 20 or 30 years there is already considerable empirical material available.

The data of most interest are those concerning the altitude of the ionosphere and the degree of ionization.

Several methods of determining the altitude are in use. A method originally employed by Appleton makes use of interferences between the so-called carrier wave and the radiation reflected by the ionosphere (see *fig. 2*). The distance between transmitter and receiver chosen is such that the intensity of the carrier wave received is about equal to that of the reflected wave. The two waves will amplify each other when $2\ (a—a') = m\lambda$ and attenuate each other for $2\ (a—a') = (m + {}^1/_2)\ \lambda$ (m being a whole number and λ the wave length).

If the wave length is decreased continuously from λ to λ', amplified and attenuated signals are received alternately. The number of times that an amplified signal is received is:

$$\mu = 2(a—a')/\lambda' - 2(a—a')/\lambda.$$

By determining μ, a, λ and λ', a' and thus also h' can be calculated. This "effective altitude" h' is greater than the actual altitude of the point where the reflection takes place, since the velocity with which the signal is propagated in the ionosphere is less than the veolcity c of light in vacuum.

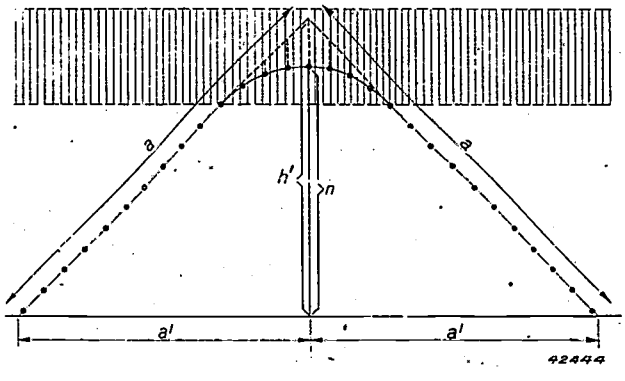


Fig. 2. Determination of the effective altitude h' of the ionosphere from interferences resulting from the difference in path $a—a'$. It may be seen that the actual altitude is smaller since the waves do not experience a sharp reflection but are gradually bent. The dots on the line indicating the path of the electric wave represent time intervals and show that the rounding-off of the angle results in no decrease in the transit time, since in the ionosphere the signal is propagated more slowly.

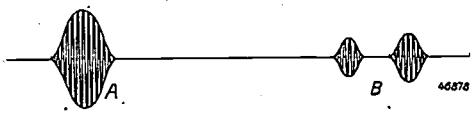


Fig. 3. The transmitted signal A and the reflected signals B in the determination of the effective altitude of the ionosphere according to the method of Breit and Tuve.

The phenomenon underlying Appleton's method, *viz.* interference of the waves that have travelled different paths from the transmitter to the receiver, is manifested in normal radio reception as the familiar fading.

Interpretation of the results of the measurements by the Appleton method is rather difficult. Therefore a different method is now usually empoyed, the "echo method" developed by Breit and Tuve, upon which, as a matter of fact, the now universally known "radar" is based.

By means of a transmitter short wave trains are sent out directly upwards at intervals of, for instance, 1/50 sec (see *fig. 3*). A receiver not far from the transmitter records the carrier wave and immediately afterwards the echoes due to reflection from the ionosphere. From the time t_e elapsing between the reception of the carrier wave and that of the reflected wave, (echo time), the effective altitude h' of the ionosphere can be calculated by the relation $h' = {}^1/_2\ ct_e$.

Some results obtained by the Breit and Tuve echo method are reproduced in *fig. 4*. The registrogram shows the echo time of the reflected signal as a function of the frequency. It is found that up to about 4 megacycles/sec ($\lambda = 75$ m) an echo time of 60 microseconds occurs, corresponding to an apparent altitude of 100 km. At frequencies above 3.5 megacycles/sec the echo time begins to increase gradually; apparently waves of these frequencies penetrate deeper into the ionosphere. The critical frequency is passed at 4 megacycles/sec. The signal which penetrates through the lowest layer of the ionosphere above this limit does not, however, disappear into space but is reflected from a higher

layer ($h \approx 200$ km). This so-called F_1 layer is much less sharply bounded than the underlying layer, which is indicated as the E layer. This is evident in the diagram from the fact that the altitude at which reflection occurs increases sharply

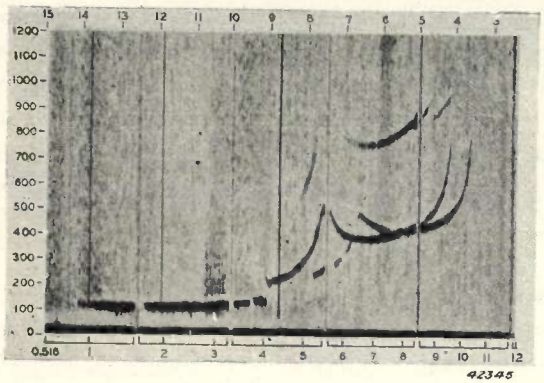


Fig. 4. Recording of the effective altitude of the ionosphere in km as a function of the frequency in megacycles/sec. when the sun is high. Registrogram of the Carnegie Insitution Washington; taken from Darrow, Bell Syst. techn. J. **19**, 455, 1940.

with the frequency. At a frequency of 5.5 megacycles/sec the signal penetrates also through the F_1 layer, reflection then occurring at the so-called F_2 layer, which has an altitude of 300 km. Not until a frequency of 10 megacycles/sec is reached are the waves able to penetrate through this layer, when the signal disappears into space.

In the frequency region in which the F_2 layer is effective the diagram shows a second echo, which can be ascribed to waves which have covered the distance between earth and the ionosphere and back twice. Furthermore a remarkable doubling of the diagram lines is visible. This latter effect is connected with the terrestrial magnetic field and will be discussed briefly farther on in this article.

The critical frequencies found for the reflections at the various layers make it possible to calculate the electron density N at the point where the waves are reflected. According to equation (7) this quantity is:

$$N = \frac{\pi m f^2_{\text{crit}}}{e^2} = 1.24 \cdot 10^{-8} f^2_{\text{crit}} \qquad (10)$$

By the combination of determinations of N with measurements of h' a fairly complete picture can be obtained of the distribution of the electrons over the different layers of the ionosphere.

Results of the investigation

Apart from systematic variations with the season of the year and the position of the sun, the state of the ionosphere is found to exhibit rather great fluctuations. Electron densities occur which are 10 times as great as the average. In general the following conditions can be determined.

At an altitude of about 50 km in the daytime there is a slightly ionized layer ("ozone" or D layer). At night the electron density in the D layer is very slight. The second layer (E layer) is permanent and lies at about 120 km altitude. Above that at 200 and 300 km altitude lie at least two other layers (F_1 and F_2) more or less distinct in the daytime but immediately after sunset merged into one, the F layer; sometimes in the daytime in winter too the separation between the F_1 and F_2 layers cannot be observed. Between the E and F_1 layers there occurs at very irregular times another layer of limited extent, which is sometimes called the E_2 layer (American: "sporadic E layer"); in the summer this is often observed in the morning and in the evening; sometimes this E_2 layer occurs in the evening at all seasons.

The following table gives a survey of the average condition of the ionosphere at noon on a summer day for a mean latitude on the earth:

layer	effective altitude (km)	critical frequency (megacycles/sec)	maximum electron density (number of particles/cm³)
D	50	<0.4	$<2.5 \times 10^3$
E	120	2.5	10^5
F_1	200	5.0	4×10^5
F_2	300	8.0	10^5

For the sake of comparison it may be noted that from the radio investigation of the ionosphere it appears that the molecular density of the atmosphere at the altitude of maximum ionization of the E layer is of the order of magnitude of 10^{12} molecules per cm³, while in the F layer the figure is about 10^{11}.

The question has arisen as to whether there are still other layers present. It is indeed very well possible that we are unable to observe all the layers; a layer can only manifest itself when its ionization is greater than that of all the underlying layers, so that its limiting frequency is higher than the lowest frequency of the waves penetrating through the lower layers. Faint indications of layers other than those mentioned have sometimes been found. It is possible, however, that deceptive phenomena may occur which would appear to indicate the existence of an ionized layer although such a layer does not actually exist. In this connection it is interesting to note that in 1927 Hals received echoes whose echo-time was of the order of magnitude of 10 to 30 sec, which would point to a reflec-

ting layer far beyond the orbit of the moon, where according to Störmer a sphere of charged particles can indeed be expected. Whether this is the correct explanation of the origin of the echoes with long transit time may be open to doubt.

The reason for the complex character of the ionosphere has not yet been determined with certainty. This phenomenon is probably connected with the absorption of different parts of the sun's spectrum at different altitudes in our atmosphere. The upper layers of the ionosphere are probably ionized not only by light but also by charged particles originating in the sun.

Which of the gases of our atmosphere are subject to this ionization is also uncertain. It might be thought that our atmosphere at 200 to 400 km altitude no longer consists mainly of nitrogen, but of lighter gases, for instance hydrogen and helium. A direct indication of the presence of these gases has not, however, been found, although they have often been sought, for example in the spectrum of the northern lights, which occur at the altitudes in question.

One of the bands of the northern lights spectrum could be ascribed to the positive ion of the nitrogen molecule, so that the occurrence of the reaction $N_2 \rightleftarrows N_2^+ + e$ is thus proved. This process is probably one of the causes of the ionization in the F layer.

Daily, annual and other periodical variations

The connection between the position of the sun and the state of the ionosphere has already been referred to in the foregoing. This connection makes it probable that the ionization is for a large part brought about by the ultra-violet radiation of the sun. For the E layer this supposition can even be confirmed quantitatively, by calculating the electron density as a function of the position of the sun on the basis of the laws of molecular equilibrium.

The following formula holds for the change in electron density per second:

$$\frac{dN}{dt} = q - aN^2.$$

In this expression q is the number of electrons formed in the ionosphere per second per cm^3 and aN^2 the number which per second and per cm^3 reunite with the positive ions; a is the so-called recombination coefficient. The fact that the velocity of the recombination is determined by the square of N is clear when remembering that the chance of recombination is proportional to the product of the concentrations of the electrons and of the positive ions. If an equilibrium condition is assumed, *i.e.* $dN/dt = 0$, the following applies:

$$N = \sqrt{q/a}.$$

The quantity q depends upon the height of the sun, which can be expressed by the angle χ made by the sun's rays with the perpendicular. If one considers, for instance, a horizontal plane in the ionosphere of 1 cm^2, it is clear that the amount of ultraviolet energy striking that plane is proportional to $\cos\chi$. Thus N is proportional to $(\cos\chi)^{1/2}$, and according to (10) this means that

$$f_{crit} \text{ is proportional to } (\cos\chi)^{1/4}. \quad . \quad . \quad (11)$$

This relation is reproduced in *fig. 5* for the sake of comparison with the results of measurements. It is found that the formula for the E layer corresponds to the observations carried out during a large part of the day. From this it may be concluded that the processes of ionization and recombination in the E layer in the daytime are indeed in equilibrium with each other.

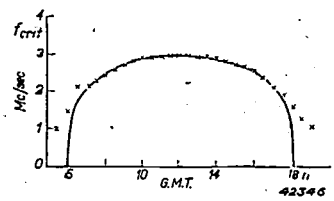


Fig. 5. Daily variation of the critical frequency (in megacycles/sec.) for the E- layer (average from March 18 to 23 incl. 1935; taken from Appleton, Proc. Roy. Soc. 162, 451, 1937). The line drawn represents the variation of the quantity $(\cos\chi)^{1/4}$, where χ is the zenith distance of the sun. Except in the hours of night the curve corresponds very well to the variation of the critical frequency. G.M.T. = Greenwich Mean Time.

Formula (11) does not meet the case so well for the variation for the F_1 layer, and it is certainly not valid for the F_2 layer. In these layers therefore the equilibrium between ionization and recombination is absent, which is understandable considering that in the more rarefied atmosphere the processes take place much more slowly, because of the smaller number of collisions. However, it may also be assumed in explanation that the ionization is determined by phenomena other than the ultraviolet rays. The two causes probably work together.

The other known periodical variations in the state of the ionosphere are likewise connected with the sun. In the course of a year the altitudes and especially the electron densities vary with the seasons (see *fig. 6*). The electron concentration in the E layer and especially in the F_1 layer is greatest in summer, which is understandable after what has been said about the influence of the height

of the sun. The F_2 layer behaves differently, the limiting frequency and thus also the electron density reaching a greater value in winter than in summer.

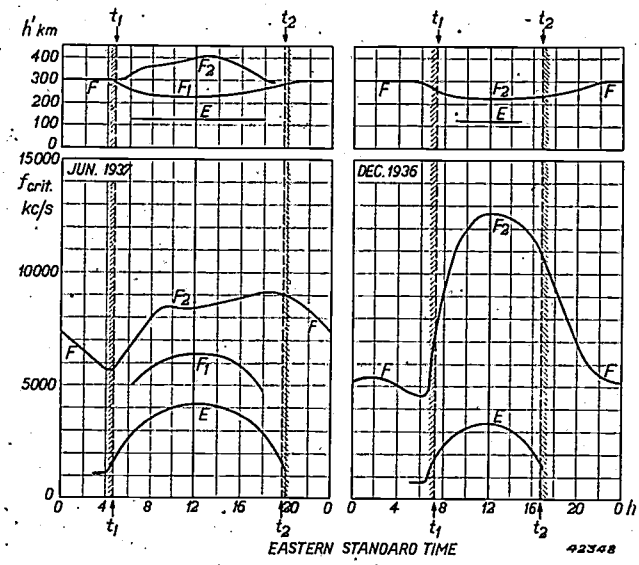


Fig. 6. Daily and seasonal variations of the critical frequency f_{crit} in kc/sec and of the effective altitude h' for the different layers of the ionosphere (borrowed from Smith, Gilliland and Kirby, J. Res. Bur. Stand. 21, 835, 1938): t_1 sunrise; t_2 sunset. Eastern standard time, 75° west longitude (practically the meridian of Philadelphia).

From this it is sometimes concluded that the F_2 layer is sensitive to temperature, becoming diffuse and rising in the summer whereas in winter it becomes denser and falls.

On the basis of his own experiments and those of his collaborators, Elias presumed that the nightly ionization at 350-400 km effective altitude can indeed be distinguished from that in the daytime and that it is to be ascribed to corpuscular radiation, or perhaps partly to eletromagnetic radiation of cosmic origin.

The eleven-year period in the number of sunspots observed every year seems to correspond to an equal period in the electron densities of the ionosphere. According to Smith, Gilliland and Kirby, during the years 1932-1937 the electron densities of the E, F_1 and F_2 layers increased with the number of sunspots (see *fig.* 7).

Further of interest are the changes in the ionosphere during an eclipse of the sun. It has been found that during an eclipse the electron density in the E and F_1 layers decreases sharply, but almost immediately after the eclipse returns to the normal value again. The electron densities of these layers are found to be roughly proportional to the non-eclipsed portion of the sun's disc; that of the F_2 layer behaves differently and varies in a smaller degree. This may be ascribed partly to the recombination coefficient being too low, but on the other hand it seems to confirm the suspicion that the ionization of the F_2 layer is caused partly by a corpuscular radiation.

Finally in 1939 Appleton and Weekes discovered lunar times in the ionization state of the ionosphere. This influence of the moon, however, is very slight.

Disturbances in the ionosphere (fade-outs)

In the year 1935 Dellinger published his detailed investigation of an important discovery by Jouast, Mögel *et al.* At certain moments it is suddenly impossible to receive any shortwave signals passing over the part of earth that is in daylight. This effect occurs so abruptly that it is often thought that something is wrong with the receiving set. The disturbance lasts from about ten minutes to an hour or longer. Dellinger pointed out the simultaneous occurrence of these fade-outs over the whole daylight half of the earth and showed that there was a connection with eruptions in the chromosphere of the sun. The explanation of the fade-outs is to be sought in the fact that when there is an eruption a large amount of ultra-violet radiation is emitted by the erupting spot on the sun, which causes intense ionization in the earth's atmosphere at an altitude of 50 km (D layer), especially through the a line of the Lyman series with a wave length of 1216 Å. As a result of the relatively high pressure at these altitudes the electrons set in vibration in the field of a radio wave will undergo continual collisions with gas molecules and lose their energy; in consequence the absorption of the radio signals is so great that they disappear and radio communication becomes impossible.

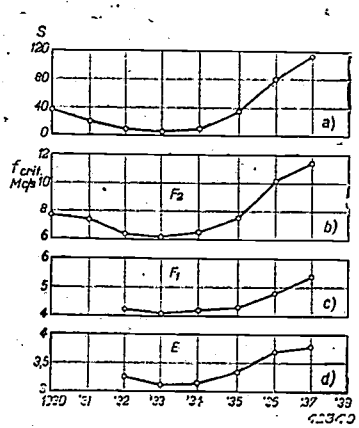


Fig. 7. Comparison of the annual average of the number of sunspots s with the noon values of the critical frequency f_{crit} (in megacycles/sec) for the different layers F_2, F_1 and E in the corresponding years (from Smith, Gilliland and Kirby, J. Res. Bur. Stand. 21, 835, 1938).

There is some indication that the reception strength of long waves increases during a fade-out, instead of decreasing. Since these waves are reflected by the *D* layer at about 50 km altitude, this would agree with the explanation that fade-out is to be ascribed to an increase of ionization in the *D*-layer.

Disturbances due to *D* layer absorption are also known which do not occur and disappear so suddenly as fade-outs, and which last for a longer time, usually several hours. The decrease in intensity of the signals, too, is usually not so great as in the case of fade-outs.

Disturbances from another cause occur during magnetic storms and upon the appearance of the aurora borealis, which again is found to be connected with the eleven-year period of the sunspots. During the occurrence of magnetic storms and the aurora borealis material particles from the sun penetrate into the ionosphere and there disturb the normal state. Investigation has shown that at a high geographical latitude the electron density of the *E* layer then increases sharply. At lower latitudes the *F* layer seems to be primarily affected; the electron density decreases and the effective altitude rises, thus indicating a diffusion of the layer.

Practical consequence for radio reception

From what has been said in the foregoing a picture can be formed of the influence of the ionosphere in the different frequency regions. In the case of long-wave daytime broadcasting (kilometre waves) the wave is, as it were, enclosed between the *D* layer and the surface of the earth. Owing to the relatively low altitude of the *D* layer the difference between the path travelled by the carrier wave and that travelled by the wave reflected against the *D* layer is usually small compared with the wave length. These waves are therefore not clearly separated from each other but form a single vibration phenomenon which is propagated parallel to the earth's surface. Fading seldom occurs.

A peculiar phenomenon, which may be caused by the ionosphere in the case of long waves, is a sort of cross-modulation, whereby the modulation of a strong transmitter can be transferred to the carrier wave of other transmitters (the so-called Luxemburg effect). This effect was first observed by Butt and Tellegen, each independently, who ascertained that the modulation of the Luxemburg transmitter could also be heard on several other transmitters lying in about the same direction from the receiver, but at greater distances. This effect is to be ascribed to non-linear phenomena in the ionosphere. Apparently the state of the ionosphere above a powerful transmitter on the long-wave is slightly affected in the rhythm of the modulation, thereby giving rise to a reciprocal action on the radio waves.

Furthermore, on the intermediate wave, between 200 and 600 m fading often occurs, especially at night and particularly in those regions where the intensities of the carrier wave and the wave reflected against the *E* layer are equal. Reception can only be called perfect in the vicinity of the transmitter (radius about 75 km) where the carrier wave is received more strongly than the reflected wave. At a very great distance, where only the reflected wave is received, no fading would be expected, since the cause of the fading phenomena lies in interferences between carrier wave and reflected wave. Nevertheless, even at great distances very marked variations in intensity may occur due to fluctuations in the reflective properties of the ionosphere and also due to interference arising from the fact that when reflected from the ionosphere the signal may reach the receiver along different paths of varying lengths.

During the day the reflected waves of the intermediate region are rather strongly absorbed by the *D* layer.

Wave lengths between 100 and 200 m are usually less suitable for radio communication over long distances. With increasing frequency the carrier wave is more and more strongly damped by absorption in the earth, an effect which is already noticeable at 100-200 m. The reflected wave is weak, since this kind of waves is very much absorbed by the ionosphere. The cause may perhaps be sought partly in the fact that the frequencies of the vibrations of the electrons around the lines of force of the terrestrial magnetic field correspond to the frequencies of this wave region. Resonance is therefore possible, and this is always accompanied by absorption. The wave lengths from 100 to 50 m and from 50 to 10 m are indeed very important for wireless communication. Especially in the latter region, however, the possibility of reception is limited by the occurrence of the dead zone. In the daytime reflection takes place mainly at the *E* layer, while at night the wave penetrates through the *E* layer and is reflected at the *F* layer.

Fading occurs to a very marked degree in this wave-length region. For receivers situated on the edge of the dead zone the average signal strength may vary considerably owing to the expansion or shrinkage of the dead zone. Reception on these

waves depends very much on the activity of the sun. Especially where the trajects pass along the zones of the aurora borealis to earth the reception is very variable.

The strong fading phenomena discussed in the previous section, which are caused by a disturbance in the normal state of the ionsophere, are particularly pronounced in this wave region.

Influence of the earth's magnetic field

In the foregoing it has already been stated that the waves in the region of 100-200 m are strongly absorbed by the ionosphere owing to the terrestrial magnetic field. A closer investigation shows that this is not the only effect of the magnetic field. Under the influence of the earth's field a so-called double refraction occurs, and one speaks, as in the propagation of light in crystals, of the ordinary and extraordinary components. That something of the sort must exist is easily seen when the behaviour at the magnetic equator is studied, where the earth's field is directed horizontally North-South. If at that spot linearly polarized radio waves are sent upwards, whose electrical vector E is also directed North-South, then in the ionosphere the electrons are set in vibration in the direction North-South and they experience in their motion no effect from the earth's field. For such polarized waves the propagation in the ionosphere is independent of the magnetic field and one therefore speaks of the ordinary component. The extraordinary component on the other hand is found when from the magnetic equator radio waves are sent upwards whose electrical vector is directed East-West; the electrons in the ionosphere set in vibration by the radio waves then experience a Lorentz force, with the result that equation (10) for the electron density at which reflection occurs is changed to

$$N = \frac{\pi m}{e^2} f(f \mp f_H) = 1{,}24 \times 10^{-8} f(f \mp f_H) \quad (12)$$

where $f_H = \frac{e}{m} \frac{H}{2\pi c}$ is the so-called Larmor frequency, *i.e.* the frequency at which the electrons can describe circular orbits in the magnetic field, which frequency is equal for all circles. The upper sign holds for $f > f_H$ and the lower for $f < f_H$. For a field strength of about 0.5 Gauss $f_H \approx$ 1.5 megacycles/sec., corresponding to a wave length of about 200 m.

When radio waves are sent upwards whose electrical vector E is directed for instance NW-SE, double refraction occurs in the ionosphere. An ordinary component is received which is polarized linearly N-S, and an extraordinary component polarized linearly E-W. When $f > f_H$ it is found from equations (10) and (12) (upper sign) that the extraordinary component is reflected at a smaller electron density than the ordinary component. As a result the ordinary component is received later than the extraordinary one. At the same time the critical frequency for the extraordinary component then lies higher than that for the ordinary one. Thus if the frequency of the signal transmitted is chosen between these two critical frequencies, the extraordinary component is still reflected but not the ordinary component.

Experiments carried out by the Carnegie Institution in Huancayo, Peru (lying on the magnetic equator) fully confirm the above statements. From this it follows in particular that the effective charged particles of the ionosphere are actually electrons; if the reflection were caused by heavier particles no measurable difference could be expected between the reflection of the ordinary and the extraordinary components.

Experiments carried out at higher magnetic latitudes give a somewhat more complicated result; in that case the reflected waves are polarized elliptically in opposite directions, at the magnetic North pole the ordinary component being polarized circularly to the left and the extraordinary component circularly to the left, while at the magnetic South pole the polarization directions of the two components are just the reverse. Observation of the opposite polarization directions in the Northern and Southern hemispheres clearly confirm these facts.

Significance of the study of the ionosphere for the physics of the sun

In the foregoing it has already been pointed out that the investigation of the ionosphere is closely connected with the physics of the sun, with meteorology and with geophysics. We shall here briefly consider the connection with the physics of the sun.

Some of the ions in the ionosphere have been formed by ionization of nitrogen molecules. This can only be accomplished by radiation of a wave length shorter than 661 Å. If the temperature at the surface of the sun is known, the intensity of the effective ultra-violet radiation can be calculated, considering the sun as a heat radiator. From the distribution of intensity in the visible region a value of at most 6500 °C follows for the temperature of the sun, and from this Saha deduced that in the wave-length region $\lambda < 661$ Å the sun would give off to the earth 10^4 light quanta per cm^2 per second. Since upon absorption by a nitrogen molecule each quantum releases one electron, in a vertical column with a base of 1 cm^2 10^4 electrons are released per second. According to Appleton and Chapman, however, 10^9 to 10^{10} electrons per cm^2 per second are necessary for the maintenance of the ionization. This result indicates that the intensity of the sun's radiation emitted in the far ultra-violet region is about 10^6 times greater than that of a glowing black body with the temperature of the sun. It is very probable that this extra ultra-violet radiation originates in the hydrogen and helium present on the sun.

The foregoing may serve to show what progress has been made in the radio investigation of the ionosphere. A much deeper insight has been obtained into the physical phenomena taking place in the ionosphere and it is not only radio technics that have profited from this, for the investigation of the ionosphere now also furnishes important material for geophysics, meteorology and astrophysics.

LITERATURE

Among the numerous articles concerning the ionosphere we would mention:

H. R. Mimno, The physics of the Ionosphere, Rev. Mod. Phys. **9**, 1, 1937.

W. Grotian, Sonne und Ionosphäre, Naturw. **27**, 555 and 569, 1939.

J. H. Dellinger, The role of the Ionosphere in radio wave propagation, Electr. Eng. Transactions. Suppl. 803, 1939.

K. K. Darrow, Analysis of the Ionosphere, Bell Syst. Techn. J. **19**, 455, 1940.

See also:

L. Harang, Das Polarlicht und die Probleme der höchsten Atmosphärenschichten, Leipzig 1940.

C. J. Bakker, M. Minnaert, A. Pannekoek, J. Veldkamp, De ver-ultraviolette en de corpusculaire stralingen der zon (The far ultra-violet and the corpuscular radiations of the sun). Ned. T. Natuurk. **9**, 209, 1942.

AN EXPERIMENTAL TRANSMITTER FOR ULTRA-SHORT-WAVE RADIO-TELEPHONY WITH FREQUENCY MODULATION

by A. van WEEL. 621.396.5 : 621.396.615.14

A transmitter working with frequency modulation has been developed for an experimental ultra-short-wave radio-telephonic link between the Philips factories at Eindhoven and those at Tilburg, Holland, on a wave length of 90.5 cm for one direction and 99 cm for the other. Modulation with frequencies from 12 to 204 kc/sec, for 48 telephone calls at once, takes place on a carrier wave with a frequency equal to 1/9 of the desired transmitter frequency, with a maximum frequency swing of 67 kc/sec. By frequency multiplication a signal is obtained with the desired transmitter frequency and with a frequency swing 9 times as large. This article points out the advantage of this method and describes how the frequencies have been chosen and the transmitting stages arranged. A brief description is also given of the construction of the transmitter, attention being drawn in particular to the distribution of the circuiting over two separate panels and the simplification of the wiring, both of which have been made possible by the application of a new method of coupling between successive stages, further explained in this article.

An experimental radio-telephony link between the Philips factories in Eindhoven and those at Tilburg [1]) has existed for several years. This has been working lately on a wave length of 90.5 cm for one direction and 99 cm for the other. The link is so arranged that the radio installation can function as an entirely automatically acting link in the telephone network. Neither the person phoning nor the operators of the telephone exchange through which the connection passes need to be aware of the fact that the calls are transmitted by wireless instead of by cable.

Since the installation was first set up there have been important developments both in the field of telephony and in short-wave transmitting technique. In telephony there is an increasing tendency to use a single pair of conductors in a cable for the transmission of a large number of calls at the same time by means of a carrier telephone system [2]). The frequency band of about 3000 c/sec width necessary for transmitting a call is modulated on a specific "carrier", the carriers for the different calls ("channels") lying 4000 c/sec apart (only one side band is used), for instance at . . . : 32 kc/sec, 36 kc/sec, 40 kc/sec, etc. It was found desirable to have the radio-telephone link referred to above equipped for such a carrier telephone system, namely for 48 channels, in connection with the available carrier telephone apparatus. The 48 channels, which cover the frequency region from 12 to 204 kc/sec, have to be modulated as a whole on the radio wave, which in this case serves as a "pair of conductors" in the link. The transmitter and receiver had therefore to be made suitable for this very wide range of modulation frequencies (up to 200 kc/sec).

At that time there was an important development in the field of ultra short waves (< 10 m), *viz* the gradual superseding of the old method of amplitude modulation by frequency modulation. The advantages of this method led us also to rebuild our ultra-short-wave link for the new method.

All these facts led to a complete reconstruction of the transmitter and receiver of this experimental communication, not only as regards electrical connections but also in the actual construction. Of the old installation only the Yagi directional aerials for transmitter and receiver [1]) could be retained unaltered.

In this article the transmitter in its present form will be described, while the receiver will be dealt with in a subsequent article.

The method of modulation

The advantages of frequency modulation over amplitude modulation are twofold: 1) In amplitude modulation of ultra short waves it is almost inevitable that also an undesired frequency modulation occurs; the difficulties created by such a mixed modulation are avoided by employing pure frequency modulation. 2) Frequency modulation, compared with amplitude modulation, gives an appreciable improvement in the ratio between the intensity of the signal and that of the fluctuation noise [3]).

[1]) C. G. von Lindern and G. de Vries: An ultra short wave telephone link between Eindhoven and Tilburg, Philips techn. Rev. 2, 171, 1937.

[2]) See for example the articles published in this periodical about carrier telephony: Philips techn. Rev. 4, 20, 1939; 6, 325, 1941; 7, 83, 104, 184, 1942.

[3]) For readers who wish to know more about frequency modulation we refer to the articles by Th. J. Weyers in the preceding numbers of this periodical: Philips techn. Rev. 8, 42 and 89, 1946.

In order to make full use of the second advantage, *i.e.* in order to reduce noise to the lowest possible level, it is necessary that the maximum frequency swing, *i.e.* the largest deviation occurring from the average frequency, should be about ten times as large as the highest modulation frequency to be transmitted. In our case this was not less than 0.2 megacycles/sec, so that the maximum frequency swing had to amount to 2 Mc/sec. Since for the undistorted transmission of a frequency-modulated oscillation, a side band of at least $1^1/_2$ times the frequency swing on both sides of the carrier frequency has to be transmitted, the transmitter and receiver would have to be adapted for a frequency band with a total width of 6 Mc/sec. Because of the difficulties that would have been involved in getting such a great band width, we confined ourselves to a frequency swing of the emitted signal equal to about three times the highest modulation frequency. In this way one arrives at a band width of 2 Mc/sec, which is well possible in practice, while the ratio between intensity of signal and that of noise is still 14 db greater than it would be with amplitude modulation.

With frequency-modulated transmitters it is customary to apply the actual modulation process to oscillator connections which oscillate at a low frequency. By frequency multiplication, where the average oscillator frequency and the frequency swing are increased in the same proportions, the desired transmitting frequency is then obtained. This method, which we also employed, has the advantage that it is easier to obtain the required proportionality between frequency swing and intensity of the modulated signal. This can be explained as follows.

Modulation is effected by detuning the oscillator circuit with the help of a so-called reactance valve [3]). This is a normal multigrid valve, between the cathode and anode of which the circuit to be detuned is connected, while part of the circuit voltage is applied to one of the grids with 90° phase displacement. The anode AC is then shifted 90° in phase with respect to the anode AC voltage, in other words the valve acts as a reactance (capacitative or inductive, according as the grid AC voltage is shifted +90° or −90° in phase). To take a specific case, let us assume that the tube acts as a capacity, C_1. The reactance $1/\omega C_1$ is equal to the ratio of the amplitudes of anode AC voltage and anode AC, while the frequency ω at which the circuit connected oscillates is determined by C_1 together with the remaining circuit capacity C_0. C_1 is now varied by applying the modulating signal voltage to a second grid of the valve, whereby the slope, and with it the amplitude of the anode AC, is changed in the rhythm of that voltage. The detuning, $\Delta\omega$, of the circuit obtained with respect to the frequency ω_0, with $C_1 = 0$ is given by

$$\Delta\omega = \frac{\omega_0 C_1}{2C_0}.$$

We will now consider the influence of the oscillator frequency to be chosen, ω_0. In connection with the linearity of the modulation, only a certain change in slope of the reactance valve is permissible, namely such that there is no deviation from that section of the valve characteristic where the slope is by sufficient approximation proportional to the grid voltage. Corresponding to the permissible maximum slope, which depends only on the properties of the valve and not on the oscillator frequency ω_0, is a certain maximum anode AC, thus — because the anode AC voltage is fixed — a certain maximum value of the reactance $1/\omega_0 C_1$, likewise independent of ω_0. The maximum detuning to be obtained $\Delta\omega$, which according to the formula is proportional to the greatest value of $\omega_0 C_1$, is therefore independent of the frequency ω_0, because C_0 may also be considered as a constant. In order to make the detuning by the reactance valve as great as possible, C_0 will be kept as small as possible, and thus will be limited to the unavoidable capacities of valve and wiring. The choice of ω_0 is then realized with the self-induction of the circuit.

Since, therefore, the same absolute frequency sweep can be obtained with a low oscillator frequency as with a high one, if in the manner described one begins with a low oscillator frequency and a correspondingly small frequency sweep, it will actually be easier to ensure the necessary linearity of the modulation. We shall revert later to the exact choice of oscillator frequency.

The connections

The connections are in push-pull arrangement for various reasons: variations in the feeding voltage are much less manifest as undesired modulation, since they act on the two halves of the modulator in the same phase instead of in counterphase; furthermore one avoids the strong high-frequency currents which otherwise flow in the earth connections and therefore (since all the points to be earthed cannot be connected to the same point of the chassis) also through the chassis, causing all kinds of undesired couplings, etc; finally, by using push-pull amplifier valves several difficulties occurring at very high voltages are diminished, because,

among other reasons, the influence of the self-induction of the cathode feeding connections is much smaller[4]). In *fig. 1* a block diagram is given of the transmitter connections. We shall first consider only that part drawn with heavy lines. The low-frequency signal from the telephone exchange (the term "low-frequency" is here used in a relative sense, the frequencies of this signal being as high as about 200 kc/sec.) is led to two reactance valves which influence the frequency of the push-pull oscillator. The signal obtained from the oscillator, *i.e.* the frequency-modulated oscillator voltage, is led to a push-pull valve, connected as a frequency triplicator. Then the signal is amplified and conducted to another frequency triplicator. The signal here reaches its final frequency, after which it is further amplified in two end stages and sent to the aerial.

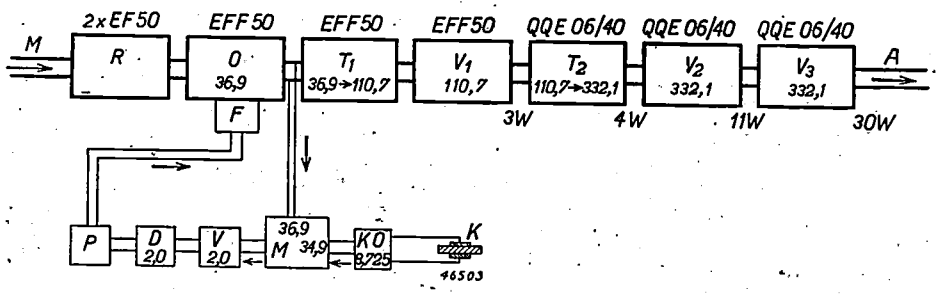


Fig. 1. Block diagram of the transmitter. The modulating low-frequency voltage is applied at M, R reactance valves, O oscillator, T_1, T_2 frequency-triplicator stages, V_1-V_3 amplifier stages. At A the frequency-modulated output voltage is fed to the aerial. The numbers in the blocks indicate the frequencies in Mc/sec at which the stages work; above each block the type of valve used for that stage is given. The part of the diagram drawn with thin lines serves to keep the average oscillator frequency constant. This is explained later (see fig. 3).

The choice of oscillator frequency

The reason for the multiplication of the frequency in stages by a factor of three lies in the nature of the connections. In push-pull connections the output voltage can contain, in principle, only the odd harmonics of the frequency of the grid AC voltage applied, so that by filtering out the harmonics in question a frequency multiplication by a factor 3, 5 or 7, etc. can be obtained. For the sake of efficient output the lowest factor has been chosen. Since the transmitting frequency, if we confine ourselves for the moment to one single call direction, was fixed at 332.1 Mc/sec (90.5 cm wave length) the possible choices of oscillator frequency were 332.1 - 110.7 - 36.9 - 12.3 - 4.1 Mc/sec, etc. We have explained above that the oscillator frequency must be chosen as low as possible. From the frequency values indicated in fig. 1 in the different stages it may be seen that the choice fell upon 36.9 Mc/sec. This still seems relatively high, but it must be taken into account that the receiver for the same connection is situated close to the transmitter. The receiver works on the superheterodyne principle with an intermediate frequency of 18 Mc/sec. In order to prevent interferences in the reception it is necessary that neither the oscillator frequency of the transmitter nor any harmonics of it shall fall in the intermediate-frequency band of the receiver, which band, according to the above figures, must extend from 17 to 19 Mc/sec. An oscillator frequency of 12.3 or 4.1 Mc/sec. would, it is true, also answer this condition, but the transmitting apparatus in question had to be so designed that later on, if necessary, the connections built up with the same stages could also be used for telephone links on other wavelengths between about 1.50 and 0.90 m, simply by changing slightly the oscillator frequency. In order to avoid once for all the danger of interferences in the receiver it was therefore decided to place the oscillator frequency above the intermediate-frequency band of the receiver. At the same time of course the possibility had to be considered of choosing that intermediate frequency itself lower; in the discussion of the receiver it will be shown, why this was not done.

It is perhaps advisable to stress the point that the requirement mentioned above, that for a favourable ratio between signal and noise the frequency sweep must be several times as large as the highest modulation frequency occurring, is only applicable for the signal emitted by the aerial. In our case, as a consequence of the frequency multiplication, the maximum frequency swing in the oscillator stage amounts to only 1/9

[4]) M. J. O. Strutt and A. van der Ziel: A new push-pull amplifier valve for decimetre waves, Philips techn. Rev. 5, 172, 1940.

of the final value, *i.e.* $1/9 \times 3 \times 200 = 67$ kc/sec. If a modulating signal with the frequency of 200 kc/sec and with the largest permissible amplitude is applied to the oscillator, the oscillator delivers an oscillation with a frequency fluctuating 200 000 times per second between 36.833 and 36.967 Mc/sec. In the case of the frequency multiplication here employed the rhythm of the fluctuation, *i.e.* the modulation frequency, naturally remains unaltered: the following stage delivers an oscillation with a frequency fluctuating 200 000 times per second between 110.5 and 110.9 Mc/sec, etc.

The valves used

The valves used are indicated over the blocks in the diagram of fig. 1. Except for those in the three last stages they are all normal receiving valves with only a low energy dissipation so that a compact assembly is possible. The last stages must of course have transmitting valves in order to produce the power required for transmission.

Nevertheless, the energy amplification in the last two stages is only slight, as may be seen from the wattage figures given in fig. 1. This is due to the high frequency at which these amplifier stages have to function. It would seem obvious to ask why the once amplified signal of 110.7 Mc/sec is not first amplified further to the desired final level and then given the necessary frequency amplification in the last stage. This, however, is impossible, because if that were done very high AC voltages (of the order of 500 Volts) would have to be applied to the grid of the last valve; in order to function as frequency multiplier the valve must work in class C, thus with very high negative grid bias. It has been found in practice that with the very small distance between grid and cathode in these short-wave tubes the high voltages mentioned lead to disturbances: breakdown may occur or the insulation between grid and cathode may be damaged (especially at very high frequencies).

Fig. 2 is a reproduction of two X-ray photographs of the type of short-wave transmitting valve used, the QQE 06/40. It is a double tetrode in which the two balanced systems have a common indirectly heated cathode and a common screen grid. The power on a wave length of 3 m is 40 W; at 1 m about 30 W.

Maintaining the constancy of the oscillator frequency

The part of the diagram in fig. 1 which is drawn with thin lines serves to keep the average oscillator frequency constant; it is given again separately in *fig 3*.

The usual method of synchronizing the oscillator vibration directly with the characteristic oscillation of a quartz crystal or a harmonic of the same could not be employed in our case since the oscillator vibration is already frequency-modulated upon its formation, and consequently when keeping the

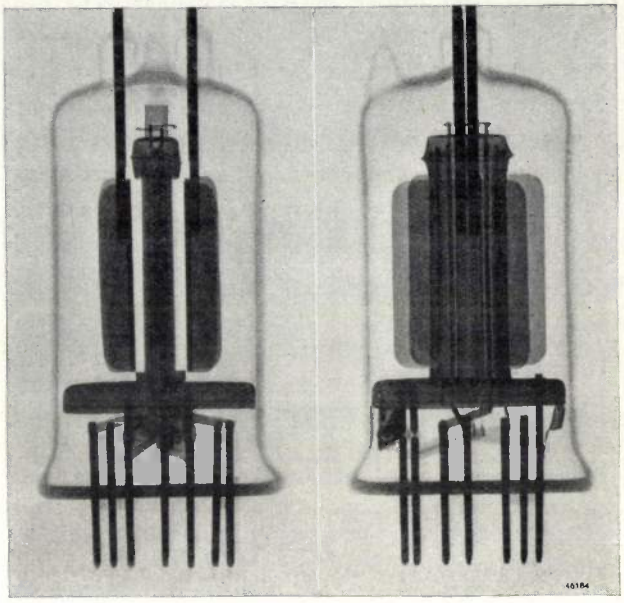


Fig. 2. X-ray photographs taken in two mutually perpendicular directions of the ultra-short-wave push-pull transmitter valve, type QQE 06/40, used in the three last stages of the transmitter. The cathode connection is made very short thanks to the two electrode systems having a common indirectly-heated cathode.

oscillator frequency constant a certain margin has to be left for the frequency swing. For that reason the following method was chosen.

A small part of the output voltage of the oscillator stage, of which the average frequency f_o must amount nominally to 36.9 Mc/sec, is tapped off and

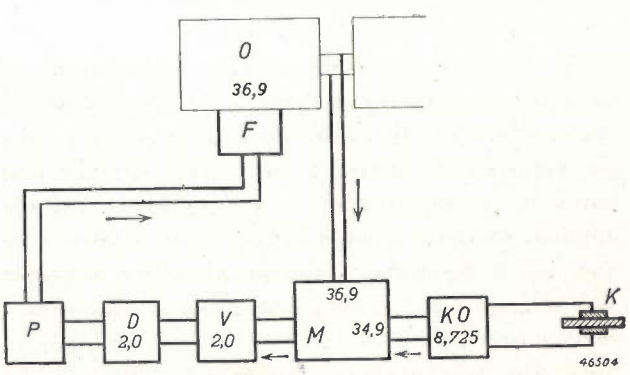


Fig. 3. Block diagram of the connections for keeping the average oscillator frequency constant. *K* quartz crystal which keeps the frequency 8.725 Mc/sec of the crystal oscillator *KO* constant, *M* mixing stage in which the fourth harmonic of this frequency (34.9 Mc/sec) is mixed with the frequency $f_o \approx 36.9$ Mc/sec of the transmitter oscillator *O*, *V* amplifier, *D* discriminator, *P* pentode, *F* coil with ferromagnetic core coupled magnetically with the circuit self-induction of the oscillator *O*. The numbers in the blocks indicate the frequencies in Mc/sec.

in a mixing valve mixed with an AC voltage of the very constant frequency of 34.9 Mc/sec from a vibrating crystal. The output voltage of this mixing stage has an average frequency $f_1 = f_0 - 34.9$ Mc/sec. If f_0 is exactly 36.9 Mc/sec $f_1 = 2$ Mc/sec; if f_0 differs slightly from 36.9 Mc/sec f_1 exhibits the same absolute difference from 2 Mc/sec. Furthermore, the "intermediate frequency" f_1 is of course frequency-modulated with the telephone frequencies in the same way as the high frequency f_0 applied. This output voltage of the mixing stage is now amplified to a discriminator connection. This produces a DC voltage proportional to the difference between the average frequency f_1 of the applied voltage and the fixed frequency $f_d = 2$ Mc/sec at which the discriminator is set. An AC voltage is superposed on the DC voltage, namely the o iginal modulation voltage, which, however, is suppressed by a low-pass filter. The DC voltage is now applied to the grid of a pentode. The anode current of this valve, which is thus proportional to the absolute deviation of the average oscillator frequency from the prescribed value of 36.9 Mc/sec, flows through a coil wound around a core of ferromagnetic material [5]) magnetically coupled with the selfinduction coil of the oscillator circuit. Owing to the fact that the anode current changes the premagnetization and thus the permeability of the coil core, it affects the self-induction of the oscillator coil in the sense that the change in the average oscillator frequency, to which the anode current is proportional, is opposed. In this way it proved possible to reduce the drift of this frequency by a factor 30.

The "fixed" frequency f_d at which the discriminator is set is, of course, not actually fixed either, but subject to some variation due to the drift of the elements of the discriminator connections with temperature, etc. The fact that this forms no obstacle to the regulating action of the whole is due to the discriminator reacting to the absolute changes of the oscillator frequency. Even if the discriminator frequency f_d should drift proportionally just as much as the oscillator frequency f_0 does with no regulation, the absolute variations of the latter are still $36.9/2 \approx 18$ times as large as the drifts of f_d.

The coupling between the successive valves

For the coupling together of the various triplicator and amplifier stages a new method has been employed which offers important advantages. In order to explain the particulars of this coupling, let us consider *figs. 4a* and *b* in which a coupling

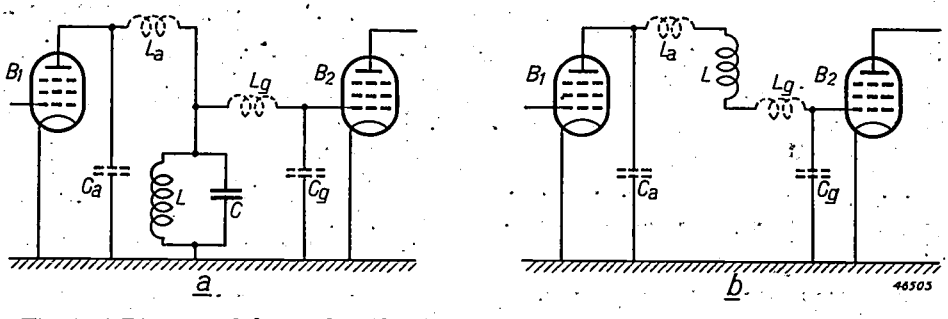


Fig. 4. *a*) Diagram of the usual method of coupling two amplifier valves B_1 and B_2; *L-C* tuned parallel circuit. *b*) New method of coupling. C_a, C_g, and L_a, L_g are the internal capacities and self-inductions of the valves involved in the coupling. The elements necessary for determining the DC voltage situation are not drawn.

according to the usual method and one according to the new method are shown side by side. For the sake of simplicity ordinary (not push-pull) stages are assumed and all elements only of importance for the DC voltage situation are omitted. In the ordinary coupling the anode of the first and the grid of the second valves are connected directly with each other, while between this point and earth a tuned parallel circuit is connected (the internal self-inductions L_a and L_g of the valves, which are indicated with dotted lines in figs. 4*a* and *b*, may for the present be disregarded). In the new coupling, on the other hand, the tuned circuit is formed by the self-induction L and the connection in series of the internal valve capacities C_a and C_g. In *figs. 5a* and *b* the situation in the two cases is shown still more concisely. Let us first consider fig. 5*b*. The high-frequency voltage between the points A and B is mainly determined by the oscillation of the circuit, but the potentials of these two points fluctuate in opposite phase with respect to earth due to the obtrusion of earth potential at point O in the middle of the circuit capacity. From this it

[5]) It is necessary to use a ferromagnetic material the losses of which are sufficiently small, even at high frequencies.

follows that also in the middle of the circuit self-induction there must be a point (P) which remains at earth potential during oscillation and thus has no

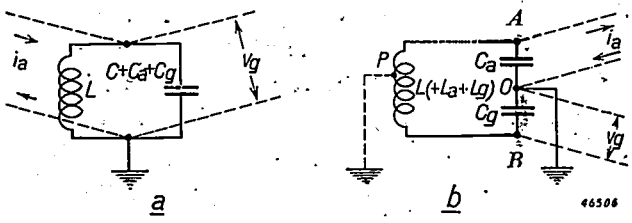


Fig. 5. *a*) More detailed diagram of the usual coupling circuit. The anode current i_a of the first valve and the grid voltage v_g for the second valve are respectively applied to and taken from the same pair of terminals.

b) The same for the new method of coupling. Here i_a is applied to the "terminals" of C_a, while v_g is taken from the terminals of C_g. Since the point O lying in the middle of the circuit capacity remains at earth potential, there is also a point P on the self-induction with the same property.

high-frequency voltage [6]). The position of P on the self-induction L is determined by the ratio of the capacities C_a and C_g. On the other hand, turning to fig. 5*a*, it is evident that in the old coupling connections no such non-voltage point can be found. Even if the self-inductions L_a and L_g are taken into account this fact is not altered.

The occurrence of a non-voltage point in the new method of coupling offers the possibility of dividing the transmitter constructionally into two parts at a non-voltage point between two valves and assembling the two parts, for example, in separate panels. The connection between the two panels then carries no high-frequency voltage (high-frequency current does, however, flow through it) and therefore no undesired couplings or radiation can occur. If such a division were made at a point which was not voltage-free, the connection would have to be shielded against the effects mentioned, and a large extra circuit capacity would thereby be introduced, in general resulting in a loss of amplification. For this reason it is very difficult to divide the connections when the old method of coupling is used.

Another advantage of the new method of coupling is of particular importance at very high frequencies. Since in this case the capacity and/or self-induction of the coupling circuit must become very small, the contributing, unavoidable capacities and self-inductions of the valves begin to play an important part. Fig. 4*a* cannot then be reduced to the simple situation of fig. 5*a*, because of the presence of L_a and L_g, which can no longer be ignored. The ordinary connections are now in principle more complicated than a simple L—C circuit, and in practice it proves difficult to obtain sufficient amplification with them. This is understandable when it is borne in mind that at very high frequencies L_g may already be approximately in resonance with C_g. The tuned parallel circuit L-C, whose task it is to furnish a high impedance for the circuit frequency, is then as it were short-circuited by the low impedance of the series circuit L_g—C_g. Measures for overcoming these difficulties are known but they only result in making the circuit still more complicated, more extensive and more difficult to manipulate. When we compare this method with the new coupling according to fig. 4*b*, we see that the presence of L_a and L_g does not alter anything in the principle of the connections and only makes it necessary to choose the external self-induction L somewhat smaller than the total circuit self-induction desired. Even if L_g or perhaps only a part of it is of itself tuned to C_g, in the main everything remains unchanged. It means really nothing else than that the above-mentioned non-voltage point then lies exactly at the valve terminal or somewhere in the internal grid feed connection, as the case may be. It is then no longer possible to divide the connections at that point, but the action of the amplifier stage is not at all affected. It would only become troublesome if a negative value should be required for the external self-induction L. But since C_a is usually much smaller than C_g, the self-induction necessary to tune C_a, *i.e.* the self-induction between the anode and the non-voltage point, is generally large enough to cover, in addition to the self-induction L_a of the internal anode feed connection, a positive external self-induction as well.

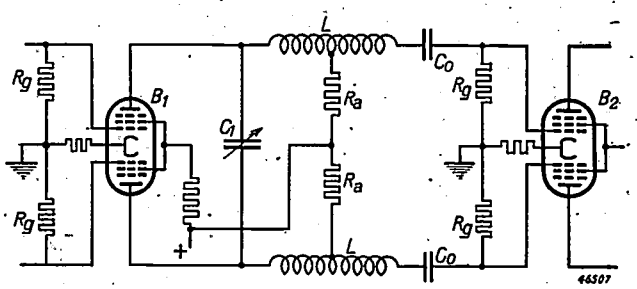


Fig. 6. Example of a push-pull amplifier stage according to the new method of coupling. Between the anodes of the push-pull valve B_1 and the control grids of the valve B_2 two self-inductions L are connected. The capacities C_0 only have the function of separating the DC voltage positions of anodes and grids. The circuit is tuned with the variable capacity C_1. The anode DC is supplied through the resistances R_a; R_a may not therefore be chosen very large; in spite of this in order not to obtain any undesired damping of the circuit these resistances are connected at the non-voltage points. There is no objection to choosing large grid leakage resistances R_g.

[6]) Strictly speaking this is not true. If the losses of the circuit elements are taken into account it is found that there is a certain residual voltage. This is very small, however, compared with the circuit voltage.

Fig. 6 shows the complete connections of a push-pull amplifier stage with the new coupling method. Several details are explained in the text of the figure.

Construction

If a transmitter is to serve as part of a carrier telephone system it is desirable that its external form should be adapted to that of the carrier telephone apparatus. In the case in question the transmitter had to be housed in a rack of certain dimensions with the components mounted on panels of a certain length and width which could be slid into the rack from front and rear. Similarity in external appearance, however, can never be complete. In the case of the carrier telephone apparatus

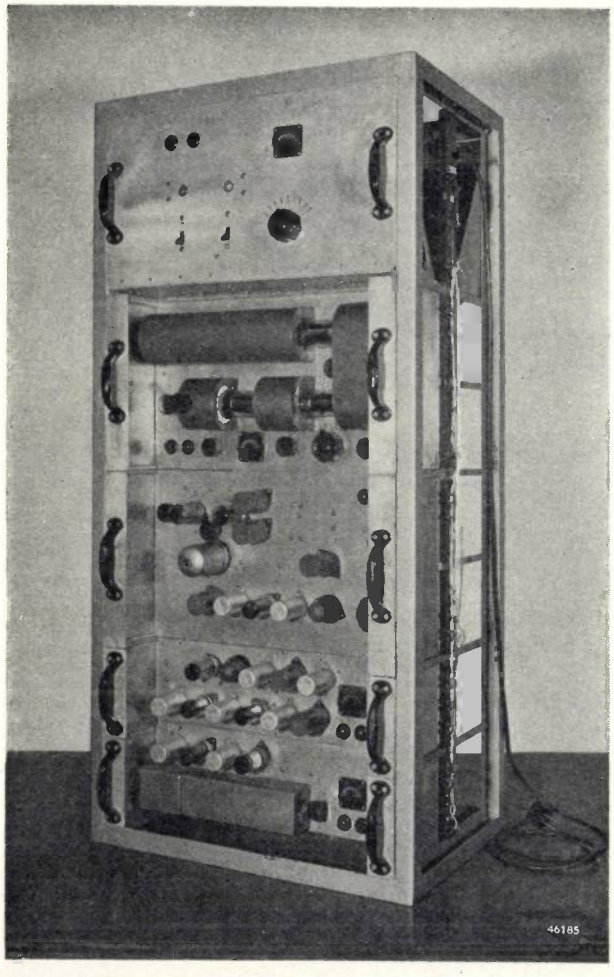


Fig. 7. The transmitter and receiver assembled in a rack. From top to bottom: a panel (of double depth) containing the high tension supply unit, two panels with the transmitter, two panels with the receiver. All the panels are constructed as sliding drawers. On the front plate of the upper transmitter panel may be seen the valves of the three last stages, within, between the coupling connections, shielded by caps. On each panel is a meter with which the cathode currents of all the valves can be checked by means of a switch. In the open spaces at the back of the rack identical drawers with a transmitter and receiver can be inserted to serve as reserves.

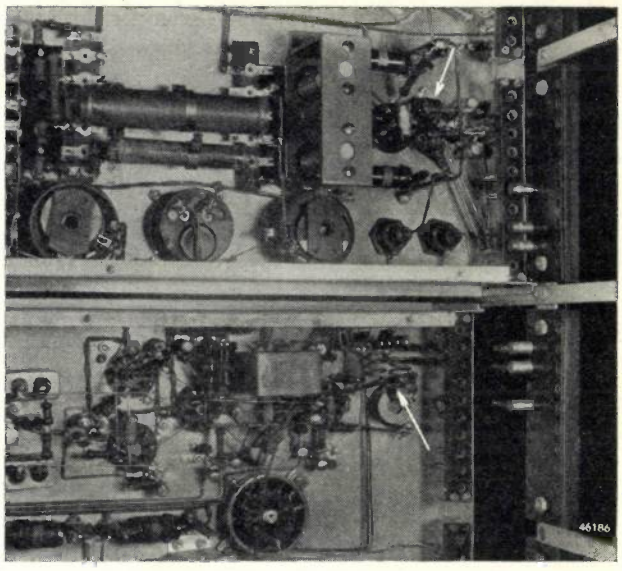


Fig. 8. Part of the two transmitter panels seen from the rear. The two small coils in each panel on the extreme right (indicated by arrows) are the parts of the two coupling coils divided into two. In the lowest drawer, pulled halfway out, may be seen the plug pins and sockets.

the available space in every panel is almost completely filled with the parts of the different filters, modulators, repeaters, etc. Because of the much higher frequencies, the components of the transmitter, however, must in general be spread out over a certain area in order to provide sufficient mutual intervals to prevent undesired couplings. One may speak here of surface assembly in contrast to the volume assembly in the apparatus for carrier telephony. In this way one arrives at the construction shown in *fig. 7*. The components are mounted side by side in each panel in a single plane, a "front plate". This lies fairly deep in the rack, so that the different valves, coil cans, Lecher systems, etc. situated on the front of the plate are well protected against undesired contact or shocks without it being necessary to place a cover plate in front of the rack, which would hinder the dissipation of the heat given off by the valves to the air.

The uppermost panel of the installation is actually double, occupying the entire depth of the rack in order to offer space for the high tension supply unit mounted therein. This is placed at the top in order that the rather large amount of heat developed in it shall not cause any difficulties in the transmitter part proper. The transmitter also contains too many components to be housed on one normal panel. As a consequence of the above-mentioned method of coupling between successive stages the transmitter could without difficulty be divided and housed in two separate panels one above the

other. This can be seen in fig. 7 underneath the high tension supply unit. *Fig. 8* shows the rear of the transmitter panels with the parts of the two coupling coils (push-pull connections, see fig. 6) in each panel. Below the transmitter panels is the corresponding receiver, also divided between two panels. On the other side of the rack two exactly identical panels with a complete transmitter and receiver can be inserted to serve as reserves, and in case of a breakdown these can be switched on by a single touch of the hand or entirely automatically.

Upon inserting a panel all connections of the various signal and feeding voltages are automatically made by means of plug pins in the rack and sockets in the panels. The lead wires to the plug pins are led out at the side, where the wiring harness for the mutual connection of the panels is mounted. The details can be seen in figs. 7 and 8. On each panel a measuring instrument is mounted on which the cathode current of each valve can be checked by means of a switch. By this means it is possible to localize the cause of any interruptions quickly, while the ageing of the valves can also easily be ascertained in good time. This is necessary because the same requirements as to reliability are made of the transmitting apparatus as are made of the telephone apparatus itself.

VOL. 8 No. 5, p. 129-160 MAY 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

AIR ENGINES

by H. RINIA and F. K. DU PRÉ.

621.412:536.8

Around the middle of the last century considerable interest was shown for a time in the air engine, but this was quickly supplanted by the development of internal combustion engines. This was not due to the thermodynamic process involved, but to the technical constructions possible in those times. Recent research work in Philips laboratories has now shown that very satisfactory results can be attained with air engines, if they are built in accordance with modern conceptions in regard to heat transfer, flow resistance, properties of materials, etc. In this article the theoretical principles of the air process are discussed.

Introduction

When reviewing the processes that have been applied in the course of years in an attempt to convert heat into mechanical work, it is remarkable that the air process at present in disuse was considered so very promising for quite a while. It was first used in 1817 by Stirling, and afterwards various air engines, some of them very large, were built and put into practical use. Some people were even of the opinion that the air engine was a serious competitor of the steam engine, that had previously come into use.

These early air engines, however, were so unwieldy, slow and uneconomical that it is not surprising, that they were entirely supplanted after the invention of the internal combustion engine. After a short while, therefore, interest in the air process faded out almost entirely.

Now it was known that according to thermodynamics the air process was to be considered as one of the most economical methods of generating mechanical energy from heat. In the previous century, however, it was technically impossible to construct a good air engine. As a result of the extensive theoretical and experimental research work that has been carried out on this subject for several years past in the Philips laboratories, it has been established that present-day technology is indeed capable of exploiting the theoretical possibilities of the air process efficiently.

By making use of modern materials and modern conceptions of heat-transfer and flow-resistance it was found possible to apply the air process in engines capable of performing 3000 r.p.m, with most satisfactory figures for weight and efficiency. In this article we shall confine ourselves to the main theoretical factors which play a part in the air process.

Principle of the air engine

We shall explain the principle of the air engine with the help of a very much simplified model. Let us imagine a cylinder divided into two parts in open connection with each other. One part, the so-called hot space, is kept at a high temperature T_h by means of a heater, while the other part, the cold space, is kept at a low temperature T_c by a cooler. In each of the two parts of the cylinder is a moving piston. A certain quantity of air is enclosed between the two pistons. We shall assume for the present that the transfer of heat between the cylinder and the air in it is so good that the air in the hot space is always at the temperature T_h and that in the cold space always at the temperature T_c.

We now cause the air in the cylinder to pass through a cycle consisting of four phases. The four

positions of the pistons at the moments when one phase passes over into the next are shown in *fig. 1a* as *I* to *IV*. The cycle is as follows:

At position *I* all the air is in the cold space, which has the maximum volume V_1.

The transition *I-II* takes place isothermally, according to our assumption about heat transfer; the air is compressed to the volume V_2.

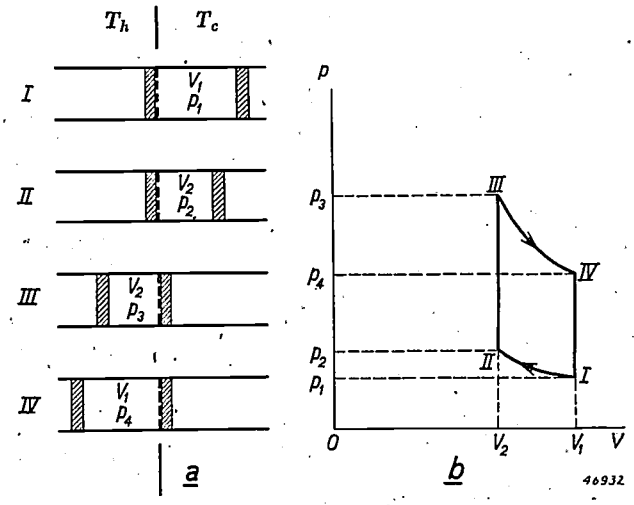


Fig. 1. Diagram explaining the action of an air engine.

a) The engine consists of a cylinder which is divided into a hot and a cold part with temperatures of T_h and T_c respectively, which are in open connection with each other. In both parts there is a piston. Between the two pistons is a definite quantity of air. In the engine, a cyclic process takes place four stages of which are indicated:
The transition *I-II* takes place isothermally at T_c.
The transition *II-III* takes place isochorically, *i.e.* with constant volume.
The transition *III-IV* takes place isothermally at T_h.
The transition *IV-I* takes place isochorically.

b) The *p-V* diagram of the cycle through which the work medium passes. Owing to the fact that during the hot expansion (*III-IV*) the pressure is higher than during the cold compression (*I-II*), the work produced during the expansion is larger than the work supplied during the compression. As a result there is a positive work surplus for the cycle as a whole. This surplus is represented by the area of the curvilinear quadrilateral *I-II-III-IV*.

At position *II* all the air is still in the cold space, which, however, now has the minimum volume V_2.

The transition *II-III* takes place at constant (minimum) volume. The air is displaced from the cold to the hot space.

At position *III* all the air is in the hot space, which has the minimum volume V_2.

The transition *III-IV* again takes place isothermally; the air expands from the minimum volume V_2 to the maximum volume V_1.

At position *IV* all the air is in the hot space, but this now has the maximum volume V_1.

The transition *IV-I* again takes place at constant (maximum) volume, the air being again displaced to the cold space.

It is not difficult to see that the process outlined can in principle be applied for an engine.

In the transitions *II-III* and *IV-I* no mechanical work at all is done. The work which has to be supplied to the one piston is exactly the same as the work done by the other (all mechanical friction being disregarded of course), because the volume of the air remains constant during these transitions. Therefore we need only consider the transitions *I-II* and *III-IV*. The first is a compression and thus work must be done. This compression, however, takes place at a low temperature and therefore also at a low pressure. The second is an expansion and thus produces work. Since this takes place at a high temperature (and because the same series of volume changes takes place as in the compression, although in reverse order) the average pressure is higher than in the compression and the work produced is greater than the work required for the compression. As a result there is a surplus of work per complete cycle.

The action of engines with one piston per cylinder usually consists in the piston being displaced during one half of a revolution under the influence of a high pressure and then moving back again during the succeeding half revolution against a lower pressure. In the presently described construction of the air engine this process is divided between two pistons, and between the two processes there is each time a workless phase. Otherwise the situation is quite analogous. In *fig. 1b* the *p-V* diagram is drawn for the cycle outlined; the phases described above can easily be recognized. The work gained per cycle is represented in the diagram by the area of the curvilinear quadrilateral *I-II-III-IV*.

From the foregoing it is evident that the surplus of work is obtained owing to the fact, that the isothermal expansion of the air takes place at a high temperature and the isothermal compression at a low temperature. This is one of the fundamental principles of the air engine, and further use will be made of it in the following.

The cyclic process as described above is of course in that form very difficult of realization technically; the four-phase piston movements would lead to very complicated constructions. It can, however, be so modified, while still retaining the principle, as to make it technically possible of achievement. Let us now imagine the two pistons as being coupled by a combination of driving rods in such a way that the movements of the hot piston are followed, with a

certain phase difference, by those of the cold piston. In *fig. 2* a very schematic representation is given of how that can be done. With this form of construction it is no longer possible to distinguish the separate phases as they begin to overlap somewhat, and in particular the expansion does not take place completely in the hot space nor the compression completely in the cold space. However, when the

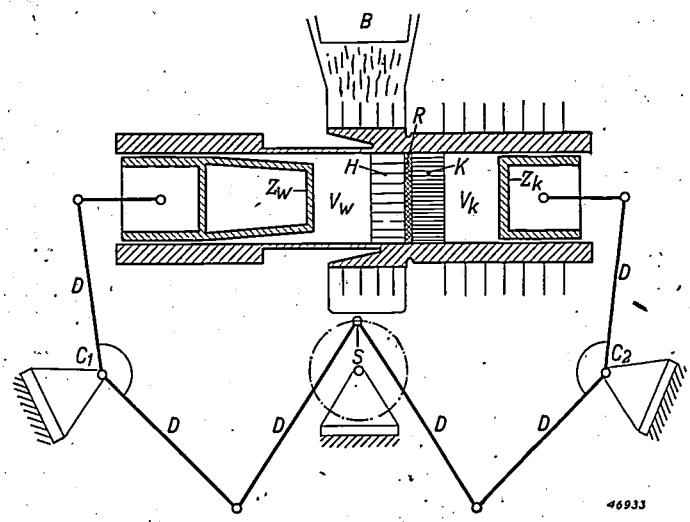


Fig. 2. Diagram of the coupling of the two pistons of the air engine. V_w hot space, V_k cold space, Z_w hot piston, Z_k cold piston, D driving mechanism, C_1 and C_2 fixed pivots, S crank shaft, B burner. In the figure the heater H, regenerator R and cooler K, which will be discussed later, are also indicated.

phase difference is suitably chosen, *i.e.* in the neighbourhood of 90°, it can be shown that the expansion does take place mainly in the hot space and the compression mainly in the cold space, so that this new cyclic process will also furnish a surplus of work per cycle, *i.e.* in this case per revolution of the crank shaft.

Variation of the air pressure p in the engine during one cycle

Before we proceed to calculate the work gained per revolution, we shall first determine how the air pressure in the engine varies, on the assumption that the engine runs uniformly with a circular frequency ω. This can always be accomplished by fitting a flywheel with a sufficiently large moment of inertia on the shaft of the engine. We will assume further, that the volumes of the hot and cold spaces then vary purely sinusoidally. The following notation will be used:

V_0 maximum volume of the hot space,
V_h volume of the hot space at any arbitrary moment,
v ratio of the maximum volumes of cold and hot spaces,
vV_0 maximum volume of the cold space,
V_c volume of the cold space at any arbitrary moment.

Finally we assume that V_h and V_c have zero as lower limit.

For V_h and V_c we may then write

$$V_h = \tfrac{1}{2} V_0 (1 + \cos \omega t), \quad \ldots \quad (1)$$
$$V_k = \tfrac{1}{2} v V_0 \{1 + \cos (\omega t - \varphi)\}. \quad (2)$$

φ being the phase difference.

Having regard to the practical applications we also assume that the engine contains a certain dead space V_s, which we shall define as that part of the total space in the engine available for air which takes no part in the movements. We assume that the air in that space has an average constant temperature of T_s.

The variation of pressure as a function of time now follows from the condition that the mass of the work medium, in this case air, must be constant. We shall calculate this mass on the assumption that air is a perfect gas. In the hot, cold and dead space there is therefore:

$$\frac{M}{R} \cdot \frac{pV_h}{T_h} \text{ gram}, \quad \frac{M}{R} \cdot \frac{pV_c}{T_c} \text{ gram and } \frac{M}{R} \cdot \frac{pV_s}{T_s} \text{ gram.}$$

In this expression p is the pressure, M the average molecular weight of air and R the gas constant. The constancy of the mass of the work medium is now expressed by the equation:

$$\frac{M}{R} \cdot \frac{pV_h}{T_h} + \frac{M}{R} \cdot \frac{pV_h}{T_c} + \frac{M}{R} \cdot \frac{pV_s}{T_s} = C'.$$

We now replace the right-hand member by $M/R \cdot CV_0/2T_c$, where C is a constant. This method of notation simplifies the calculation.

In the expression obtained we now substitute the expressions for V_h and V_c.

After slight reduction this gives:

$$\frac{C}{2p} = \frac{(1 + \cos \omega t)\, T_c}{2T_h} + \frac{v\{1 + \cos (\omega t - \varphi)\}\, T_c}{2T_c} + \frac{V_s}{V_0} \cdot \frac{T_c}{T_s}.$$

We now introduce the following quantities:
τ, the temperature ratio: $\tau = T_c/T_h$;
s, the reduced dead space: $s = V_s/V_0 \cdot T_c/T_s$.
(In practice for example $T_h = 875°$ K and $T_c = 350°$ K, so that $\tau = 0.4$; s is usually about 0.5).
The equation thus becomes:

$$\frac{C}{p} = \tau \cos \omega t + v \cos (\omega t - \varphi) + \tau + v + 2s, \text{ or}$$
$$\frac{C}{p} = (\tau + v \cos \varphi) \cos \omega t + v \sin \varphi \sin \omega t + \tau + v + 2s.$$

For this we may now write:

$$\frac{C}{p} = \sqrt{\tau^2 + v^2 + 2\,\tau\,v\,\cos\varphi}\,\cos(\omega t - \Theta) + \tau + v + 2s,$$

where Θ is determined by:

$$\operatorname{tg}\Theta = \frac{v\sin\varphi}{\tau + v\cos\varphi}.$$

If we now introduce the abbreviations:

$$\sqrt{\tau^2 + v^2 + 2\,\tau\,v\,\cos\varphi} = A;$$

$$\tau + v + 2s = B \text{ and } \frac{A}{B} = \delta,$$

we find for p:

$$p = \frac{C}{B}\,\frac{1}{1 + \delta\cos(\omega t - \Theta)}.$$

From this it follows, that the maximum value of p is determined by:

$$p_{\max} = \frac{C}{B}\,\frac{1}{1-\delta}.$$

With the help of this last expression we may write for p:

$$p = p_{\max}\,\frac{1-\delta}{1 + \delta\cos(\omega t - \Theta)}. \quad \ldots \quad (3)$$

The variation of pressure found can be represented very simply in polar coordinates. Equation (3) is the equation of an ellipse with a focus at the origin, when p is chosen as radius vector and ωt as vectorial angle. This ellipse is shown in *fig. 3*.

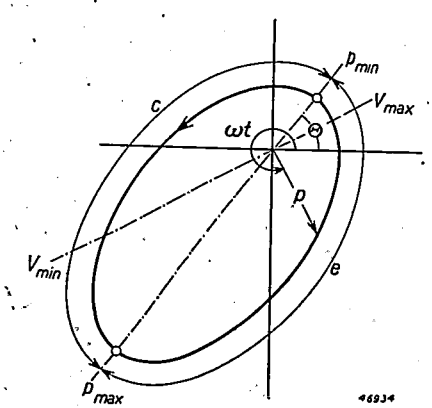


Fig. 3. Polar diagram of the variation of the pressure p in the engine (with sinusoidally moving pistons with a constant phase difference) as a function of the time t. The relation is represented as an ellipse with one focus at the origin. The vertices of this ellipse thus represent the highest and the lowest pressures, $p_{\max}$ and $p_{\min}$. The points corresponding to the positions where the total volume ($V = V_h + V_c + V_s$) is a maximum or a minimum lie on a line through the origin. The slope of this line with respect to the horizontal axis is smaller than the corresponding slope Θ of the axis of the ellipse. Thus the average pressure during the transition from $V_{\min}$ to $V_{\max}$ is higher than during the transition from $V_{\max}$ to $V_{\min}$. Arc e represents the expansion, arc c the compression.

The minimum and maximum pressures are reached when $\omega t = \Theta$ and $\Theta + 180°$, respectively. These values are represented by the vertices of the ellipse, which therefore lie on a line through the origin making an angle Θ with the zero line. In the figure it is also indicated for what value of ωt the whole work volume,

$$V = V_h + V_c + V_s$$

reaches a maximum or a minimum. These values can easily be determined with the help of formulae (1) and (2).

The corresponding points in the diagram also lie on a line through the origin, which, however, as follows from the calculation, makes an angle smaller than Θ with the zero line, provided $\varphi > 0$ and $\tau < 1$. This condition, as we already know, is always complied with in an air engine, and as a consequence the average pressure upon transition from $V_{\min}$ to $V_{\max}$ is greater than upon transition from $V_{\max}$ to $V_{\min}$. From this it immediately follows that in a full cycle of the engine positive work is gained.

The result obtained may be summed up as follows. An installation of two cylinders in open connection with each other, each closed by a piston with a closed quantity of air between the two pistons, one cylinder being kept at a constant high temperature and the other at a constant low temperature, acts as an air engine as soon as the pistons are made to move sinusoidally with a constant phase difference, with the variations in the volume of the hot cylinder preceding those in the cold cylinder.

Power produced

When a work medium goes through a cyclic process the work gained is in general given by the expression:

$$\oint p\,\mathrm{d}V. \quad \ldots \quad (4)$$

This quantity can be interpreted geometrically as the area of the closed curve which represents the cycle in the p-V diagram.

In our case $V = V_h + V_c + V_s$, and since V_s is constant, we may also write for the expression above:

$$\oint p\,\mathrm{d}V_h + \oint p\,\mathrm{d}V_c.$$

Each of these integrals may in the same way be conceived as the area of a closed curve in the p-V diagram. We must then plot, not the corresponding values of p and V, but those of p and V_h and of p and V_c, respectively.

The latter two integrals can also be interpreted physically, namely as the work performed separately by the hot and cold pistons respectively.

Formula (3), which represents the variation in pressure, in combination with formulae (1) and (2) for the variations of V_h and V_c, now enables us to construct the three p-V diagrams and to calculate the quantity of the work gained per revolution.

The three p-V diagrams are shown in *fig. 4* for the case that $V_0 = 2300$ cm³, $v = 1$ and $p_{max} = 40$ kg/cm².

the work per revolution by the number of revolutions per second $\omega/2\pi$, and we therefore obtain:

$$N = \frac{\omega}{2\pi} \oint p \, dV = \frac{\omega}{2\pi} \oint p \, (dV_h + dV_c).$$

If we substitute in this expression the values found for p, V_h and V_c (given by (1), (2) and (3) we obtain after simplification:

$$N = \frac{\omega}{2} p_{max} V_0 v (1 - \tau) \frac{\sin \varphi}{A} \frac{1 - \delta}{\delta} \left(\frac{1}{\sqrt{1 - \delta^2}} - 1 \right) \quad . \quad . \quad . \quad (5)$$

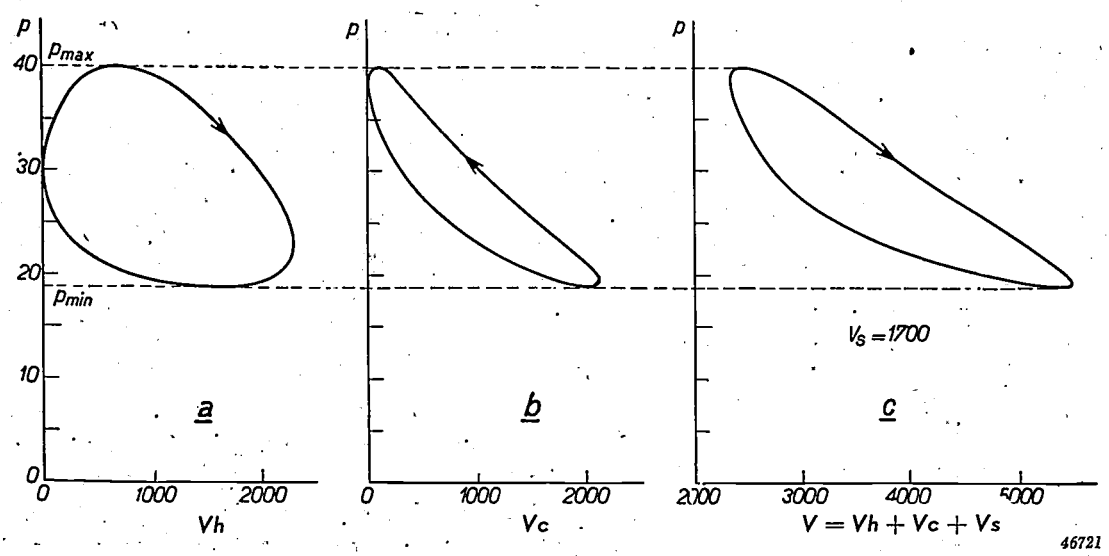


Fig. 4. p-V diagrams for an air engine with two pistons moving sinusoidally with a constant phase difference.
a) The diagram for the hot space (volume V_h).
b) That for the cold space (volume V_c).
c) The diagram for the whole working volume, (volume V, dead space, V_s).
$(V_h)_{max} = (V_c)_{max} = 2300$ cm³; $p_{max} = 40$ kg/cm², $\varphi = 90°$.
From the cyclic direction in the diagrams it may be seen that the hot piston produces work per revolution, that the cold piston consumes work in that time interval and that there is a total positive surplus of work per revolution.

Fig. 4*a* represents the diagram for the hot air, fig. 4*b* that for the cold air, fig. 4*c* finally giving the diagram for the whole working volume. From the cycle in the diagrams it can be seen that in case *a* positive work is done, and in case *b* negative work, while in case *c* (which is the sum of the amounts of work in *a* and *b*) it is again positive.

From the formulae given it is easily deduced that the surplus of work is only positive if the variations in the volume of the hot space are ahead in phase of those of the cold space. If this phase difference (the quantity φ in our calculations) were chosen equal to zero, the p-V diagrams would shrink to sections of curves and no work would be gained. If φ becomes negative the engine consumes work; this is applied in refrigerators, which are mentioned at the end of this article.

The power N, *i.e.* the work produced by the working medium per second, is obtained by multiplying

If all quantities are expressed in cgs units we of course obtain N in ergs/sec. By division by 7.36×10^9 N is obtained in the more customary unit HP. It has been found that the quantity N calculated in this manner gives a good approximation of the power actually produced by the air process. Of course this does not mean that the amount N thus obtained represents the effective power of the engine, because a not inconsiderable percentage is always consumed in the engine itself as a result of friction of the pistons and the other moving parts and from several other causes. In the case of small engines this may be around 25%, whilst in larger multicylinder engines it is less [1]).

[1]) Considerations about the variations of pressure and power analogous to those given above can be found in publications of Schmidt of 1862 and 1871, which refer to the models of air engines then known.

In *fig. 5* the power values calculated according to (5) are plotted as a function of the phase angle φ. (It must be kept in mind that this expression contains φ not only because the factor sin φ occurs in it, but because δ and A are also functions of φ.)

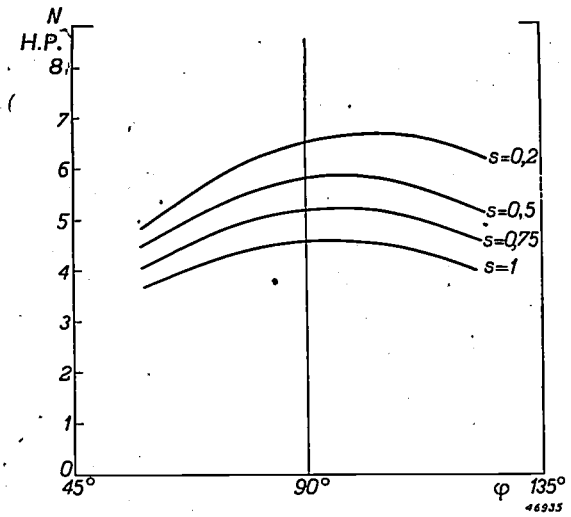


Fig. 5. Relation between the power N of an air engine (calculated on the assumption of isothermal compression and expansion) and the phase difference φ between the variations of the volumes of hot and cold space, for different values of the reduced dead space s. $V_0 = 1$ litre, $p_{max} = 10$ kg/cm², $n = 1000$ r.p.m., $v = 1$.

The diagram has been calculated for an engine with maximum pressure $p_{max} = 10$ kg/cm², a hot volume $V_0 = 1$ litre and a speed of $n = 1000$ r.p.m. It is further assumed, as is always approximately true in practice, that the hot and cold volumes are equal, thus $v = 1$. Since the power is proportional to p_{max}, V_0 and ω, the graph can also be used directly for engines with different values of those three parameters.

This diagram shows that the optimum value of the phase angle lies between 90° and 110°, while it is important not to choose the reduced dead space too large.

Regulation of the power produced

The formulae derived also enable us to realize how the power of air engines can be regulated. In the case of an engine of given construction and dimensions the maximum pressure occurring depends upon the amount of air taking part in the process. By varying that amount the maximum pressure can therefore also be varied. Now it is found from formulae (3) that when p_{max} is varied the pressure p at every moment changes in the same ratio and thus also the integral $\oint p\,dV$ and with that the power. From this it follows, that the power of the air engine can be varied by varying the amount of air taking part in the process.

The necessary heat Q and the thermodynamic efficiency

We shall now pass on to a discussion of the thermodynamic efficiency η of the air engine and, in doing so, take the opportunity to draw attention to several points of essential importance in the construction of such an engine.

We define the efficiency in the usual way as:

$$\eta = \frac{N}{Q},$$

where N is the power and Q the heat necessary per second. Here we will quote the following well-known Carnot theorem of thermodynamics: When a system passes through a reversible process, taking up heat from the outside only at the temperature T_h and giving off heat only at the temperature T_c, that process produces mechanical work with an efficiency of:

$$\eta = \frac{T_h - T_c}{T_h} = 1 - \tau \quad . \quad . \quad . \quad . \quad . \quad (6)$$

So far we have been assuming that the air in the hot space always has the temperature T_h, thus that the expansion is isothermal, and likewise that the air in the cold space is always at the temperature T_c and thus that the compression is also isothermal. Although this is not quite correct we shall for the present keep to these assumptions, so that then we may say that the efficiency of the engine would be represented by the above formulae as soon as the process becomes reversible. In the form, in which we have so far been discussing the engine, however, that will certainly not be the case. When, as is assumed in fig. 1, the air is pressed at a constant volume (isochoric) from the cold to the hot space (*II-III*) and then returned (*IV-I*), we are dealing with irreversible processes which are accompanied by loss of energy. As a result of these processes a certain amount of heat is transferred per revolution from the temperature T_h to T_c, which produces no work at all. Because of this the value of Q will be larger than in the theoretical case and η correspondingly smaller. What has been observed here for the schematic case of fig. 1 holds equally so for the engine of fig. 2, which more nearly approaches the practical case.

It is, however, possible to cause the transitions from the hot to the cold space and *vice versa* to take place reversibly, at least theoretically. For that purpose a so-called regenerator is introduced between the two spaces, in which the temperature of the air changes gradually from T_h to T_c as it flows through. In this way the exchange

of heat always takes place between bodies (namely the air and the respective part of the regenerator) which differ only very little in temperature. Thus the heat given off by the air remains stored in the regenerator and as the air flows back after half a cycle this heat is again used to bring the temperature of the air up to T_h. It is nowadays possible to make regenerators having an efficiency of 95% and more; by efficiency in this case is meant the percentage of the heat contained in the air inflow that is stored in the regenerator and given back to the air as it returns. The part of the heat, not stored, is carried off by the cooler and thus lost for the cycle.

If we now assume that the regenerator used is ideal and, moreover, disregard the flow resistances, we may say that the cyclic process applied is reversible.

In that case, therefore, the efficiency is given by the equation (6).

It is interesting to note that the process discussed, which clearly differs from the Carnot cycle, has, nevertheless, the same efficiency. We further call attention to the fact that the size of the dead space has no effect on the efficiency.

With the aid of the value found for η, the heat to be supplied to the work medium per second is determined by:

$$Q = \frac{N}{1-\tau}.$$

The heat to be dissipated is:

$$Q - N = \tau Q = \frac{\tau N}{1-\tau}.$$

This is the theoretically lowest possible amount. If the process is not reversible the heat to be dissipated is greater and the mechanical work and thereby the efficiency smaller.

We have already pointed out that the efficiency of the engine will be considerably lower than the efficiency of the process taking place in the engine, since owing to mechanical frictions the engine itself consumes an appreciable percentage of the power N furnished, while, moreover, the supply of heat always involves certain losses.

The internal process, however, even in the case of the engines already constructed, reasonably satisfies the equation $\eta = 1-\tau$, so that the internal efficiency can only be affected by the choice of the temperature of the hot and the cold spaces. In order to obtain a high internal efficiency it is important to provide that $\tau = T_c/T_h$ is as small as possible, which in practice means that T_h should be as high as possible and T_c as low as possible. For practical reasons, however, T_h is not chosen higher than 900 to 1000° K. T_c is usually around 300 to 350° K. Since the pressure does not occur in the expression for the efficiency it can be chosen quite independently, being dependent only on the dimensions of the engine and the properties of the materials used.

Until now we have expressly assumed that the air in the hot space is always at the temperature T_h and that in the cold space always at T_c. In order to accomplish this as nearly as possible it is not sufficient to keep the walls of the two spaces at the temperatures prescribed, transfer of heat from the wall to the air not being sufficient for that. A good air engine must therefore be provided with a specially constructed heater and also a specially constructed cooler. (In their construction provision has to be made on the one hand for an efficient and rapid heat transfer to the air, while on the other hand the flow resistance may not be too large, since it consumes power.) The heater and cooler are placed between the hot space and the regenerator and between the cold space and the regenerator, respectively (see also fig. 2), and their construction, to which we shall return in a later article, is such that after passing through them the whole volume of air is at the temperature T_h or T_c respectively. This arrangement guarantees that the air always flows into the hot space at the temperature T_h and into the cold space at T_c.

This, however, does not ensure that the air also remains continuously at those temperatures. Let us consider for example the expansion in the hot space. In a high-speed engine the expansion takes place within a fraction of a second and consequently the parts of the air at some distance from the wall will not receive the necessary heat quickly enough and will therefore drop in temperature. (Imagine here, for example, that the heater is a ring surrounding the hot space, so that heat may be fed to the work medium also through the wall of the hot space.) These parts of the air will not expand isothermally, but approximately adiabatically. Similarly for the compression in the cold space, the air there will assume a temperature higher than T_c. These adiabatic processes mean that the whole process is no longer reversible. At the end of the expansion the air in the hot space has an average temperature lower than T_h. As this air leaves the hot space again it passes the heater, which is at the temperature

T_h: thus it passes through a finite temperature difference. The same is true for the cold space, *mutatis mutandis*.

The deviation from reversibility has the result that the actual efficiency is no longer equal to the theoretical efficiency given by equation (6). The effect is not large, however, (of the order of 10%), for two reasons. In the first place the decrease in temperature resulting from the adiabatic expansion is partly compensated by the fact that during the time of the whole expansion air at the temperature T_h still flows from the heater into the hot space. While it is true that this mixing of air of the temperature T_h with air of an average lower temperature is itself an irreversible process, it, nevertheless, reduces the magnitude of the effect, because the latter is determined entirely by the fall of temperature taking place. (Here again a quite analogous situation obtains in the cold space.)

The second reason why the effect is slight lies in the relatively low expansion ratio ($p_{max} : p_{min}$) which is employed in air engines. It lies between 2 and 2.5.

We shall not go here into the quantitative determination of the effect, for this leads to rather complicated calculations.

This discussion will now be concluded with a brief summing up of what goes on in the motor with respect to energy. For the sake of simplicity we again assume that in the hot space the temperature T_h always prevails and in the cold space T_c. For the isothermal expansion of the hot air heat is necessary, and this is supplied by the heater. In the regenerator the air is cooled, but the heat is stored and used again later for heating the air; it may therefore be left out of consideration in the balance of energy. In the expansion all the heat taken up is converted into mechanical energy. In the compression a part of it is used and again converted into heat, which is given off in the cooler. The net result of all the changes is that heat is continually being taken up from the heater by the work medium and partly converted into mechanical work and partly given off to the cooler.

The air process applied in a refrigerator

As already noted, the direction of the air cyclic process can be reversed. This means that it can also be made to take place in such a way that upon supplying a certain mechanical power $N = (1-\tau)\,Q$ per sec. an amount of heat τQ is taken from a reservoir at the low temperature T_c and an amount of heat Q is given off to a reservoir at the high temperature T_h. This reversion can be accomplished by choosing the phase difference φ already mentioned negative. Work is then not only taken up, as we have already noted, but the flow of heat also changes direction, *i.e.* heat is now removed from the cold space and given off to the hot space. This takes place with the same high efficiency that characterizes the air process in the engine. It is clear that the air process in this way functions as a refrigerator, and it has been found in practice, that this can in fact easily be realised. When the elements corresponding to heater, regenerator and cooler are given suitable dimensions very satisfactory refrigerators are obtained. In an experiment in which hydrogen was used as work medium a temperature of 80° K was attained in one stage.

CARRIER SUPPLY IN AN INSTALLATION FOR CARRIER TELEPHONY

by D. GOEDHART and G. HEPP. 621.395.44

In an installation for carrier telephony AC voltages of frequencies, which are whole multiples of 4 kc/sec, have to be supplied to a number of apparatus in the different telephone channels. The requirements made of these "carrier voltages" as to constancy of frequency and amplitude, freedom from distortion, etc. are explained in this article. For example the carrier frequencies for one and the same channel in the transmitting and in the receiving stations may differ by not more than about 1 c/sec. All the carrier frequencies can be excited collectively by means of circuits which give a highly distorted AC voltage, for example a periodic impulse with a fundamental frequency of 4 c/sec. This voltage also contains all the harmonics of 4 kc/sec, so that the desired carriers can be selected by a set of filters. The fundamental frequency of the impulse generator is kept contant by a "master oscillator", while the equality of the fundamental frequencies in the transmitting and in the receiving stations is ensured by synchronization of the two master oscillators by means of a transmitted synchronization signal. In the discussion of a carrier supply equipment designed by Philips a discussion is given of the circuits of the impulse generator, of the choice of impulse width, of the avoidance of frequency division in the impulse generator and of the manner of connecting the series of carrier filters with the impulse generator.

In ordinary long distance telephony the alternating currents from the microphone, which contain the speech frequencies of about 100 to 4000 c/sec, are amplified and transmitted directly over a pair of conductors of the cable between transmitting and receiving stations.

In carrier telephony the low-frequency speech vibrations are first modulated in the transmitting station, on a carrier of higher frequency, in much the same way as in radio broadcasting. By using a series of carriers with different frequencies, each of which "carries" a call (provides a telephone channel), a large number of calls can be transmitted simultaneously over each pair of conductors.

Systems have been developed with 4,12 or even more carriers (channels). A system designed by Philips before the war works with 17 channels. Compared with ordinary telephony, carrier telephony has the advantage that for a given number of simultaneous calls a much smaller number of pairs of conductors is needed, which means an appreciable saving of copper in the cables, although at the expense of a more complicated apparatus in the terminal stations.

The fundamentals of carrier telephony and several component parts of the apparatus employed were dealt with more or less extensively in this periodical in the years 1940-1942 [1]). Among the details which were not discussed at that time was the manner in which the necessary carriers are excited. This forms the subject of the present article. Other components of a carrier-telephony installation will be discussed in articles to be published in subsequent numbers of this periodical.

With respect to the examples to be discussed we shall keep to the apparatus already developed, to which the articles mentioned above apply [1]), *viz.* the 17 channel system with so-called single modulation. It has to be pointed out, however, that in the course of the past few years considerable developments have been achieved in the Philips laboratories with respect to carrier telephony. Systems for a large number of channels have been designed which differ considerably from those so far dealt with, both fundamentally and constructionally. The new developments, which have been made possible mainly because of the fact, that improved materials and component parts have meanwhile become available, will be dealt with in a new series of articles to follow the present ones.

For the sake of a better understanding of what is to be discussed here, it is desirable first to recall the most important features of a carrier-telephony installation. In *fig. 1* the main parts of the apparatus for two channels between the stations *A* and *B* are indicated. The microphone currents with the frequency q arriving from a subscriber *A* at station *A* are sent through a low-pass filter (*LFZ*), which, roughly speaking, has the task of suppressing the frequencies above 3400 c/sec, to the modulator

[1]) See for example Philips techn. Rev. 6, 325, 1941 (Fundamentals); 7, 83, 1941 (modulators); 7, 104, 1942 (filters); 7, 184, 1942 (equalization). (Volume 7 not yet published in English).

(*Mod.*), to which at the same time an AC voltage with the higher frequency p_1 (the carrier) is applied.

The modulator then delivers an output voltage which contains among other frequencies the "side-bands" $p_1 + q$ and $p_1 - q$ of the carrier. Following the modulator is a bandpass filter (*BFZ*), whose attenuation as a function of the frequency is of such a character that practically only one side-band is passed, thus only $p_1 + q$ if it is desired to use the higher side-band. The AC voltage thus obtained, which only contains components within the frequency band $p_1 + q$ (from about $p_1 + 300$ to about $p_1 + 3400$ c/sec), is combined with the similarly obtained voltages in the frequency bands $p_2 + q$, $p_3 + q$, ... of calls of other subscribers, and after amplification put on a pair of conductors of the cable. At the other end of the cable, at the station B, a series of bandpass filters (*BFO*) is connected in parallel on the same pair of conductors, each of which filters passes only one of the frequency bands $p_1 + q$, $p_2 + q$, $p_3 + q$, ... In this way the different channels are separated again. The voltages $p_1 + q$ of the call of subscriber A are passed by the corresponding band filter and applied to the demodulator (*Dem*). This apparatus is identical with the modulator at the transmitting end, and, as in that case, the carrier p_1 is here added. This results in a voltage at the output of the demodulator with the frequencies $p_1 - (p_1 + q) = -q$, which are the original low-frequency speech vibrations. In a final low-pass filter (*LFO*) these are freed of any undesired frequencies still present and finally passed on to the subscriber B_1.

The microphone currents from B_1 are transmitted to A_1 in exactly the same way over a different pair of conductors. As will be seen, a series of AC voltages p_1, p_2, p_3, ... must be available at both stations for supplying the modulators and demodulators. The apparatus serving for the excitation of these carrier voltages must satisfy the following main conditions:

1. The frequencies of the carriers must be exactly the same at the transmitting station as at the receiving station, and must be extremely constant;
2. the carrier voltages must be free of harmonics and other undesired components;
3. the energy available must be sufficient for supplying the numerous modulators, demodulators and possible other apparatus [2]) which may be located at the station;
4. the amplitude of the carriers must not vary too much with time;
5. the apparatus must have the greatest possible reliability when in operation.

In the following we shall go deeper into these

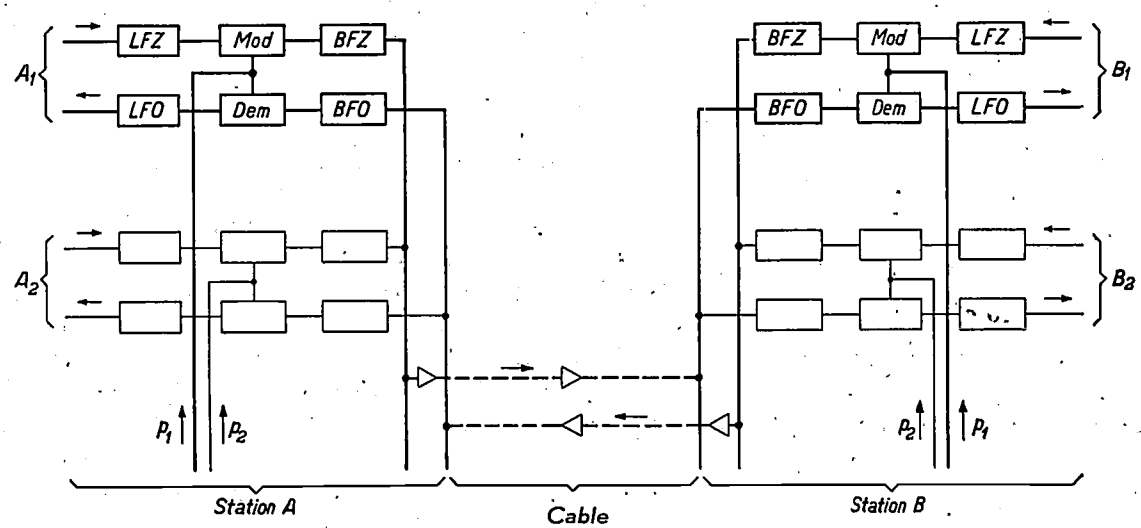


Fig. 1. Diagram of a carrier telephony connection between two stations A and B, drawn for two channels. The parts of one channel (with the carrier p_1) in which a call is taking place between the subcribers A_1 and B_1 are drawn with a heavy line. *LFZ* low-pass transmitting filter, *Mod* modulator, *BFZ* transmitting bandpass filter, *BFO* receiving bandpass filter, *Dem* demodulator, *LFO* low-pass receiving filter. At p_1, p_2, the carriers for the various channels are applied.

[2]) For example the carrier voltage in each channel can also be used for signalling, *i.e.* for the transmission of the dialling signals bringing about the connection with the desired subscriber.

conditions and explain how they are satisfied in practice. After that we shall discuss the most important parts of an apparatus constructed by Philips.

Stability of the carrier frequencies

Requirements

It is easy to understand why the carrier frequencies have to be exactly the same at the transmitting and receiving ends. Suppose that a call (frequencies q) is modulated in station A on a carrier p_1, but in station B it is demodulated with a slightly different carrier frequency $p_1 - \Delta p$. At the receiving end, instead of the original frequency q, the frequency $q + \Delta p$ is obtained: all the frequencies q of the speech spectrum are shifted by the same amount Δp. This shift in the spectrum is accompanied by a change in sound, which in the case of speech results in the first instance in a change in the individual voice so as to make it unrecognizable, while with larger shifts the speech even becomes unintelligible. Experiments have shown that a shift of ± 10 c/sec must already be considered as prohibitive. Music is even much more sensitive to a frequency shift: it looses its harmonic character and becomes dissonant [3]), a frequency shift of less than 1 c/sec already being disturbing. Since in carrier telephony installations the transmission of music has indeed often to be taken into account, for example for the connection of a broadcasting studio with a concert hall or with radio distribution centres, where channels with frequency bands two or three times as broad are reserved, the difference of about 1 c/sec is therefore the maximum difference permissible between the carriers in two stations.

This same high degree of accuracy is required both for the lowest and for the highest carrier frequencies employed. The necessary relative precision is therefore greatest in the case of the highest carriers; for example with a carrier of 68 kc/sec it amounts to about 10^{-3} percent. In order to comprehend what this means, imagine that this AC voltage were used for running electric clocks: the two clocks should then not differ by more than about one second per day.

Method of supplying the carriers

How then does one set to work in order to ensure this agreement between the carrier frequencies in two stations? It would be technically possible to use for the excitation of each carrier an oscillator whose frequency is kept extremely constant by some suitable construction (among other methods, by placing the frequency-determining element in a thermostat). It would then suffice to check the oscillators at regular intervals with a kind of "tuning fork" and to readjust them. Where we have a large number of channels, however, this becomes too laborious, so that other methods have had to be sought. According to recommendations of the C.C.I.F. (Comité consultatif international de téléphonie) whole multiples of 4 kc/sec should be chosen for the carrier frequencies. Thus for example in the 17-channel system mentioned the first 17 harmonics of 4 kc/sec were taken for the carriers: 4, 8, 12, . . . 60, 64, 68 kc/sec. Now, since a strongly distorted AC voltage with the fundamental frequency 4 kc/sec in general also contains all the harmonics thereof, it seems obvious to supply all the carriers mentioned collectively in each station by generating one such strongly deformed AC voltage and then separating the components by means of filters. In order to obtain constant carrier frequencies it is then only necessary to keep the fundamental frequency of the composite AC voltage sufficiently constant. This trouble, however, can also be saved, since in the first instance it is not actually a question of keeping the carrier frequencies absolutely constant, provided they are identical in the transmitting and receiving stations. If provision is made, by the transmission of a suitable synchronization signal, that the fundamental frequency of the distorted AC voltage is exactly the same in the two stations, then all the carriers of the two stations automatically correspond.

The apparatus which excites the AC voltage mentioned with all its harmonics will be dealt with further on. Here we would only observe that this generator of harmonics is kept at the correct fundamental frequency by a "master oscillator" tuned accurately to 4 kc/sec. From this generator the signal which is sent over the cable to synchronize the corresponding master oscillator in the other station is also derived. As synchronization signal the AC voltage of 4 kc/sec itself is not used, nor a harmonic of it which coincides with one of the other carriers, because owing to the modulation of these carriers with speech this would lead to interferences. It is preferable to use an unmodulated harmonic of 4 kc/sec that is higher than the highest carrier employed in the installation.

Summarizing we may represent the method

[3]) It may be expressed by saying that the ear has not an arithmetic, but a logarithmic perception of pitch. Thus upon transposing a piece of music all the frequencies should not be shifted by the same amount, but multiplied by the same factor. Experiments on the effect of frequency shifts have been described by J. F. Schouten, The perception of pitch, Philips techn. Rev. 5, 286, 1940.

outlined for the excitation of the carriers by a diagram as given in *fig. 2*.

When by means of the synchronization of the master oscillators the equality of the carriers in two stations has been ensured, a drift of the fundamental frequency is in itself not disturbing [4]). Nevertheless, in case the synchronization should fail, it is desired to be able at least to transmit speech intelligibly. Therefore measures are still taken to keep the frequency of the master oscillator in each station satisfactorily constant (see the thermostat in fig. 9), and it is regularly checked and readjusted so that even without synchronization the highest carrier in each station can never shift more than a few c/sec.

carrier voltage p, after being filtered out, is first fed to an amplifier capable of supplying the necessary power for all the modulators, demodulators, etc. of the channels working with p. In this amplifier distortion may occur to a larger or smaller degree, with the result that in addition to p, the harmonics $2p$, $3p$, etc. are also present in the output voltage. Finally there is the fact, that all the carrier filters, amplifiers, etc. are assembled in the same bay, so that if there is an undesired coupling in the apparatus one or more of the other carriers can be induced on the circuit for the carrier p.

Such a "contamination" of the carriers may lead to very undesirable disturbances. If a weak, unwanted voltage with the frequency P reaches

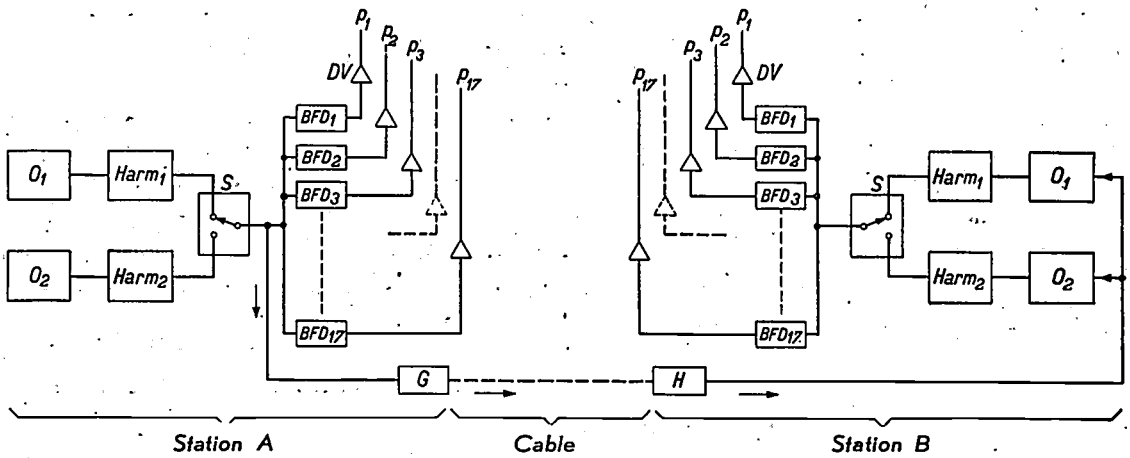


Fig. 2. Diagram of a carrier supply equipment, for instance for 17 channels at transmitting and receiving stations. O_1 master oscillator with reserve oscillator O_2, both tuned to 4 kc/sec; $Harm_1$ generator of harmonics with reserve generator $Harm_2$, $BFD_{1\text{-}17}$ carrier filters, DV carrier amplifiers; S automatic changeover; G filter for the synchronization signal, whose frequency lies above the highest carrier frequency used; H apparatus with which a frequency of 4 kc/sec is derived from the synchronization signals and fed to the oscillators O_1 and O_2.

Purity of the carrier frequencies

For various reasons the carrier voltage with the frequency p will never be quite purely sinusoidal. In the first place the carrier filters which have to select the different carriers (*BFD* in fig. 2), in addition to their "own" carrier, also transmit to a very small extent the adjacent carriers $p + 4$ and $p - 4$ and more distant ones. In the second place, in large stations from which many pairs of conductors go out, each with a complete carrier system, each

the modulator in addition to the desired carrier p, upon modulation with the speech frequency q, the bands $P + q$ and $P - q$ will also occur in addition to the desired sideband $p + q$. If P is the adjacent carrier $p + 4$, $P + q = p + 4 + q$ is suppressed by the transmitting band filter. $P - q = p + 4 - q$ however, falls in the same frequency region as $p + q$.

This modulation product is thus passed in the normal way by the transmitting and receiving band pass filters of the channel and after demodulation with the carrier p gives rise to a frequency $4 - q$ at the receiving end. All the speech frequencies q of the channel are therefore transmitted in the same channel again with "inversion", at $4 - q$, which results in a disturbing unintelligible noise. With some systems a disturbance may arise when P is a carrier situated further away from p, that is to say in the event

[4]) The condition also holds that the frequency spectrum may not be shifted too much with respect to the frequency characteristic of the band filters, since otherwise the highest or the lowest frequencies are more strongly attenuated than is permissible. But the tolerance in this case (about 30 c/sec shift) is still much greater than the one actually employed, which will be explained in the following.

that the transmitting bandpass-filters produce ample attenuation only immediately on either side of their band of transmission, with but little attenuation at greater frequency distances [5]). In such a case the modulation products $P + q$ and $P - q$ will be relatively little attenuated when they reach the cable. These frequencies are then passed at the receiving end by the receiving band filters of the channels with carrier P and $P - 4$ in the normal way, and after demodulation give the original speech frequency q respectively the inverted speech frequency $4 - q$ belonging to the channel p. While the latter frequency in the channel $P - 4$ again results in a disturbing unintelligible noise (unintelligible cross talk), the occurrence of q in the channel P, in addition to the frequencies Q of the call belonging to P, leads to an intelligible reproduction of the call intended for channel p. This "intelligible cross talk" is worse than the unintelligible sort, since it not only causes a disturbance of the call in channel P but also endangers the necessary secrecy of telephone calls.

Experience has shown that all these disturbing effects can be sufficiently restricted if care is taken that the admixtures with each carrier lie at least 60 db below the level of the carrier itself. This is obtained by the use of carrier filters, with sufficiently sharp cut off, by efficient shielding of the parts in the carrier supply bay and by the use of carrier amplifiers with extremely slight distortion. Amplifiers with strong back coupling are used.

Constant amplitude

If the amplitude of the carrier voltage on a modulator changes, and if the modulating voltage is small compared with the carrier voltage, the modulated voltage obtained will vary only slightly in intensity (in first approximation not at all). The fluctuations which occur normally in the carrier amplitude as a result of variations in mains or battery voltages or of the ageing of valves, can therefore be tolerated without difficulty as long as they do not exceed a value of for instance 10 percent. The effect upon the carrier amplitude of a variation in the loading of the carrier supply equipment — for instance when fewer channels are in use at night than in the daytime — can be limited sufficiently by keeping the internal impedance of the outputs of the carrier supply equipment small.

Reliability during operation

Serious consideration must be given to the possibility that a carrier voltage, or even the whole carrier supply may fail due to a disturbance. If one carrier fails, only the channel (or channels) working with that carrier is put out of operation. Provision must be made for an alarm signal to warn the operators immediately, so that the defect can be repaired. If, however, the whole carrier supply fails, the transmission on all the channels of all the pairs of conductors comes to a standstill. For such a calamity it is not enough to warn the operators. Therefore, in addition to all the precautions, which are taken to minimize the chance of such a disturbance, special circuits are provided which upon the failure of a vital piece of apparatus automatically and without any interruption switch over to a reserve apparatus. This holds in particular for the master oscillator and the generator of harmonics in the above apparatus (fig. 2). The only other reserve needed here is a single reserve amplifier for all carrier amplifiers. The fact that so little reserve apparatus is necessary is an additional advantage of the method of combined excitation of the carriers over the use of separate oscillators, in which a separate reserve oscillator would have to be provided for every wave length.

The practical construction of an apparatus

In the discussion of the construction of an apparatus for carrier supply we shall limit ourselves mainly to the most essential parts, namely the generator of harmonics and the filters.

Choice of the form of voltage to be excited

There are all kinds of distorted AC voltages containing all the harmonics of the fundamental frequency. The choice of the form of voltage to be excited however, is limited by the desire, that the different harmonics should not differ too much in amplitude. If one of the harmonics has a much stronger neighbour a very complex and expensive filter is needed for that harmonic in order to suppress this neighbour sufficiently. This holds especially for the highest harmonics where the neighbouring frequencies are relatively closest together.

A very suitable form of voltage is the periodic impulse, see *fig. 3*. If this is very narrow it contains all the harmonics in practically the same strength, or better, in the same weakness, because with impulses of short duration the power is, of course, only small. With a greater width of impulse the amplitude, at least that of the lower harmonics, increases proportionally, but at the same time with increas-

[5]) The attenuation necessary in these frequency regions in the channel p is provided by the low-pass filter *LFO*.

ing order of harmonics the amplitude shows a steadily more pronounced decrease [6]). If one considers especially the highest harmonics to be used as carrier, it will be seen that, under the influence of the two opposite effects they will at

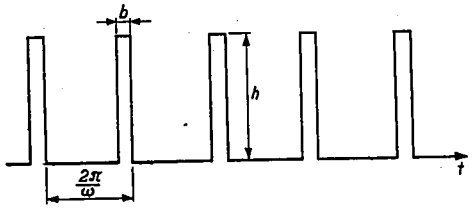


Fig. 3. Periodic impulse from whose harmonics the desired carrier frequencies can be filtered out. It is characterized by the period $(2\pi/\omega)$. the impulse width b and the impulse height h.

first become stronger and then weaker with increasing width of impulse (*fig. 4*). That width of impulse at which the amplitude of the highest harmonic assumes its maximum value is the most suitable one for our purpose, because in the first place that harmonic is in any case the weakest of the series, and in the second place it is most necessary for that harmonic to obtain a sufficiently large amplitude, since in the corresponding carrier filter the transmission region must be relatively the narrowest and therefore the attenuation in the transmission region is the strongest.

The above-mentioned optimum width of impulse can easily be calculated. If the height of the impulses is h and if a is the relative width of impulse (*i.e.* the duration b of an impulse divided by the period $2\pi/\omega$), then the variation of the voltage $f(t)$ of the periodic impulse can be given by the Fourier series:

$$\frac{\pi}{2h} \cdot f(t) = \frac{\pi a}{2} + \frac{\sin \pi a}{1} \cos \omega t + \frac{\sin 2\pi a}{2} \cos 2\omega t + \ldots$$

If m is the order of the highest harmonic to be used, we want to choose a such that the amplitude $\sin m\pi a/m$ of this harmonic is a maximum. The condition for this is plainly that $\sin m\pi a = 1$, therefore $a = {}^1/_2\, m$.

It should be pointed out that with this width of impulse the desire for practically equal amplitudes for all the harmonics is also very satisfactorily fulfilled. For the highest harmonic the amplitude is $1/m$, for the lowest $\sin (\pi/2m) \approx \pi/2\, m$. The ratio between these amplitudes is $\pi/2$, *i.e.* the largest difference in level in the series of carriers amounts to only about 4db.

[6]) *Cf.* for example: J. F. Schouten, Synthetic sound, Philips techn. Rev. 4, 167, 1939.

Circuits of the impulse generator

All kinds of circuits can be used to excite periodic impulses. One of the simplest, the principle of which is also used in the Philips 17-channel system already mentioned, is shown in *fig. 5*. The circuit includes an amplifier valve V strongly back-coupled with the transformer T_1. Thus when anode current begins to flow through its primary winding the transformer gives the control grid of the valve a positive voltage. Due to the positive grid voltage the anode current becomes stronger, the grid becomes still more positive, the anode current increases still more, and so on. The resulting very rapid growth of the anode current may be compared with the breakdown in gas discharge. At a given moment the anode current can no longer

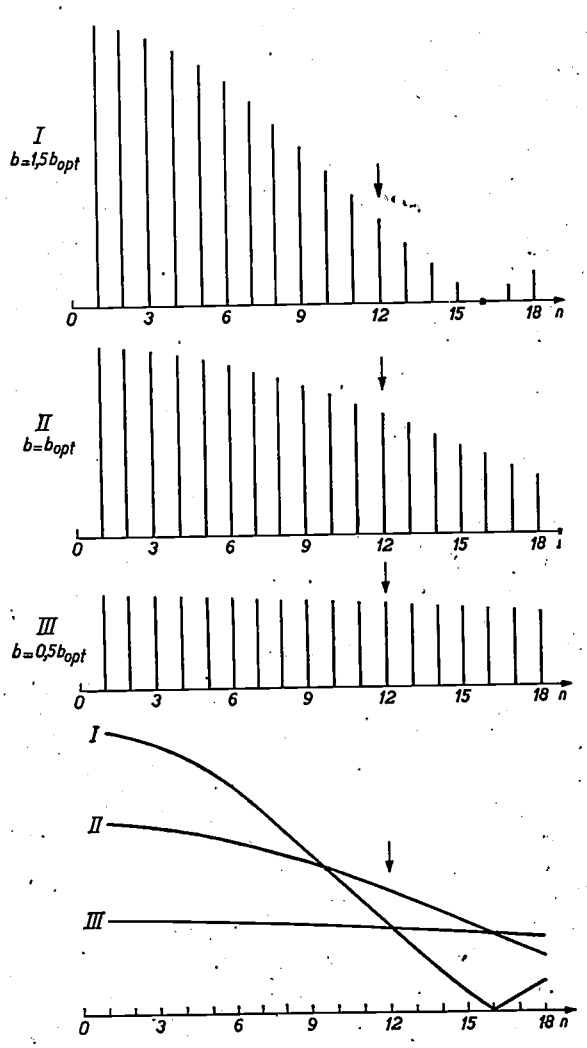


Fig. 4. Frequency spectrum of a periodic impulse with a given fundamental frequency and height. In the upper three figures the spectrum for three different values of the impulse width b is drawn, while in the lowest figure the envelopes of these three spectra are shown. It is clear that each harmonic assumes a maximum amplitude at a given impulse width. The optimum width for the 12th harmonic is here indicated by b_{opt}.

increase, since it is limited by the properties of the valve. When, however, the increase in the anode current ceases, the positive grid voltage delivered by the transformer fails, the anode current begins to fall and the transformer then delivers a negative

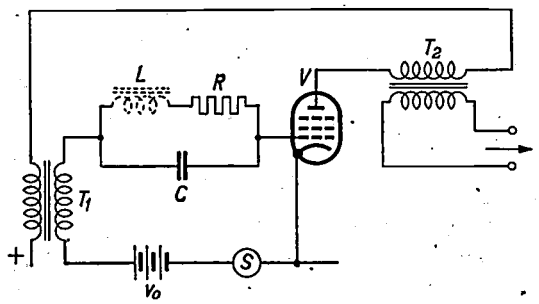


Fig. 5. Circuit for the excitation of periodic impulses. Due to the back-coupling *via* the transformer T_1, the amplifier valve V delivers an anode current in the form of short impulse. The "blocking" condenser C with the leakage resistance R provides that these impulses follow each other with approximately the correct period, while the period is accurately regulated by a synchronization voltage S. The periodic impulse obtained is supplied to an output valve *via* the transformer T_2.

voltage to the grid which accelerates the fall and in a short time causes the anode current to cease entirely. The anode current thus behaves as an impulse, whose width can be regulated for instance by a suitable choice of the transformer selfinductance.

Without the "blocking condenser" C and the leakage resistance R indicated in fig. 5 a second impulse would immediately follow the first one and so on. During the time that the grid was positive, however, the condenser was being charged by the grid current. After the interruption of the transformer voltage making the grid positive, the charge on the condenser causes a negative grid voltage which continues even after the end of the impulse and keeps the valve "blocked" until the greater part of the charge has been equalized over the leakage resistance. Then a new impulse follows, with a renewed charging of the condenser, this process repeating itself at a frequency mainly determined by the values of R and C. But the properties of the valve and the supplying voltage also play a part here. The fundamental frequency of the periodic impulse obtained in this way, which can be applied to the carrier filters *via* the output transformer I_2, is of itself not yet very constant. In order to obtain the required high degree of constancy, the impulse generator is synchronized by the above-mentioned master oscillator, which gives a very constant frequency of 4 kc/sec. For this purpose that oscillator voltage, together with a strong negative bias

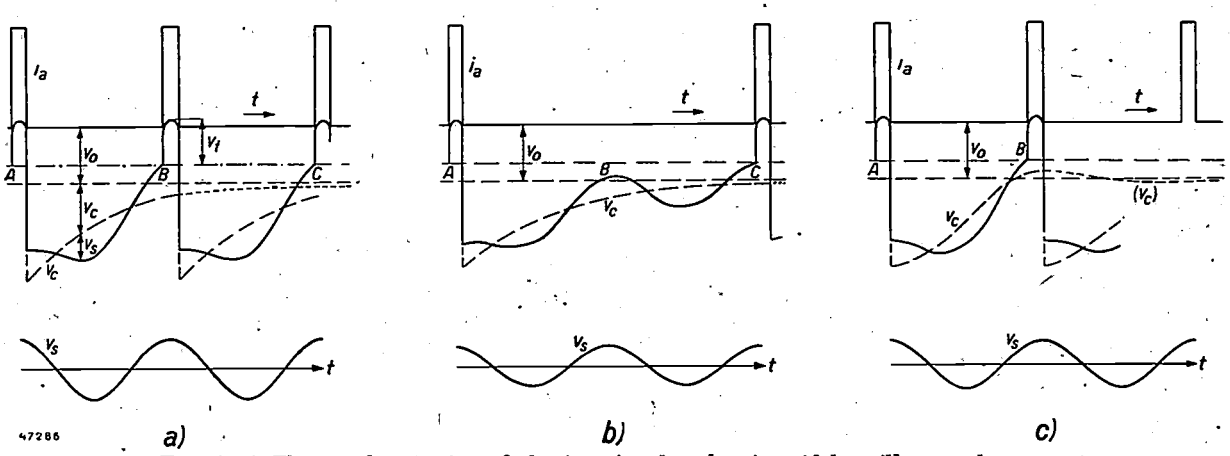


Fig. 6. *a*) The synchronization of the impulses by the sinusoidal oscillator voltage v_s drawn below. Below the time-axis (in the upper figure) the grid voltage of the valve V in fig. 5 is plotted, above the axis the anode current. The negative grid bias v_0 prevents the valve from "breaking down", even when the blocking voltage v_c of the blocking condenser has entirely disappeared. By the superposition of v_s on v_c, the bias v_0 is exceeded at each positive peak of v_s (for instance at B), and the valve is free to function. The dot-dash line indicates the level of the grid voltage at which the valve breaks down. v_c is the grid voltage contribution occurring at that moment from the transformer T_1, in fig. 5. The part of the curve v_c indicated by dots would only be traced if no break-down occurred.

b) The occurrence of frequency division. If v_s is not large enough, the breakdown at the desired moment B fails to happen and occurs only in the next (C) or a still later cycle of the synchronizing voltage.

c) The suppression of frequency division. By including the self-induction L, indicated by a dotted line, in the circuit of fig. 5 the "blocking" voltage v_c assumes the character shown here; at the moment B it does not block, but promotes the breakdown of the valve. If v_0 is so large or v_c so small that the moment B still passes without breakdown, the impulse generator fails entirely and the reserve generator goes into action.

is applied to the grid of the amplifier valve. The bias, prevents the valve from "breaking down" of itself; only when that voltage is compensated by the synchronizing voltage, thus at the positive peaks of the latter, can "breakdown" occur, so that it is forced into the rhythm of the synchronization voltage (see *fig. 6a*).

The avoidance of frequency division

A familiar phenomenon in such circuits is the occasional occurrence of frequency division. With insufficiently high synchronization voltage the breakdown at the prescribed moment sometimes fails to occur, owing to the fact, that the necessary level has not been reached, and it then occurs only in the next period, because on account of a further discharge of the blocking condenser the level has then risen somewhat (fig. 6*b*). In that case an impulse is obtained only once in two cycles of the synchronization voltage, and with still lower synchronization voltage sometimes only once in three, four or more cycles.

When the circuits are used for carrier supply this phenomenon is very undesirable. With frequency division of 1 to 2 the impulses would have a fundamental frequency of 2 kc/sec instead of 4 kc/sec, and the voltage would be too low for all carriers, but particularly the filters would be unable to suppress the extra harmonics lying at a distance of only 2000 c/sec away, so that a loud whistling tone would occur in all channels.

In order to prevent this, it is possible to introduce, in series with the leakage resistance R, a self-inductance L (shown with a dotted line in fig. 5) of such dimensions that "over discharge" (opposite charging) of the blocking condenser occurs at the moment, at which the following breakdown should occur. The "blocking" voltage thus changes its sign at that moment and promotes the breakdown. If in spite of this there is no breakdown, at the following positive peak of the synchronization voltage there will be even less chance of breakdown since then the "blocking" voltage, which has in the meantime fallen again, no longer helps, or at least only to a less extent (see fig. 6*c*). The valve thus remains blocked, *i.e.* the impulse generator fails entirely. This is much less serious than if it continued to function with frequency division, because upon its failure the reserve generator is brought into action by the automatic switching device, and the installation continues to operate without disturbance, while at the same time the operators are warned by an alarm signal of the defect in the generator thrown out of action.

Connection of the carrier filters to the impulse generator

In order to obtain the necessary carrier energy, the impulse voltage excited by the circuits already described is amplified by an output pentode in whose anode circuit the series of carrier filters is included. Each separate harmonic selected by one of the filters is then amplified again in the previously mentioned carrier amplifier and can then be applied as carrier to the various apparatus of the corresponding channel.

The connection of the filters to the output pentode mentioned will now be considered somewhat more closely.

The impulse voltage is applied to the control grid of the pentode in such a way that during each impulse the latter reaches zero potential, and in the intervals is strongly negative. A large anode current then flows through the loading impedance (the series of carrier filters) only during the impulses, while in the intervals the valve can deliver no current. This amounts to the same thing as if between the impulses the carrier filters were cut off by a switch from the source. A peculiar difficulty arises from this, which we shall attempt to explain. For the sake of simplicity we may consider the carrier filters as a series of single

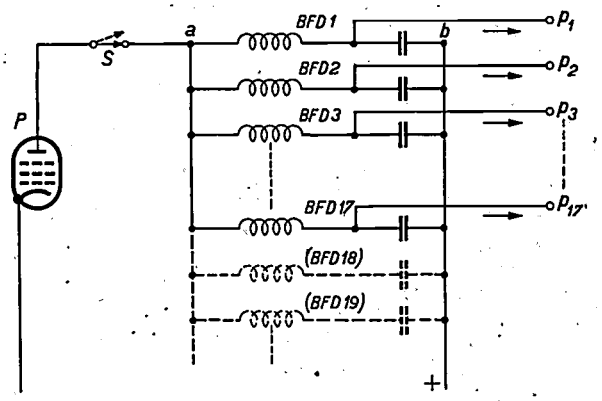


Fig. 7. Circuit of the series of carrier filters (BFD_1 to BFD_{17}), which are here represented simply as ordinary L-C-series circuits, with the output pentode P. The latter behaves as a source of current, which is switched on during the voltage impulses and off during the intervals by a switch S. The filtered carrier voltages are drawn off at p_1, p_2, ...

L-C series circuits, tuned to 4, 8, 12, ... kc/sec, each with a resonance resistance R and all connected in parallel, see *fig. 7*. One might also take a series of L-C parallel circuits connected in series, but in practice it is preferable, for instance because of the greater reliability, to work with elements connected in parallel, thus series circuits. (Actually of course each filter is very much more complicated than a single series circuit.) Now suppose that we wish to select only one carrier frequency of the impulse voltage, and thus include only one such series circuit in the anode circuit of the pentode. The action of that circuit then comes to nothing. A sinusoidal current should flow through the circuit, but during the greater part of the time, namely in the intervals between the impulses, no current at all can flow, since, as we have seen, the filter is then as it were part of an open circuit. If, however, we connect the whole series of filters in parallel in the anode circuit the situation is different. If the resonance resistance R

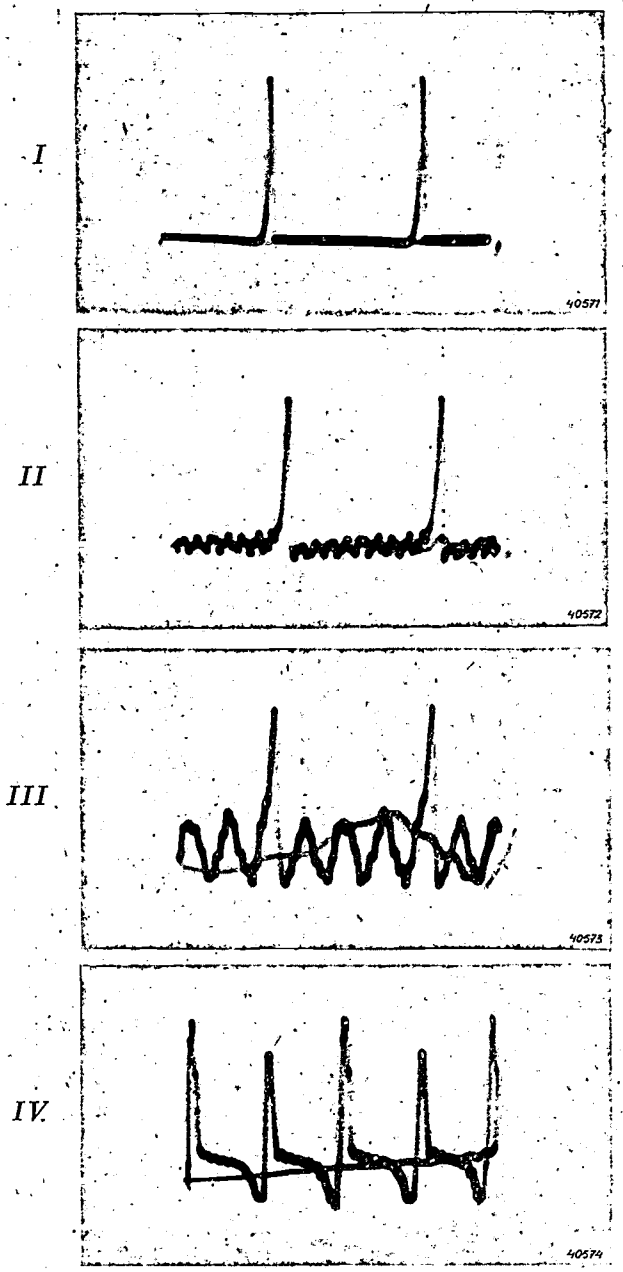


Fig. 8. Oscillograms of the voltage measured on a carrier supply bay between the points *a* and *b* in fig. 7.

I) The ideal case, obtained when there is a filter with very small resonance resistance R for all the harmonics contained in the impulse. (Recorded by replacing the whole series of filters by a resistance.)

II) The actual case in normal operation. The slight fluctuations in the time between the impulses indicate that for the higher harmonics the added extra filters and the correction network represent approximately, but not exactly, the desired low resistance R.

III) The carrier filter for the fourth harmonic has been removed. The remaining system of filters now has a much higher resistance for that harmonic, so that a considerable voltage with that frequency occurs between the points *a* and *b*. (This phenomenon can be used as a simple method of localizing any defect in one of the filters.)

IV) The carrier filters for all the even harmonics have been removed. Analogous to the case under (*III*) all the corresponding frequencies now occur in the voltage between *a* and *b*. Together these form another periodic impulse, but with a fundamental frequency twice as high as the one supplied to the output pentode. Thus a frequency doubling, as it were, has been realized by the remaining set of filters, when we take off the voltage between *a* and *b*.

is small enough so that practically all the current of a given frequency flows through the circuit tuned to that frequency, and if that current can be disregarded in all the other filters together, the system behaves as a resistance R for each of the harmonics and therefore for the whole impulse voltage. Thus we have a current in impulse form flowing through the connecting wires to the filters, *i.e.* in the time between two impulses the source need not supply any current, and it makes no difference whether or not the source is switched off from the filters in the time between two impulses.

The practical importance of this consideration is easy to understand: the desired action of the filters is ensured only when there is a filter present in the series for each of the harmonics occurring in the impulse voltage. Only then do all the currents of different frequencies flowing in the filters add together in such a way that in the interval between two impulses the total current is zero. If a number of filters are missing the other filters deliver less power of their own frequency. Thus even if we wish to use only a limited number of harmonics of the periodic impulse for carriers, we must have a filter for each of the remaining harmonics. In practice when all the harmonics up to a certain order are used, it is sufficient to take only a few more circuits than those necessary for the carrier frequencies, because, as appears from the Fourier series given above, the higher harmonics become rapidly weaker and thus have less effect; all the circuits for the very highest harmonics together can be finally replaced by a simple correction network.

The above can be demonstrated by means of a simple experiment. In a carrier supply bay of a 17-channel system with carriers from 4 to 68 kc/sec all the filters above 56 kc/sec are taken out of connection. The power of the carrier at the output of the 56 kc/sec filter thereby decreases no less than 50 percent.

It is interesting to note that the whole combination of filters behaves as a resonance circuit for an impulse voltage, in the same way as a single L-C-circuit behaves with respect to a sinusoidal voltage. In particular, when the attenuation constants (R/L) of all the filters are made equal, after the removal of the external impulse voltage the combination can be seen to oscillate with an impulse voltage, which decreases according to a power of e. The output voltage of the set of filters can also be used instead of the master oscillator to synchronize the impulse generator. The frequency of the latter is kept constant by the tuning of the filters in the same way as the sinusoidal vibration of an ordinary back-coupled oscillator is kept constant by an L-C-circuit in the anode circuit. Finally a "frequency multiplication" can even be realized, by removing all the filters of even order. A new set of filters is thus obtained, which forms as it were an "impulse resonance circuit" for double the fundamental

frequency of the impulse applied to it, and one then indeed obtains over the filters an impulse voltage with this double frequency. This surprising fact is illustrated by the last of the oscillograms in *fig. 8*. These oscillograms are explained in the text below the figure.

In order to obtain the greatest possible power out of the output pentode it is necessary here to make the resistance R of the filters equal to the internal resistance of the valve during the impulse. Actually, however, part of this optimum adaptation is sacrificed and R is made somewhat larger. The resultant decrease in power is scarcely worth mentioning, and the advantage is obtained that during the impulse the voltage drop in the valve, which is naturally small, since in a pentode a large anode current can already flow with a very small anode voltage, becomes still smaller, for example only 10 V, while the rest of the total feeding voltage, for instance 190 V, acts upon R. In this way an impulse height is obtained and with it an amplitude of the carriers which is affected practically only by fluctuations in the feeding voltage (mains voltage) and hardly at all by the properties of the valve or the magnitude of the control grid voltage. Thus the above mentioned requirement of constancy of the carrier amplitudes within 10 percent can easily be satisfied.

We give in *fig. 9* a photograph of the master-oscillator which determines the fundamental frequency of the whole series of carriers, and the corresponding reserve-oscillator, both mounted in the carrier supply bay of finished installation for carrier telephony. Some details are explained in the text below the figures.

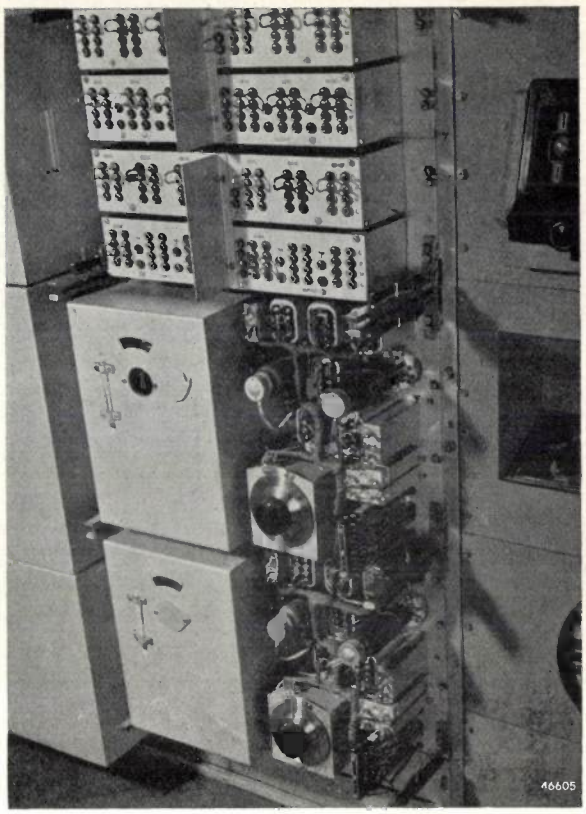


Fig. 9. The master oscillator with the (identical) reserve oscillator. In each oscillator in addition to the oscillator valve itself may be seen a second amplifier valve which serves as coupling valve, *i.e.* it prevents the oscillator valve being affected by the apparatus following it. The oscillation circuit is placed in a thermostat under the cover to the left. The temperature of this thermostat can be set by opening the window in the cover (in the upper oscillator it is open) and it can be read off on the thermometer beside it. With the rotating knob at the bottom of each oscillator the frequency can be very precisely adjusted (entire scale $\approx$ 1 c/sec).

MEASUREMENT OF THE ADHESIVE FORCE OF LACQUERS

by P. KOOLE.

620.179.4:620.198:667.7

The adhesive force of a lacquer is defined as the force necessary to scrape off a 1 cm wide strip of lacquer from an under-layer by means of a chisel. A description is given of a simple apparatus for measuring this force.

A layer of lacquer applied to a metal surface either to protect it from atmospheric influences or for esthetic reasons must, in order to fulfil its function, also have sufficient mechanical strength. This means that the lacquer must possess sufficient elasticity and hardness so that, when struck, it will not crumble or be scratched and that it must adhere firmly to the underlayer so as not to flake off or be easily abraded. The mechanical properties first mentioned, elasticity and hardness, can be measured fairly simply by the usual methods, but when judging the quality of a layer of a lacquer it is of no less importance to know the other property mentioned, the adhesive force of the lacquer, and there is as yet no generally accepted method of measuring this property. To begin with, there is no general agreement about what should be regarded as a measure of the adhesive power. One would be inclined to define the adhesive power as the force acting perpendicular to the surface of the lacquer which is necessary to remove 1 cm^2 of the lacquer film from its under-layer. In practice, however, this definition has no value, since an appropriate method of measuring such a force involves very great difficulties and, moreover, a lacquer film is in practice never subjected to such a treatment. The adhesive power can more appropriately be measured by the force which has to be exerted on a knife or chisel to scrape off, for example, a strip of lacquer 1 cm wide from its underlayer. It is of course true that in this force, besides the adhesive force proper determined by the nature of the lacquer and of the under-layer, the strength properties of the lacquer itself are discounted, and it depends therefore partially on the thickness of the layer of lacquer. Nevertheless, this force lends itself well to measurements on a large scale and corresponds somewhat to the forces occurring in practice which may cause mechanical injury to the lacquer film.

The simplest method of measurement based upon this definition consists in drawing a plate with a 1 cm wide lacquer film along a guide under a chisel slightly more than 1 cm wide and measuring the force which has to be exerted on the plate (see *fig. 1*). According to the sharpness of the chisel, a certain pressure on it is necessary in order to prevent it from sliding over the lacquer film. Due to this pressure, not only does a frictional force occur between the chisel and the under-layer — this could be determined, together with the frictional force between the plate and the guide, by repeating the experiment without the lacquer film — but at the same time the chisel has a tendency to penetrate into the base material, no matter how small the angle of inclination (α in fig. 1) is chosen.

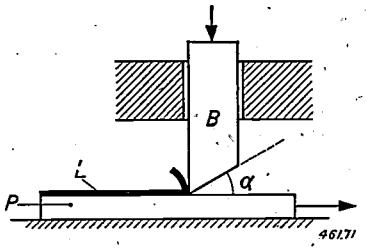


Fig. 1. The adhesive force of the lacquer film on the plate P can be measured by drawing the plate along a guide under the chisel B in such a way that the lacquer film is just scraped off the metal.

In order to avoid this, the chisel must be fixed rigidly and before each test the cutting edge must be so adjusted that it just shaves over the surface of the metal. The surface of the metal, however, must be perfectly smooth and flat, as otherwise the chisel will cut into the under-layer at some points and at others leave a thin layer of lacquer behind. It is therefore obvious that one cannot use any arbitrary lacquered plate for this test, but rather the lacquer film should be applied to a smooth under-layer specially prepared for the purpose. The requirements with regard to flatness of the under-layer prove to be so severe that in the set-up according to fig. 1 it was not even possible to grind the plate used as under-layer to a sufficiently accurate degree of flatness. We therefore decided to measure the adhesive force of the lacquer film on a cylindrical surface, which it is possible to grind without any great difficulty to a degree of accuracy within 1 micron, which proved to be sufficient for our purpose.

With a cylindrical surface as under-layer it is, moreover, possible to apply a very simple method of operation. The relative motion between lacquer film and chisel can then be brought about by fixing the chisel to a lever rotating about the same axis as that of the cylinder with the lacquer film. In this way we arrived at the arrangement shown in *fig. 2*. The cylindrical surface is formed by the rim of a disc 1 cm thick, upon which the lacquer is applied with the aid of a spray stencil, the sides and a narrow strip of the cylindrical surface itself

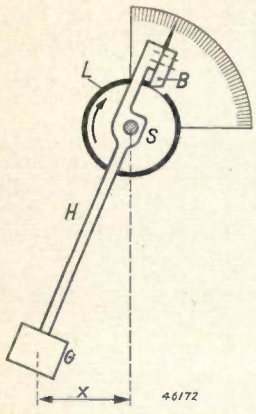


Fig. 2. Arrangement for the measurement of the adhesive force. S steel disc, accurately ground to circular shape, L lacquer film, H lever carrying the hard steel chisel B, G counter-weight.

being left bare. This bare strip serves as starting plane for the chisel, which is so adjusted that it just scrapes over the metal. The disc is now rotated slowly by means of a motor (at about 1 r. p.m.) in a direction opposite to the cutting direction of the chisel. The lacquer film tends to carry the chisel along with it, but to do so it must raise a counter-weight on the chisel lever. The largest couple that the lacquer film can exert on the lever is the product of the adhesive force and the radius of the disc. The couple of the counter-weight ($C \cdot x$) increases with increasing movement of the lever and at a given point the adhesive force is overcome and the lacquer is scraped off the disc. At this angle, which can be read off on a graduated arc, the lever remains in equilibrium during the rotation of the disc. The graduated arc can be calibrated directly in units of adhesive force (kg/cm).

Since the chisel also experiences a frictional resistance along the metal, and since a frictional force also occurs at the fulcrum of the lever, the experiment has to be repeated without the lacquer film on the disc. The difference between these two measured forces gives the adhesive force.

Fig. 3 is a photograph of an apparatus which has been in use for some time and has proved very satisfactory. With the regulatory resistance, which may be seen below, the number of revolutions of the motor can be adjusted to a suitable value. It may also be seen in the photograph that three different counter-weights can be used with the apparatus. This makes it possible to cover a wide range of measurements, since the adhesive forces of different kinds of lacquer vary considerably.

As has already been noted, the adhesive force according to our definition depends also on the thickness of the lacquer film. In order to obtain comparable values, therefore, it is necessary always to measure with the same thickness of lacquer, for instance 0.05 mm. The inevitable slight variations in thickness of the layer obtained by spraying on the lacquer — even for an experienced worker it will not be possible to keep to the thickness mentioned to within less than 0.01 mm — can in this set-up be eliminated very simply by applying a somewhat thicker layer and then, after it has dried, turning it to exactly the required thickness by means of the chisel. The necessary adjustment of the chisel is obtained by placing a metal strip of the required thickness under the chisel at the starting place.

The true adhesive force between lacquer and under-layer will not depend upon the thickness of the lacquer, but it will undoubtedly depend upon the nature of the surface of the under-layer. In our measurements, as has been seen, a smoothly ground under-layer must always be used, while in practice the surface to be covered will always be more or less rough.

It will sometimes even be roughened purposely, because the adhesion of the lacquer is then better. It is not *a priori* certain whether two kinds of lacquer, which adhere equally well to the smooth surface, will also adhere equally well to the rough surface. If, however, the adhesive force of a series of different lacquers is determined in the manner described, an order of evaluation is obtained which agrees perfectly with practical experience. Thus even although the relation between the adhesive force measured with the apparatus and the adhesion to be expected in practice is not entirely certain the values measured are quite useful as comparative figures for the quality of the lacquers to be used.

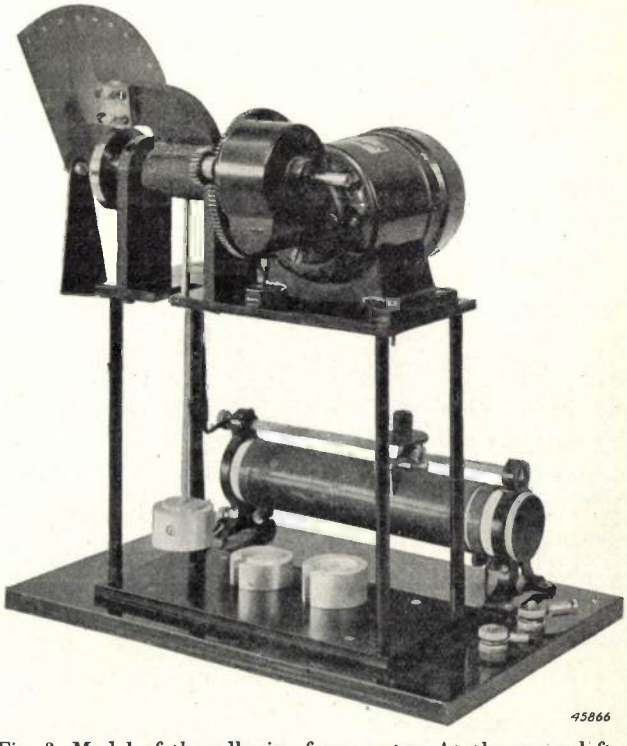


Fig. 3. Model of the adhesive-force meter. At the upper left may be seen the steel disc from which the lacquer film is scraped off, and the scale on which the necessary force is read off. At the upper right are the motor and the tooth-wheeled transmission for driving the disc. At the bottom a counter-weight may be seen suspended on the lever, and two counter-weights for different measuring ranges.

FLAT CAVITIES AS ELECTRICAL RESONATORS

by C. G. A. von LINDERN† and G. de VRIES.

538.565:538.566.5

After some introductory remarks about the characteristic vibrations which may occur in Lecher systems which are short-circuited at one end, this article deals with conical flat cavity resonators short-circuited around their outer edge and whose behaviour, as regards the rotation-symmetrical vibrations, corresponds entirely to that of the Lecher systems first mentioned. Subsequently the rotation-symmetrical characteristic vibrations of flat cavity resonators of more general forms are discussed and the variation of current and voltage with the radius is drawn for a flat cavity resonator. Resonance resistance and quality factor are calculated. Further it is indicated how the quality factor and the resonance resistance can be improved by making the cavity resonators thicker than those for which the theory given here applies unconditionally. In conclusion some examples are given of flat cavity resonators and their employment in high-frequency technique, for example for frequency stabilization and as output and input electrodes for short-wave transmitting valves.

In the excitation of high-frequency oscillations in radio transmitters, especially for very short waves, use is made of specially formed resonators; because at very high frequencies the ordinary oscillation circuits consisting of concentrated capacities and self-inductions possess too low a resonance resistance and too small a quality factor [1]). In addition to the so-called resonance cavities, Lecher systems consisting of two parallel conductors of uniform diameter as well as empty spaces with metal walls, which we call cavity resonators, are often used in high-frequency technology. In a Lecher system of a given length l stationary waves can be generated having a wave length of $4l$, provided the Lecher system is excited at one end by an electromagnetic oscillation, while the other end is short-circuited [2]). We shall now examine the free oscillations which may occur in cavity resonators of a general form, but in this article shall confine ourselves to flat cavity resonators, which have one dimension much smaller than the other two. When such flat cavity resonators are constructed in the form of solids of revolution, electromagnetic vibrations may occur in them which do not depend upon the angle of revolution, but only upon the radius. Such rotation-symmetrical oscillations thus actually depend upon only one coordinate. They can therefore still be treated in practically the same way as the oscillations of a Lecher system, which are of course also considered as depending exclusively on one coordinate, namely the length, since the transverse dimensions may be neglected compared with the length.

[1]) See for example our article "Resonance circuits for very high frequencies", Philips techn. Rev. 6, 217, 1941, in which, *inter alia*, different equivalent definitions of the quality factor are explained in more detail.

[2]) The behaviour of a Lecher system with respect to travelling and stationary waves was discussed extensively by us in Philips techn. Rev. 6, 241, 1941.

We shall first discuss the characteristic vibrations which may occur in a homogeneous Lecher system short-circuited at one end. Then we shall extend our discussion to include flat cavity resonators with double conic cross-section, after which we shall have something to say about the characteristic vibrations of flat cavity resonators of more general forms. After that we shall examine the influence of the resistance, giving values for resonance resistance and quality factor. These theoretical considerations will then be concluded with a discussion of the possibility of building up cavity resonators not having one dimension much smaller than the others, by piling up flat cavity resonators one upon another. Finally some examples are given of the practical use made of the cavity resonators here discussed in high-frequency technology.

Characteristic vibrations of Lecher systems

Let us consider a section of a Lecher system consisting of two parallel conductors of uniform cross-section and a given length l, which are connected at one end. We shall ignore the energy losses for the present, while for the capacity and self-induction per unit of length we introduce the symbols C^I and L^I. For the current i and the voltage V as functions of the time t and the coordinate of length x, the following equations hold:

$$\frac{\partial i}{\partial x} = -C^I \frac{\partial V}{\partial t} \text{ and } \frac{\partial V}{\partial x} = -L^I \frac{\partial i}{\partial t} \quad . \quad . \quad . \quad (1)$$

By differentiating these equations with respect to x and t four equations are obtained:

$$\left.\begin{aligned} \frac{\partial^2 i}{\partial x^2} &= -C^I \frac{\partial^2 V}{\partial x \partial t}; \quad \frac{\partial^2 i}{\partial x \partial t} = -C^I \frac{\partial^2 V}{\partial t^2}; \\ \frac{\partial^2 V}{\partial x^2} &= -L^I \frac{\partial^2 i}{\partial x \partial t}; \quad \frac{\partial^2 V}{\partial x \partial t} = -L^I \frac{\partial^2 i}{\partial t^2}, \end{aligned}\right\} \quad (2)$$

from which by combination two corresponding differential equations of the second order (the so-called vibration equations) for i and V can be derived:

$$\frac{\partial^2 i}{\partial x^2} = + L^I C^I \frac{\partial^2 i}{\partial t^2} \text{ and } \frac{\partial^2 V}{\partial x^2} = + L^I C^I \frac{\partial^2 V}{\partial t^2} . \quad (3)$$

If now for the sake of simplicity we assign the coordinate $x = 0$ to the short-circuited end of the Lecher system (*fig. 1*) and to the open end $x = l$,

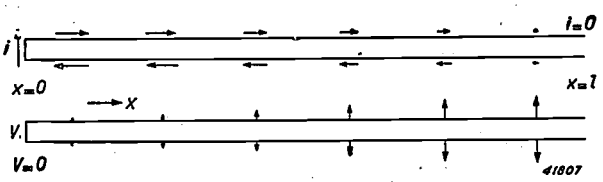


Fig. 1. Behaviour of current i and voltage V along a Lecher system short-circuited at $x = 0$ and open at $x = l$, while it vibrates at the fundamental frequency.

we require only those solutions of the vibration equations (3) for which the voltage V is zero for $x = 0$. For the characteristic vibrations of the short-circuited section of Lecher system there is the additional condition that at the open end $x = l$ the current i must be zero. The simplest stationary wave, which satisfies the vibration equations (3) and also these two conditions, is, except for a constant factor, the following [3]):

$$\left.\begin{aligned} i &= \sin \frac{\pi t}{2\, l \sqrt{L^I C^I}} \cos \frac{\pi x}{2l}, \\ V &= - \sqrt{\frac{L^I}{C^I}} \cos \frac{\pi t}{2\, l \sqrt{L^I C^I}} \sin \frac{\pi x}{2l}. \end{aligned}\right\} \quad (4)$$

The variation of current and voltage along the Lecher system according to (4) is represented in fig. 1.

The coefficient of the time t in the expression (4) is the angular frequency $\omega = 2\pi\nu$. For the frequency ν_1 of the simplest characteristic vibration the following thus holds:

$$\nu_1 = \frac{1}{4\, l \sqrt{L^I C^I}}, \quad (5)$$

In the same way it may be seen from the expressions (4) that the length λ_1 of the largest stationary wave along the Lecher system short-circuited at one end is four times as great as the length l. The dependence of x on current and voltage can be represented by a cos or sine of $2\pi x/\lambda_1$, so that from (4) it does indeed follow that:

$$\lambda_1 = 4\, l \quad (6)$$

If (5) and (6) are multiplied by each other the well-known result follows that the product of the characteristic frequency ν_1 and the corresponding wave length λ_1 is equal to $1/\sqrt{L^I C^I}$, and this is the velocity of propagation v of the travelling waves which may occur along the Lecher system:

$$\nu_1 \lambda_1 = v = 1/\sqrt{L^I C^I} = \frac{3 \cdot 10^{10}}{\sqrt{\varepsilon \mu}} \text{ cm/sec.} \quad (7)$$

The expressions represented by (4) for current and voltage of the short-circuited Lecher system represent the so-called fundamental vibration of the latter. In addition to this, overtones also occur as characteristic vibrations of the same Lecher system (*fig. 2*), which also satisfy the vibration equations (3) and the conditions for voltage and current mentioned:

$$\left.\begin{aligned} i &= \sin \frac{(2k+1)\pi t}{2\, l \sqrt{L^I C^I}} \cos \frac{(2k+1)\pi x}{2l}, \\ V &= - \sqrt{\frac{L^I}{C^I}} \cos \frac{(2k+1)\pi t}{2\, l \sqrt{L^I C^I}} \sin \frac{(2k+1)\pi x}{2l}, \end{aligned}\right\} \quad (8)$$

where k may assume the values of the whole numbers. The characteristic frequencies ν_{2k+1} of these overtones lie harmonically and are the odd multiples of the fundamental frequency ν_1, namely:

$$\nu_{2k+1} = \frac{2k+1}{4\, l \sqrt{L^I C^I}} = (2k+1)\, \nu_1 \quad (9)$$

The corresponding wave lengths are:

$$\lambda_{2k+1} = \frac{4l}{2k+1} = \frac{\lambda_1}{2k+1}, \quad (10)$$

so that the length l of the Lecher system is equal to an odd number of quarter wave lengths. In the case of homogeneous Lecher systems we are therefore only concerned with the familiar harmonic overtones which also occur with a vibrating string. This, however, is no longer the

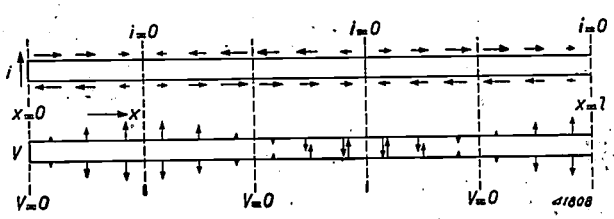


Fig. 2. Behaviour of current i and voltage V for the third characteristic vibration λ of a Lecher system short-circuited at one end. ($k = 2$ in equations (8), (9) and (10)), in the case of fundamental vibration.

[3]) The magnitude and sign of the coefficient in the expression for the voltage V is of course so chosen that the expressions (4) also satisfy the differential equations (1) of the first order for current and voltage from which we started in this discussion. $\sqrt{L^I/C^I}$ is often called the "wave resistance" (*cf.* Philips techn. Rev. **6**, 241, 1941).

case when the capacity and the self-induction depend upon the position, *i.e.* when the Lecher system is not homogeneous.

Then the position of the overtones is anharmonic, as we shall see later in this article.

Conical flat cavity resonators

If we were to imagine the homogeneous Lecher system short-circuited at one end, which we have considered until now, to be rotated so as to form a flat box, with or without a hole in it (*fig. 3*), we

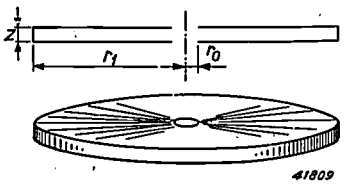


Fig. 3. Flat cavity resonator with thickness z and radius r_1 of the short-circuited outer edge. The radius of the hole is r_0. The flat cavety resonator originates by the revolution of a homogenious Lecher System, short-circuited at one end.

might expect that the rotatory-symmetrical characteristic vibrations of such a flat cavity resonator would exhibit a far-reaching similarity with those of a Lecher system short-circuited at one end. However, this does not give us the two-dimensional cavity resonator with the simplest characteristic vibrations, since in the vibration equations for current and voltage as functions of the radius r, the capacity and self-induction per ring of 1 cm width play a part in this rotatory-symmetrical case and these quantities, which we shall again call C^I and L^I, are not constant as in the homogeneous Lecher system, but depend upon the radius r of the ring being considered. It will be obvious that C^I is directly proportional to r, while L^I is inversely proportional to it; the expressions are as follows:

$$C^I = \frac{r}{2z} \text{ and } L^I = \frac{2z}{c^2 r}, \quad \ldots \quad (11)$$

where z represents the thickness of the flat cavity resonator, the dielectric is a vacuum and c is the velocity of light.

A case, which is indeed mathematically entirely analogous to the homogeneous Lecher system, is a conical flat cavity resonator, for which the distance z between the two conductors is proportional to the radius r of the ring being considered (*fig. 4*). According to the ratios (11) C^I and L^I then do indeed again become independent of r. For the conical flat cavity resonator, therefore, we obtain exactly the same harmonic characteristic vibrations as for the homogeneous Lecher system, only in this case we are not dealing with a coordinate x which passes from $x = 0$ at the short-circuited

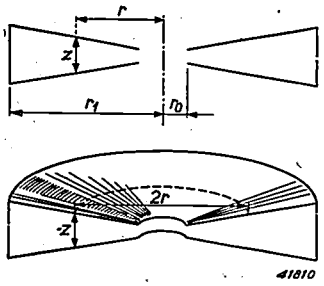


Fig. 4. Conical flat cavity resonator with thickness z at radius r. Outer radius r_1, inner radius r_0.

end to $x = l$ at the open end, but with a coordinate r which passes from $r = r_0$ at the open inner edge to $r = r_1$ at the short-circuited outer edge. Because of this the solutions of the vibration equations become somewhat more complicated than (4) and (8), but one is still dealing with an odd number of quarter sine waves of current and voltage (*fig. 5*), which now lie along the radius of the conical flat cavity resonator instead of along the length of the homogeneous Lecher system, so that in general the following is valid:

$$r_1 - r_0 = \frac{2k+1}{4} \lambda_{2k+1}, \quad \ldots \quad (12)$$

which is quite analogous to relation (10) for the homogeneous Lecher system.

Characteristic vibrations of other flat cavity resonators

Flat cavity resonators whose thickness z depends in any arbitrary manner on r do not behave, as far as the electromagnetic characteristic vibrations are concerned, entirely like a Lecher system of uniform cross-section, as is the case for the conical flat cavity resonator just discussed.

In order to describe the axially symmetrical

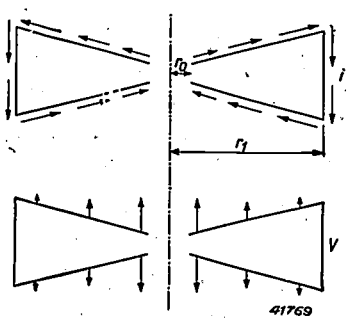


Fig. 5. Variation of current i and voltage V with the radius for a conical flat cavity resonator i: r_0 inner radius, r_1 outer radius.

characteristic vibrations which may occur in a flat cavity resonator with a given cross-section, we refer to the treatment of the characteristic vibrations of a so-called non-homogeneous Lecher system [4]), for which the cross-section and/or the distance between the two conductors depends upon the position (*fig. 6*). With Heaviside we shall now assume

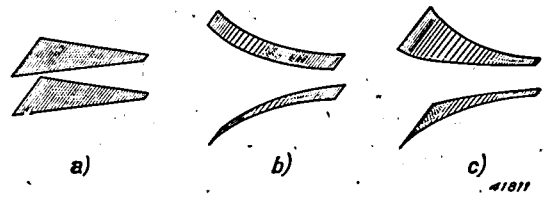


Fig. 6. Different forms of generalized Lecher systems consisting of:
a) two parallel strips of varying width,
b) two strips of uniform width at varying distance apart and
c) two strips which not only become wider but also diverge.

that the quantities C^I and L^I vary only slowly with the coordinate, which for the sake of simplicity we shall immediately call r, in order to obtain equations, which are also valid for the flat cavity resonators. We then obtain the following for the variations of current i and voltage V:

$$\frac{\partial i}{\partial r} = -C^I \frac{\partial V}{\partial t} \text{ and } \frac{\partial V}{\partial r} = -L^I \frac{\partial i}{\partial t} \quad . \quad . \quad (13)$$

These equations (13) are now again differentiated with respect to r and t:

$$\frac{\partial^2 i}{\partial r^2} = -\frac{\partial C^I}{\partial r}\frac{\partial V}{\partial t} - C^I \frac{\partial^2 V}{\partial r \partial t} \text{ and } \frac{\partial^2 i}{\partial r \partial t} = -C^I \frac{\partial^2 V}{\partial t^2};$$

$$\frac{\partial^2 V}{\partial r^2} = -\frac{\partial L^I}{\partial r}\frac{\partial i}{\partial t} - L^I \frac{\partial^2 i}{\partial r \partial t} \text{ and } \frac{\partial^2 V}{\partial r \partial t} = -L^I \frac{\partial^2 i}{\partial t^2}.$$

By combination of these equations, in which (13) has to be substituted for the first derivatives, about the same differential equations of the second order can again be found as vibration equations for i and V:

$$\left.\begin{aligned} \frac{\partial^2 i}{\partial r^2} - \frac{1}{C^I}\frac{\partial C^I}{\partial r}\frac{\partial i}{\partial r} - L^I C^I \frac{\partial^2 i}{\partial t^2} &= 0 \\ \frac{\partial^2 V}{\partial r^2} - \frac{1}{L^I}\frac{\partial L^I}{\partial r}\frac{\partial V}{\partial r} - L^I C^I \frac{\partial^2 V}{\partial t^2} &= 0 \end{aligned}\right\} \quad . \quad (14)$$

It may be seen that compared with the vibration equations (3) there is a middle term added due to the slow change of C^I and L^I with the coordinate r.

In order to obtain closed solutions of (14) we shall now assume as a special case that the thickness z of our flat cavity resonator is proportional to an arbitrary power n of the radius r (*fig. 7*):

$$z = h\left(\frac{r}{r_1}\right)^n \quad . \quad . \quad . \quad . \quad . \quad . \quad (15)$$

According to (11) capacity and self-induction per ring 1 cm wide in a vacuum are:

$$C^I = \frac{r}{2h}\left(\frac{r_1}{r}\right)^n \text{ and } L^I = \frac{2h}{c^2 r}\left(\frac{r}{r_1}\right)^n \quad . \quad (16)$$

thus, respectively, inversely and directly proportional to r^{n-1}. For this special case therefore the vibration equations (14) for current and voltage become:

$$\left.\begin{aligned} \frac{\partial^2 i}{\partial r^2} + \frac{n-1}{r}\frac{\partial i}{\partial r} - \frac{1}{c^2}\frac{\partial^2 i}{\partial t^2} &= 0, \\ \frac{\partial^2 V}{\partial r^2} + \frac{1-n}{r}\frac{\partial V}{\partial r} - \frac{1}{c^2}\frac{\partial^2 V}{\partial t^2} &= 0. \end{aligned}\right\} \quad . \quad . \quad (17)$$

Just as for the simple vibration equations (3), we now look for solutions of (17) which are harmonic in the time t. For current i and voltage V we again take as variation with time sin and cos $2\pi ct/\lambda$, respectively. For the amplitudes of current (i_m) and voltage (V_m), which depend only upon the coordinate r, when we introduce $x = 2\pi r/\lambda$ as new independent variable we obtain the following equations:

$$\left.\begin{aligned} \frac{\partial^2 i_m}{\partial x^2} + \frac{n-1}{x}\frac{\partial i_m}{\partial x} + i_m &= 0, \\ \frac{\partial^2 V_m}{\partial x^2} + \frac{1-n}{x}\frac{\partial V_m}{\partial x} + V_m &= 0. \end{aligned}\right\} \quad . \quad . \quad (18)$$

For the case where $n = 1$ the middle term becomes equal to zero, so that the vibration equation (3) is again obtained for current and voltage. For $n = 1$ we have in fact the simple case, already discussed, of the conical flat cavity resonator with i and V varying harmonically in t and r.

Just as it is well known that the solution of (18) without the middle term are sines and cosines of

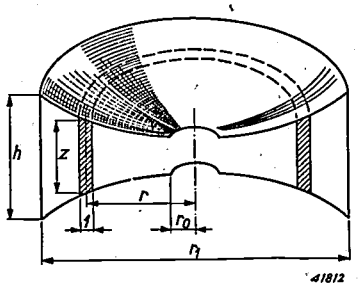


Fig. 7. Flat cavity resonator with thickness h at the outer edge, outer radius r_1 and inner radius r_0.

[4]) For a recent review of the theory of non-homogeneous Lecher systems see K. W. Wagner, Die Theorie ungleichförmiger Leitungen, Arch. Elektrotechn. **36**, 69, 1942.

x, it is also known that Bessel functions of the first and second sorts (J_p and N_p, where p denotes the "order" of the Bessel function) multiplied by powers of the coordinate x are solutions of (18). It is clear that not all solutions of (18) also satisfy the differential equations (13) with which we began. Those, for which this is indeed the case, are given by the following expressions:

$$\left.\begin{aligned} i_m &= x^{1-n/2}\,[A\,J_{n/2-1}(x) + B\,N_{n/2-1}(x)],\\ V_m &= -\sqrt{-1}\,\zeta\,x^{1-n/2}\,[AJ_{n/2}(x) + BN_{n/2}(x)]; \end{aligned}\right\} \quad (19)$$

A und B are arbitrary constants here and ζ is the wave resistance (*cf.* footnote [4]):

$$\zeta = \sqrt{\frac{L'}{C'}} = \frac{60\,h}{r_1^{\,n}}\,r^{n-1}\ \text{ohm}.$$

We must now choose A and B in (19) such that their expressions satisfy the boundary conditions.

In the first place the voltage must be zero at the short-circuited outer edge of the flat cavity resonator, *i.e.* $V_m = 0$ for $r = r_1$, or $x = 2\pi\,r_1/\lambda = b$. This boundary condition thus gives:

$$AJ_{n/2}(b) + BN_{n/2}(b) = 0 \quad . \quad . \quad . \quad (20)$$

By filling in the value of B from (20) in (19) we obtain:

$$i_m = Ax^{1-n/2}\,[N_{n/2}(b)\,J_{n/2-1}(x) - J_{n/2}(b)\,N_{n/2-1}(x)], \quad . \quad . \quad (21a)$$

$$V_m = -\sqrt{-1}\,A\,\zeta\,x^{1-n/2}\,[N_{n/2}(b)\,J_{n/2}(x) - J_{n/2}(b)\,N_{n/2}(x)] \quad . \quad . \quad (21b)$$

In the second place in the case of a flat cavity resonator with a hole in it (*i.e.* $r_0 > 0$) the current must be zero at the inner edge. This gives us the relation

$$N_{n/2}(b)\,J_{n/2-1}(a) - J_{n/2}(b)\,N_{n/2-1}(a) = 0, \quad (22)$$

where $a = 2\pi r_0/\lambda$. In the case of a flat cavity resonator without a hole (*i.e.* $r_0 = 0$), the current in the centre ($i = 0$) for $n > 0$ need not be zero because the top and bottom planes do not then touch each other. It certainly cannot, however, be infinite at that point, as would follow for $x = 0$ and $n \geqq 2$ from expression (21a).

It is found by closer inspection of this expression that the current remains finite for $x = 0$ and $n \geqq 2$ provided

$$J_{n/2}(b) = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (23)$$

For $n < 2$, except for $n = 1, -1, -3, \ldots$, the same condition (23) results in the fact that the current (21a) at the centre not only remains finite but is moreover equal to zero. In order to attain this for $n = 1, -1, -3, \ldots$, instead of (23), the condition must be fulfilled that

$$N_{n/2}(b) = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (24)$$

Now since $a = 2\pi r_0 \nu/c$ and $b = 2\pi r_1 \nu/c$ and r_0, r_1 and n are fixed for a given flat cavity resonator, equation (22) (resonator with hole) or (23) or (24) (without hole) are only satisfied for certain values of the frequency ν. These special values of ν, which are thus the roots of equations (22) (23) or (24), respectively, represent the series of characteristic frequencies of the flat cavity resonator in question. The smallest positive root represents the fundamental tone.

It remains to be mentioned, that the values of the characteristic frequencies calculated from (22) pass over continuously into those calculated from (23) or (24) when a in (22), *i.e.* the radius of the hole, is allowed to approach zero.

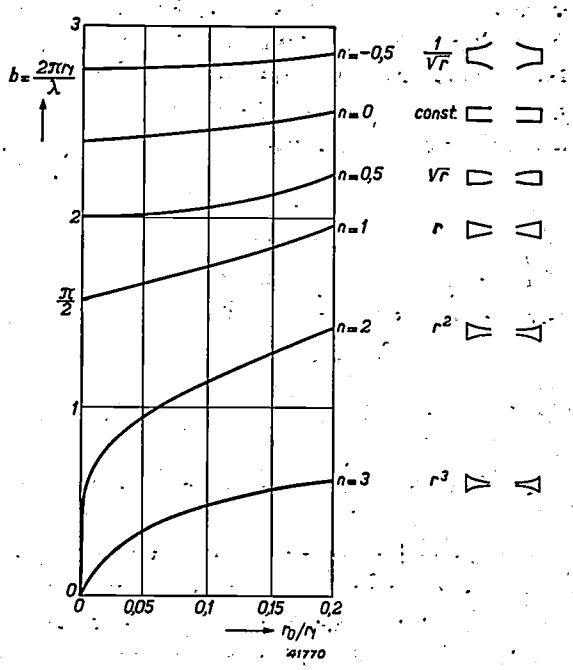


Fig. 8. The quantity $b = 2\pi r_1/\lambda$, which is proportional to the characteristic frequency, as a function of the ratio r_0/r_1 of the radii with different values of n for flat cavity resonators. For $n = 1$ and $r_0 = 0$ the quantity b becomes $\pi/2$, so that the wave length λ is then equal to four times the outer radius r_1.

In *fig. 8* it is now shown how at different values of n according to (22) the value of $2\pi r_1/\lambda$ for the fundamental tone varies with the quotient of inner and outer radii: $r_0/r_1 = a/b$. In the limiting case of a conical flat cavity resonator the wave length λ_1 of the fundamental tone is four times the difference between outer radius r_1 and inner radius r_0. It is evident from fig. 8 that this difference $r_1 - r_0$ is, however, in general by no means equal to a

quarter wave length. For instance in the case of a flat cavity resonator (fig. 3) with no hole ($n = 0$ and $r_0 = 0$), for which at the fundamental tone the following holds: $b = 2\pi r_1/\lambda_1 = 2.405$, the wave length becomes

$$\lambda = \frac{2\pi r_1}{2.405} = 2.61\, r_1.$$

If we now wish to know the wave lengths of several overtones for such a flat cavity resonator we can easily find them from the higher roots of equation (23) where n must be set equal to zero, so that we obtain:

$$J_0\,(2\pi r_1/\lambda) = 0. \quad . \quad . \quad . \quad . \quad . \quad . \quad (23')$$

The first three zero points of (J_0) lie respectively at $2\pi r_1/\lambda = 2.405$; 5.520; 8.653 (*cf. fig. 9*), from which then follow for the three longest wave lengths at which a flat cavity resonator with no hole can execute characteristic vibrations:

$$\lambda_1 = 2.61\, r_1;\ \lambda_2 = 1.14\, r_1;\ \lambda_3 = 0.73\, r_1\,. \quad (25)$$

The quotient of λ_2 and λ_3 is 1.56, which is not much smaller than $5/3 = 1.67$, which would be the ratio with a harmonic position of the characteristic frequencies. The frequencies of the higher overtones for the flat cavity resonator with no hole are indeed found to be progressively more nearly in the ratios of the successive odd numbers. Thus for example $\lambda_3/\lambda_4 = 1.36$ and $\lambda_4/\lambda_5 = 1.27$ while $7/5 = 1.40$ and $9/7 = 1.29$. It is only the ratio of the fundamental tone to the overtones which is far from harmonic, since $\lambda_1/\lambda_2 = 2.29$, which differs very much from 3!

In conclusion, for the simple case of the flat cavity resonator with no hole ($n = 0$; $r_1 = 0$) we shall examine the way in which the current and voltage are distributed along the radius on the basis of equations (21) and (23'). When we make use of the relation $J_{-1}(x) = -J_1(x)$ and omit the constant factors: we obtain the following:

$$\left.\begin{aligned} i_m &= x\, J_1\,(x)\,, \\ V_m &= J_0\,(x)\,. \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (26)$$

The variation of current and voltage with the radius according to (26) is drawn in fig. 9. In the variation of the voltage we find the zero points of J_0 already mentioned in the derivation of the characteristic frequencies [5]), which are at the same time the points of the maxima of the current curve xJ_1.

The formulae given here lose their physical significance as soon as the absolute value of n becomes too large. The quantities C^I and L^I can then no longer be considered as varying slowly with r (*cf.* (16)), so that the assumption for which the initial equations (13) were derived is no longer satisfied.

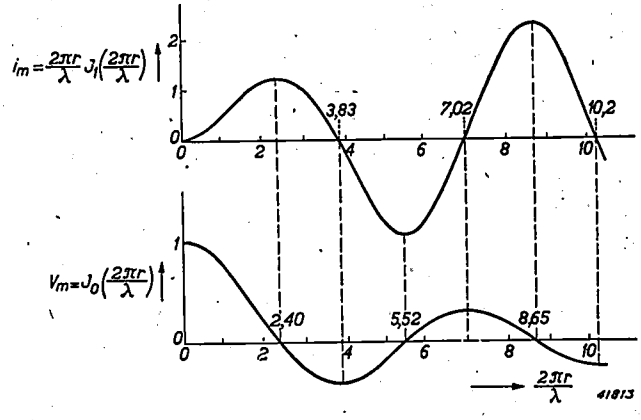


Fig. 9. Variation of the current amplitude i_m and the voltage amplitude V_m with the radius r of a flat cavity resonator. The behaviour from $x = 0$ to 2.405 gives current and voltage of the fundamental tone, from $x = 0$ to 5.520 the same for the second characteristic vibration, from $x = 0$ to 8.653 for the third characteristic vibration, and so on. In the same way at the zero points of the current curve xJ_1 lie the maxima for the voltage J_0.

Resonance resistance and quality factor

In the foregoing considerations we have assumed that we were concerned with flat cavity resonators having no ohmic resistance. Upon resonance no current then flows at the open inner edge, independent of the voltage acting on it, so that we would be dealing with an infinite resonance resistance. Now, however, we wish to calculate the resonance resistance and the quality factor which can be obtained in practice with flat cavity resonators, in which case, therefore, the energy losses should actually be taken into account.

The quality factor Q of a resonance circuit, except for a factor π, is the reciprocal of the logarithmic decrement and is thus a measure of the time in which a free vibration dies out in that circuit. The energy losses can now be taken into account by first determining the distribution of current and voltage as if there were no losses, and afterwards calculating the corresponding losses with this current and voltage distribution, without allowing them to influence that distribution. We shall follow the same method for the flat cavity resonators.

Like the self-induction L^I and the capacity C^I per cm width of ring (*cf.* fig. 7), the resistance R^I per cm width of ring now depends upon the radius r of the ring in question. It becomes

$$R^I = \frac{1}{2\pi r \sigma \delta}, \quad . \quad . \quad . \quad . \quad . \quad (27)$$

[5]) The short-circuited outer edge of the flat cavity resonator may of course be situated at each of these zero points.

where σ is the specific conductivity and δ the depth of penetration due to the skin effect for the frequency in question, which is given by the following expression:

$$\delta = \frac{1}{2\pi}\sqrt{\frac{c\lambda}{\sigma}}.$$

It is evident from this that for a conical flat cavity resonator also the resistance is by no means constant, but varies along the radius. The analogy with the homogeneous Lecher system thus indeed holds except for a factor. 2π, is itself the quotient of the field energy and the quality factor Q, the resonance resistance (in ohms) becomes:

$$Z = \frac{1}{2\pi}\frac{QV^2}{\text{field energy}} = 120\, Q\,\frac{h}{r_1}\,\mu_3, \quad (29)$$

where μ_3 again represents a numerical factor which is shown in *fig. 10c* for different values of n as a function of r_0/r_1.

In *fig. 11* for the case of a flat cavity resonator ($n = 0$) the quality factor Q is shown as a function

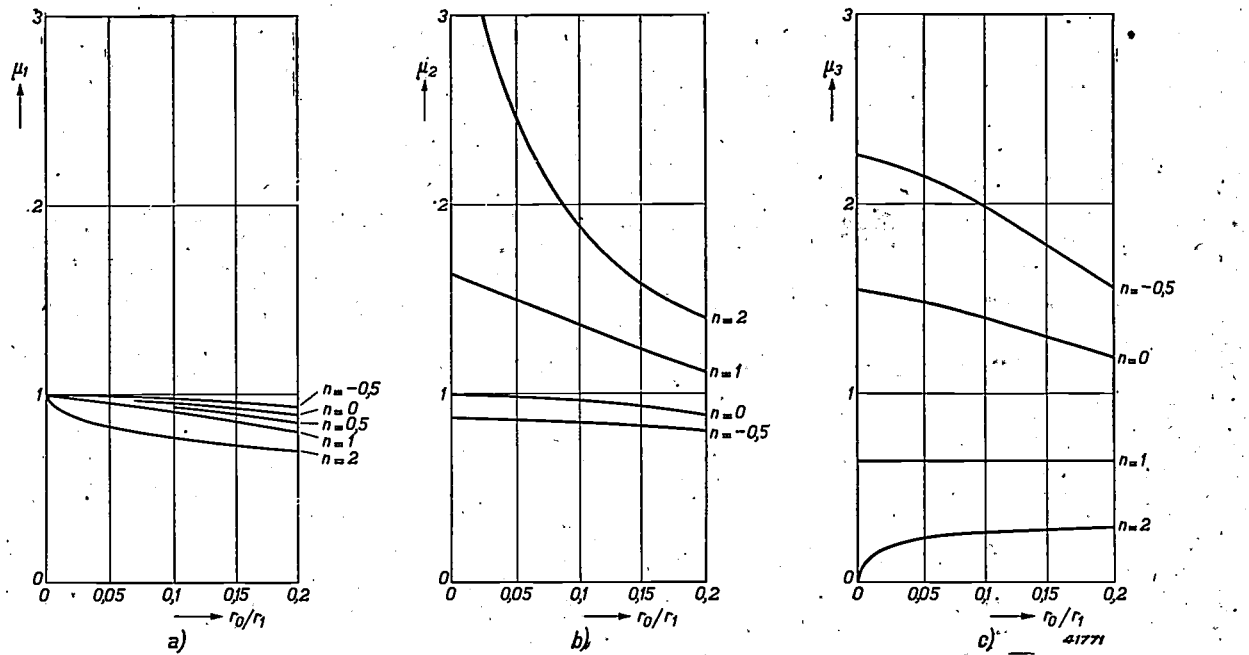


Fig. 10. The auxiliary quantities μ_1, μ_2 and μ_3 for different values of n as functions of r_0/r_1.

for the character of the vibrations and waves, but not for the energy consumption by the conical flat cavity resonator. Upon calculating the quality factor Q complicated formulae are obtained. For the fundamental vibration the quality factor can be written in the following form:

$$Q = \frac{h}{\delta}\,\frac{\mu_1}{\mu_2 + \dfrac{h}{r_1}}; \quad \ldots\ldots \quad (28)$$

μ_1 and μ_2 here represent numerical factors which we have represented in *fig. 10a* and *b* for different values of n, as functions of r_0/r_1.

In the derivation of equation (28) we began with the fact that the quality factor Q is equal to 2π times the quotient of field energy and energy dissipated per period (see footnote [1])).

The resonance resistance of the flat cavity resonator may be considered as the quotient of the square of the effective voltage V for $r = r_0$ and the heat developed in it per second. Since the latter, of the ratio r_0/r_1 of the radii. The continuous (almost) straight line which is nearly horizontal is calculated according to equation (28). The small circles indicate the values determined experimentally. They

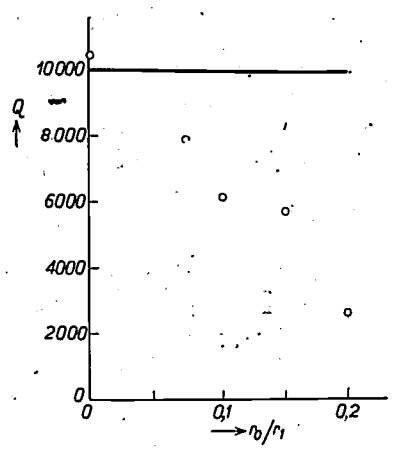


Fig. 11. The quality factor Q as a function of the ratio r_0/r_1 of the two radii of a flat cavity resonator ($n = 0$) with a thickness of 4 cm and an outer radius of 40 cm. For large holes it is found that smaller values of Q are measured than those calculated.

are derived from the resonance widths $\Delta\omega$ and the resonance frequencies ω_0: the quality factor may also be defined as $\omega_0/\Delta\omega$. It may be seen that for large holes the theory yields too large values of Q. This need not be surprising because in the preceding discussion the radiation losses were completely ignored.

The influence of the radiation losses, like that of the heat development, could now be taken into account beginning with the assumption that the radiation, like the ordinary ohmic resistance, does not change the established current distribution for a loss-free flat cavity resonator. It would, however, be necessary to take into account that the currents on the outside of the flat cavity resonator can no longer be disregarded as soon as the hole becomes larger. In the foregoing we have only considered the field inside the resonator. There the electric field is fairly homogeneous and has a vertical direction. Around the hole in the resonator, however, the lines of force are somewhat bent (*fig. 12a*) so that a small part of them even passes from the lower outside surface to the upper outside surface. There are therefore also charges on the outside of the resonator and their intensity varies at the same frequency as that at which the resonator vibrates. Consequently currents occur which flow on the outside but also continue on the inside (fig. 12*b*). In our calculation of the current distribution in the resonator, however, we assumed that $i = 0$ when $r = r_0$, so that a correction must actually be introduced here. This results not only in corrections of the characteristic frequency but also in an increase in the radiation losses. We shall not discuss this rather complex problem any further here.

Improvement of the quality factor and the resonance

As may be seen from formula (28) for the quality factor (Q), this quantity is in the first instance proportional to the ratio between the thickness of the flat cavity resonator and the depth of penetration δ due to the skin effect. Our previous considerations always referred to very thin cavity resonators for which $h \ll r$. If h is taken small enough to justify this treatment the quality factor will often be inadequate. This is not an insurmountable difficulty, since the cavity resonator can as required be made somewhat thicker and then be constructed as being built up of a number of thin resonators, for which the foregoing considerations are immediately valid. By such a piling up of flat cavity resonators, therefore, it is possible to deal also with thicker cavity resonators, without it being necessary to pass over to the general theory of cavity resonators of three dimensions.

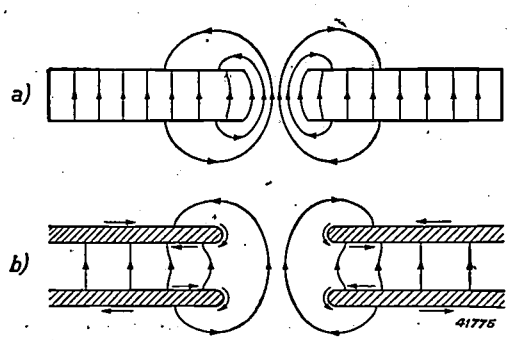


Fig. 12. *a*) Paths of the electrical lines of force between the plates of a flat cavity resonator. *b*) Electric currents induced in the vicinity of the hole as a result of the bending of the lines of force.

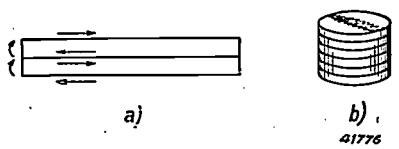


Fig. 13. By the stacking of thin cavity resonators (*a*) more general resonators can be built up (*b*).

As is shown diagrammatically in *fig. 13*, for the case where $n = 0$ a cylindrical cavity resonator can very easily be obtained by the piling up of thin flat cavity resonators, the height of the final resonator no longer being small compared with the radius. In this cylinder, however, there are still partitions which are at the same time the top surface of one of the thin cavities and the bottom surface of the next. Since all these flat cavity resonators vibrate with the same characteristic frequency and all possess an equally intense vertical electric field, the situation is now simply that the currents belonging to the upper and lower cavities in such a partition are everywhere equal and opposite, so that the partitions could be removed from the cylindrical cavity resonator. One point will, however, then be altered; the heat development, which the electric currents in the partitions of the piled up thin cavities would produce, no longer occurs! This reduction of the losses by the removal of partitions is not merely an imaginary experiment but a true fact: as long as the partitions still remain in our cylindrical cavity resonator the heat losses will occur in it, because as a result of the skin effect the currents then actually flow only in the top and bottom layers of such a partition, so that they will by no means cancel each other as far as the development of heat is concerned.

Thus by the removal of the partitions we reduce the energy losses due to heat development compared with the total vibration energy, and it is therefore understandable that for such a more general cavity

resonator obtained by "stacking" we can obtain better values for the quality factor Q and the impedance Z than were possessed by the flat cavity resonator with which we began. If we calculate Q for a cylinder of height h and radius r, we obtain

$$Q = \frac{h}{\delta} \cdot \frac{h}{r_1 + h} \quad \ldots\ldots \quad (30)$$

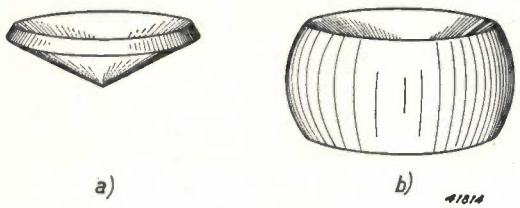


Fig. 14. Thin cavity resonators in the shape of conical shells can also be stacked.

In a similar way, by the stacking of the thin cavity resonators shown in *fig. 14a* in the form of conical shells, we can obtain the cavity resonator shown in fig. 14*b*, which has become very important in short-wave technology. In the discussion of several practical examples we shall also encounter such a cavity resonator (see fig. 20).

Several practical examples

For a copper conical flat cavity resonator (*fig. 15*) with $n = 1$, a thickness $h = 10$ cm at the outside edge and radii $r_0 = 0.6$ cm and $r_1 = 30$ cm, at a wave length of 122 cm, we measured a quality factor $Q = 10\,200$, while the value 11 400 was calculated. For the impedance Z a value of 273 600 ohms was calculated. If in this flat cavity resonator a power of 25 watts is developed in the form of heat, the current at the outer edge is $i = 130$ A, the voltage on the inner edge of the cavity amounts to $V_{r0} = 2600$ V, while the magnetic field strengths at the inner and outer edges are $H_{r0} = 1.4$ gauss and $H_{r1} = 0.9$ gauss. It was actually found that a small transmitter which could deliver a power of 25 W gave a current of about 100 A in the short-circuit ring, which could be concluded from the voltage generated in a loop.

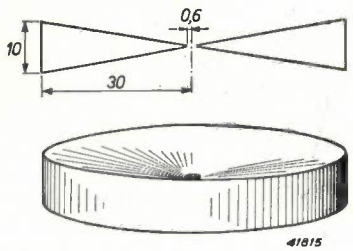
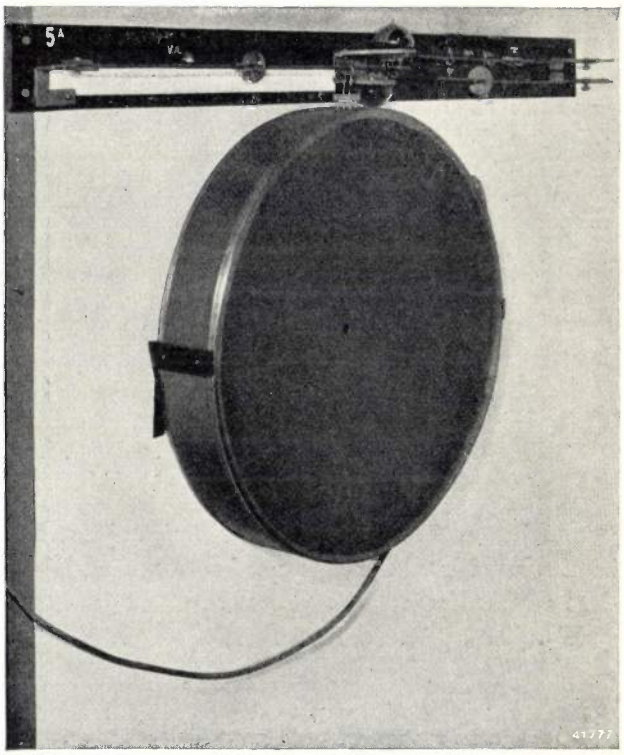


Fig. 15. Conical flat cavity resonator constructed with $n = 1$. The dimensions are given in centimeters.

For the case of a flat cavity resonator ($n = 0$) with a thickness $h = 4$ cm and an outer radius $r_1 = 40$ cm and different radii of the hole: $r_0 = 0$; 3; 4 cm, measurements gave the following values of the quality factor: $Q = 10\,350$; 7 700; 6 100.

Frequency stabilization with flat cavity resonators

Since a satisfactory quality factor Q can be obtained with flat cavity resonators, they can very well be used for keeping the frequencies of oscillators constant. The cavity resonator is then loosely coupled with the oscillator, so that in certain frequency regions the wave length is determined much more by the tuning of the resonator than by the rest of the connections joined to it. In *fig. 16* a conical flat cavity resonator is shown coupled with a Lecher system of variable length l_1. In *fig. 17* the



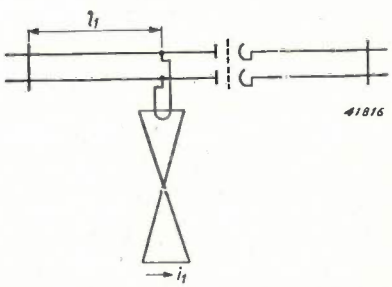


Fig. 16. A conical flat cavity resonator is coupled with a Lecher system of variable length l_1, in order to keep the frequency of the oscillator constant. In the photograph the loop by means of which the current i_1 is measured can just be seen on the upper right-hand edge of the flat resonator.

wave length λ at which this system vibrates and the amplitude i_1 of the current flowing in the short-circuit ring of the resonator are plotted as functions of

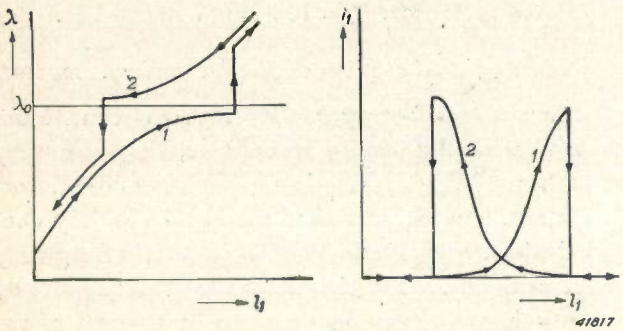


Fig. 17. The wave length λ and the current i_1 occurring with the connections shown in fig. 16 as functions of the length l_1 of the Lecher system. The figures represent a portion of the λ-l_1 plane and the i_1-l_1 plane, respectively, the points $\lambda = l_1 = 0$ and $i_1 = l_1 = 0$ falling outside the figures. Over a wide range the influence of l_1 on the wave length is found to be only slight, since the latter is mainly determined by the wave length λ_0 of the conical flat resonator. It is further evident from the figure that the state of oscillation is not determined unambiguously by the length l_1, but that it also depends upon the way in which a given tuning is reached.

the length l_1 of the Lecher system. Over a wide range the length l_1 of the Lecher system is found to have only very little effect on the wave length λ at which the whole system vibrates. The latter wave length is then determined almost exclusively by the wave length λ_0 of the free vibration of the conical flat cavity resonator. In this figure the disadvantage is also evident that the oscillator frequency is not wholly determined by the circuits alone but also depends upon the way in which the momentary state of the circuits was reached, so that the previous history of the state at a given moment also plays a part. This is a result of the non-linearity of the transmitting valves used [6]).

Cavity resonators as output and input electrodes with short-wave transmitting valves

When the high-frequency oscillation energy is taken from radio valves by means of the anode upon which the electrons must themselves finally impinge with high velocities, in the case of high-power valves a large amount of heat is developed in the anode. In the construction of such valves this should be taken into account, be it at some cost of considerations connected with high frequency. For this reason transmitting valves for very short wave lengths are at present often constructed so that the electrons pass along a pair of rings in which they induce charges. They then give off high-frequency oscillation energy to the rings without themselves striking the rings. The collision energy, however, is taken up by the anode, which is earthed for high frequency and may thus have any desired large dimensions.

In *fig. 18* such a so-called induction tube oscillator is shown. A beam of electrons whose intensity is controlled in a high-frequency rhythm (*fig. 19*) by a grid (*hf*) passes through a slit in a flat cavity

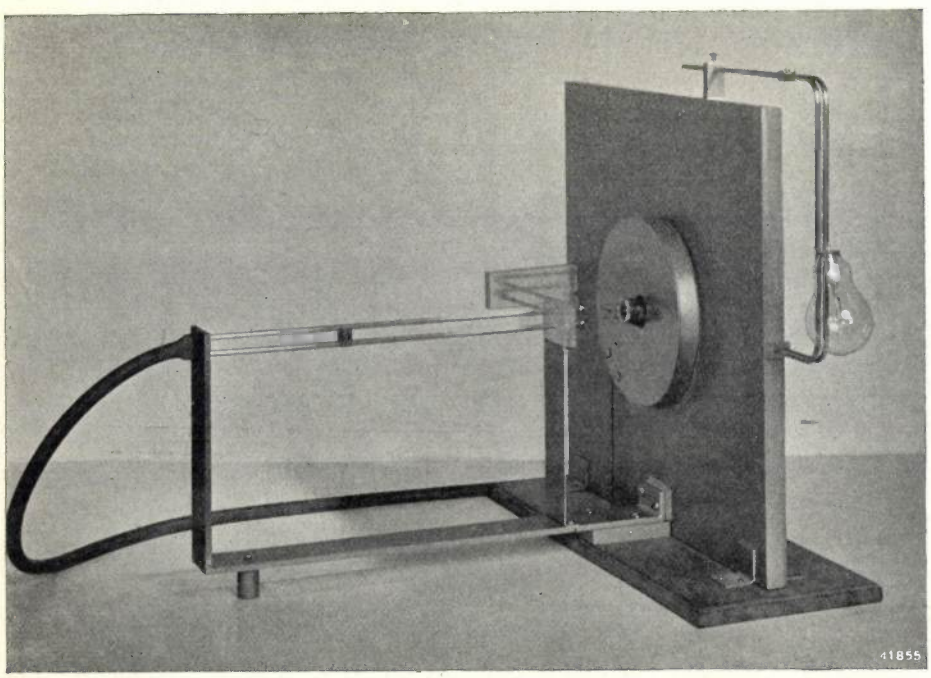


Fig. 18. Induction tube oscillator for waves of 65 cm, equipped with a vertically placed, rectangular flat cavity resonator for taking off the high-frequency oscillation energy. On the left-hand side of the flat resonator may be seen one of the magnetic coils which serve to keep the electron beam concentrated. The Lecher system extending to the left serves for tuning the grid circuit, while the Lecher system extending toward the rear conducts part of the oscillation energy excited in the flat resonator back to the grid, so that the electrical oscillations are maintained. The Lecher system to the right takes from the resonator the energy dissipated in the lamp serving as loading impedance.

[6]) *Cf.* for example Balth. van der Pol, Trillingshysteresis bij de triodegenerator met twee graden van vrijheid (Vibration hysteresis in the triode generator with two degrees of freedom). T. Ned. Radio Gen. **1**, 125, 1921.

resonator (S), upon which electric charges are induced, while the electrons of the beam themselves finally impinge on the anode (a). Seen from the point of view of the cavity resonator, the induction tube seems to possess a fairly high adaptation resistance. In order to ensure a good transmission

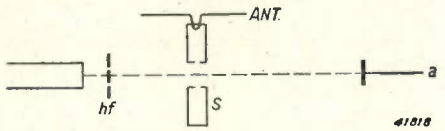


Fig. 19. Diagrammatic representation of an induction tube amplifier with a flat cavity resonator S as output electrode. The electrons are controlled by the high-frequency grid *hf* and move toward the anode *a*. The aerial (*ANT*) is coupled to the resonator, which is excited by the pulsating electron current passing through it.

of the high-frequency energy to the aerial, the impedance of the unloaded oscillation circuit must be high compared with the adaptation resistance. Therefore a cavity resonator with its large impedance is especially suitable to play the part of oscillator circuit here. The high-frequency energy is thereby taken from the cavity resonator by coupling the aerial to it with the help of a loop.

The resonator then functions mainly as transformer between the high-frequency electron beam supplying the energy and the aerial radiating it.

If it is impossible to obtain an adequate quality factor and resonance resistance with the thin, flat cavity resonator shown in figs. 18 and 19 in the induction tube oscillator, these quantities cannot be easily increased by using a somewhat thicker, flat resonator, since in order to keep the electron beam in the correct direction magnetic coils are situated along it at fairly short intervals and a resonator has to be fitted between the coils. It is, however, possible to obtain a larger quality factor and a higher resonance resistance by using a resonator like the one shown in fig. 14*b* whose thickness at the centre remains sufficiently small while at distances farther away from the electron beam the resonator becomes much thicker. In *fig. 20* an induction tube amplifier with such a thicker resonator is shown.

For the generation of high powers at very high frequencies much use is being made at present of a control mechanism with which the velocities of the electrons allowed to pass are varied and not the number. In their further progress the faster electrons will then overtake the slower ones, so that accumulations occur whereby the same effect is attained as with amplitude control. In the application of such so-called velocity modulator valves different cavity resonators must be introduced around the

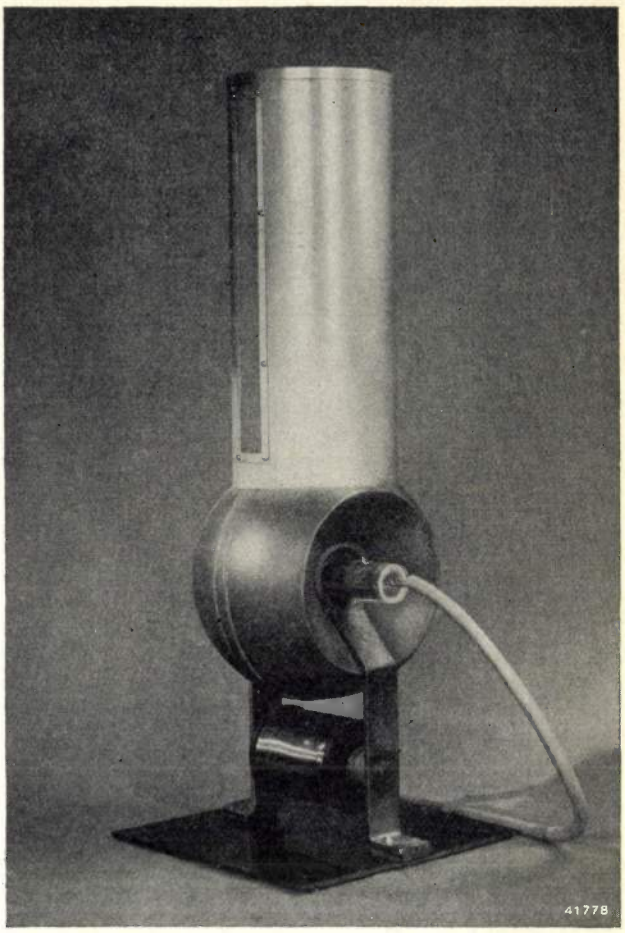


Fig. 20. Induction tube amplifier for 45 cm waves, equipped with a thicker cavity resonator like that shown in fig. 14*b* with which a larger quality factor and higher resonance resistance can be obtained than is possible with thinner resonators. In the cylinder on top of the cavity resonator is an artificial aerial in the form of an incandescent lamp. The coil below the resonator provides the magnetic field which keeps the electron beam concentrated.

path followed by the electrons from cathode to anode. Therefore thin, flat cavity resonators are often used here.

In *fig. 21* a velocity modulator amplifier is shown diagrammatically. In this case instead of being controlled by the grid *hf* of fig. 19 the electron beam is controlled by the slit of the input resonator S_1. Some distance farther along the path of the electrons is the output resonator S_2, so that

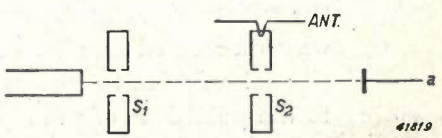


Fig. 21. Diagrammatic representation of a velocity modulating valve amplifier with flat input and output cavity resonators: S_1 and S_2. The electrons, on their path to the anode *a*, are controlled in a high-frequency rhythm by S_1, while then in passing S_2 they excite it in a high-frequency rhythm. The latter resonator passes on its oscillation energy to the aerial (*ANT*).

the electrons will arrive there with a variation in density sufficient to excite oscillations in S_2 with reasonable efficiency. The electrons themselves finally reach the anode a, while the flat resonator S_2 gives off its high-frequency oscillation energy to the aerial coupled with it.

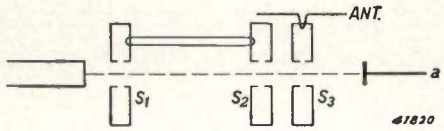


Fig. 22. The flat input and output cavity resonators S_1 and S_2 are coupled to each other in order to make it possible for the system to oscillate independently. The high-frequency oscillation energy for the aerial (ANT) is obtained from the pulsating electron current on its way toward the anode a by means a third flat cavity resonator S_3.

If the flat input and output resonators (S_1 and S_2) are coupled with each other, this system can be made to oscillate independently (*fig. 22*). The high-frequency oscillation energy is then finally taken from this velocity modulator oscillator with the help of a third flat resonator S_3 coupled with the aerial. A demonstration model of such a generator for very short waves is shown in *fig. 23*.

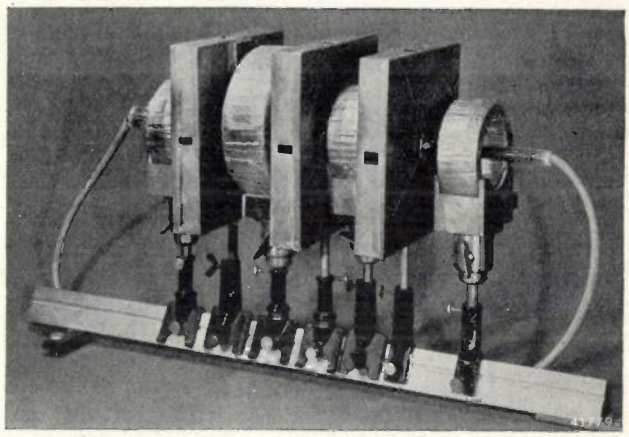


Fig. 23. Velocity modulator for waves of 40 cm in which use is made of three flat cavity resonators (*cf.* diagram of fig. 22). The cathode is on the left and the anode on the right. The size of the opening in the cavity resonator at the right S_3 can be regulated by turning the spokes visible in the photograph in order to change the characteristic frequency. The magnetic coils serve to keep the electron beam concentrated.

---

latest

The ~~last~~ number of *Philips Research Reports* (No. 3 of volume 1, April 1946) contains following articles:

R12: K. F. Niessen: On the error in the determination of the median plane of a radiobeacon in a tilted airplane.

R13: B. D. H. Tellegen: Network synthesis, especially the synthesis of resistanceless four-terminal networks.

R14: H. B. G. Casimir: On Onsager's principle of microscopic reversibility.

R15: M. Gevers: The relation between the power factor and the temperature coefficient of the dielectric constant of solid dielectrics.

R16: T. Jurriaanse, F. M. Penning and J. H. A. Moubis: The cathode fall for molybdenum and zirconium in the rare gases.

R17: G. W. Rathenau and J. L. Snoek: Apparatus for measuring magnetic moments.

Anyone interested in these articles can apply to the Administration of the "Natuurkundig Laboratorium" Kastanjelaan, Eindhoven, Holland, where a limited number of reprints are available for distribution gratis. For subscription to Philips Research Reports application should be made to the publishers of "Philips Technical Review".

VOL. 8 No. 6, p. 161-192 JUNE 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

CONTACT ARC-WELDING

by P. C. van der WILLIGEN.

621.791.753.41

One of the difficulties in electric arc-welding has always been the fact, that the distance between the electrode and the workpiece must be kept constant. It is, however, possible to make electrodes, which can be kept resting on the workpiece, so that the welding can be done by dragging the rod over the workpiece. Philips have now developed a method of touchwelding, whereby the arc is automatically ignited and if necessary reignited. This is achieved by transferring a large part of the metal from the core wire in the form of a fine powder to the originally insulating coating, which thereby becomes more or less conductive. Since the electrode can be kept resting on the workpiece right from the start and, during the welding, is continuously in contact with it both mechanically and electrically, we speak of contact arc-welding and contact electrodes. With these contact electrodes welds can be made quicker and more reliable, than with ordinary electrodes. Furthermore it is also possible to use the free arc with these electrodes.

The method of electric arc-welding widely used to-day, where the deposited material comes from the electrode, which is held close to the workpiece, was first introduced by Slavianoff in 1892. In those days direct current was used exclusively, since, with alternating current of a safe voltage, the arc between the iron electrode and the workpiece was always being extinguished. Only after 1908, after Kjellberg had introduced a coating of slag-forming substances around the rod, was it possible to use alternating current for welding. The coating emits so many electrons, that the welding arc, which with A.C. of 50 cycles is extinguished 100 times per second, continues to burn; or, more accurately, it is reignited 100 times per second. The coating, however, has a still more important function, a metallurgical one. It improves very much the mechanical properties of the deposited metal, which properties are poor when uncoated wire is used. In the course of years this improvement has proceeded so far that the mechanical properties of the deposited material of the most modern electrodes, such as the Philips type Ph 55, are equal to those of the best kinds of structural steel.

As far as welding with A.C. and the mechanical properties are concerned, the introduction of the coating has meant a great advance. Arc welding, however, has always remained an "art", which can only be mastered by much practice. The introduction of the coating has indeed made it easier to do overhead welding. The cup formed during the welding serves to direct the molten drops of metal, just as the barrel of a gun directs the projectile [1]). As to welding in the down-hand positions, however, the coating has scarcely made it any easier. While on the one hand the welder has the convenience of being able to check the speed of travel by following the slag behind the arc, on the other hand the coating and the slag hide the melting-down process and the melting pool from his view.

Touch-welding

The great difficulty in free arc welding has always been in maintaining the arc. The welder must continually take care, that the distance between electrode and workpiece is only a few millimeters and that it remains constant.

When the distance is too small the rod "freezes" to the workpiece, *i.e.* the solidifying drops form a rigid connection. When the distance is too large, the arc burns irregularly and the quality of the work suffers, and with A.C. there is much chance of the arc being extinguished.

[1]) *Cf.* for example J. Sack, Philips Techn. Rev. 1, 26, 1936 and 4, 9, 1939.

In the course of years, however, it has been noticed that with certain types of electrodes the coating can rest on the workpiece during the welding; this is called touch-welding.

It was found that this procedure was only possible with heavily coated electrodes. This can be understood, when one considers that in welding with a rod with a thinner coating only a short cup is usually formed around the arc. When the edge of such a cup is allowed to rest on the workpiece, the drops of metal flowing from the core wire short circuit and freeze the rod to the workpiece.

only would this, however, be an expensive method, but at the same time difficulties would occur in the welding because of the larger amount of slag.

It is to be pointed out that fig. 1 is only a rough sketch; for instance the formation of the crater (a small dimple in the workpiece underneath the arc) has been disregarded. When high currents are used this crater is rather deep and in that case touch-welding with electrodes of type Ph 50 and 55 easily succeeds. It has never been possible to arrive at a general application of touch-welding because of the limitations just mentioned and some others.

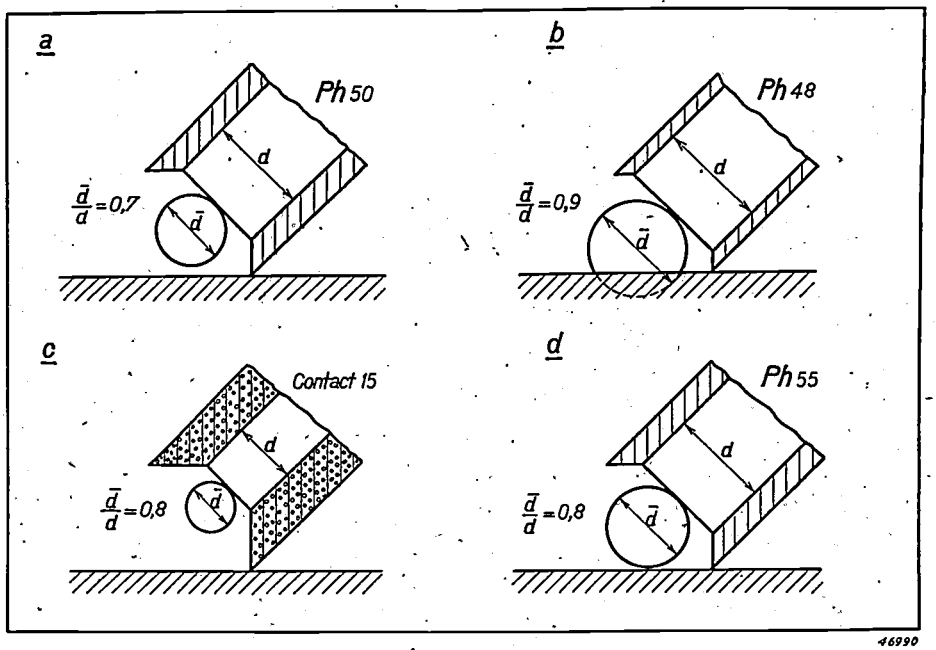


Fig. 1. Diagrammatic representation of the possibility or impossibility of touch-welding for different kinds of electrodes manufactured by Philips. The diameter of the iron core is d; $\bar{d}$ is that of the droplet. For the method of determining the size of the droplets see J. Sack, The Welding Industry, July 1939.

If one tries to determine the relation between the thickness of the core wire and that of the coating necessary to prevent this, it is found that it depends upon various factors. If we assume that the welding is done at an angle of 45° to the workpiece and that the cup is formed at an angle of 45° to the axis of the electrode, we see from *fig. 1a* that the relation between the size of drop of the molten metal and the dimensions of the cup is such, for instance with the electrode type Ph 50, that touch-welding is just possible, while with type Ph 48 (see *fig. 1b*) it will be quite impossible.

From this the conclusion might be drawn, that to make touch-welding possible it would be sufficient to give the electrodes a heavy coating. Not

The new development: contact arc-welding

For all these reasons, research work was carried out in this laboratory with the object of perfecting the properties necessary for touch-welding, to such a point, that welding could be done with much less trouble and that practically no errors could be made in welding. This includes the requirement that the weld bead should have a flat and uniform appearance. This investigation has now led to a new method of welding for which the name "contact arc-welding" was chosen, for the reasons which will be explained. We will first sketch the development of this method. Although, in principle, other types of electrodes could have been used in this development, we chose in fact the Ph 55 as

starting point, because its mechanical properties surpass those of other types of electrodes (*i.a.* because it has a low oxygen and nitrogen content). Ductility and impact value [2]) are very high. By using the Ph 55, therefore the chance of cracks in the weld is small, which is especially important in welding hard steels and also dynamically heavily loaded structures. For welding hard steels and steel containing much sulphur it is, moreover, important that the electrode Ph 55 is insensitive within wide limits to the content of carbon and sulphur of the workpiece [3]).

The fact that, in spite of all these good properties, the electrode Ph 55 is not universally employed is due to the fact, that much care is required to make good welds with this type of electrode. If, for instance, the welder uses too long an arc, porosity may result in the weld. If a welding transformer is used with a low open voltage (< 60 V) difficulty is experienced from the extinguishing of the arc. The welding is thereby interrupted and has to be started again, which leads to inhomogeneities in the weld.

Therefore, the electrode Ph 55 had to be altered in such a way as to overcome these objections, while still retaining the important advantages mentioned. The first step in this direction was to try to perfect the quality necessary for good touching, without increasing the quantity of slag-forming coating material. This was achieved by applying the principle of making the coating heavier, (which principle, of course, is by no means confined to the use of the Ph 55 rod) by transferring part of the metal of the core in a finely divided form to the coating. The ratio of metal to slag in the welding rod may remain the same. A limit is set, to the transfer of metal from the core to the coating of the rod, by the consideration that the core, which has to carry the current, may not be too thin, as otherwise the resistance becomes too high and the permissible arc current too low.

In the case of the newly developed electrodes, which are denoted by the word "Contact" (thus for example: Contact 15), since they are in continuous contact with the workpiece, about half of the metal of the core is transferred in scattered form to the coating. In *fig. 1c* it is indicated diagrammatically that, with electrode Contact 15, touch-welding should be quite easy. The external diameter of the electrode is more than twice as great as the diameter of the core, so that the coating is extremely heavy. For the sake of comparison we have shown in *fig. 1d* an ordinary electrode Ph 55, which has the same external diameter and the same weight. These two electrodes deposit the same amount of iron in the same time.

Different kinds of electrodes are distinguished not only according to their type, but also according to the thickness of their core, which follows the name of the type. This thickness determines the arc current and in the case of ordinary electrodes also the amount of metal deposited per second. It must be borne in mind that, in the case of the new contact electrode, only half the molten metal comes from the core, the other half being supplied by the coating. In the case of the electrode Contact 15, therefore, the diameter of the core is $\sqrt{2}$ times smaller than that of an electrode Ph 55 depositing the same amount of metal. The new electrode Contact 15-5, therefore, corresponds to the ordinary electrode Ph 55-7.

With the electrode Contact 15 the ignition of the arc has also become much easier. It is not done in the usual way by tapping and breaking off the cup, but entirely automatically. This simplification in starting is closely connected with the new composition of the coating.

In the development of the contact electrodes, as we have seen, much metal has been transferred from the core to the coating. When the amount of metal, transferred to the coating, is approximately the same as that remaining in the core, the conductivity of the coating is found to have become so high that when the coating, under tension, is brushed over the workpiece sparking occurs, such as takes place from the trolleys or shoes in electrical traction when there is icing. A current is then flowing of the order of magnitude of 0·1 A. When the coating is not brushed over the edge, but is held against a definite spot on the workpiece, the current will rise rapidly and at a value of about 1 A will suddenly pass to an arc: the welding arc is then automatically ignited. The oscillogram of the current flowing with Contact 15 is illustrated in *fig. 2*; the arc is usually ignited within one second. When in the hot condition, ignition takes place even more quickly, and obviously the voltage applied also has some influence on the speed of ignition.

Thus the electrode Contact 15 has been made self-starting. Not only has it become unnecessary for the welder to start by tapping, but the reignition, after an interruption of a few cycles, also takes place automatically. Such an interruption

[2]) P. C. van der Willigen, Philips Techn. Rev. 6, 97-104, 1941.

[3]) J. ter Berg, Philips Techn. Rev. 7, 91-93, 1942.

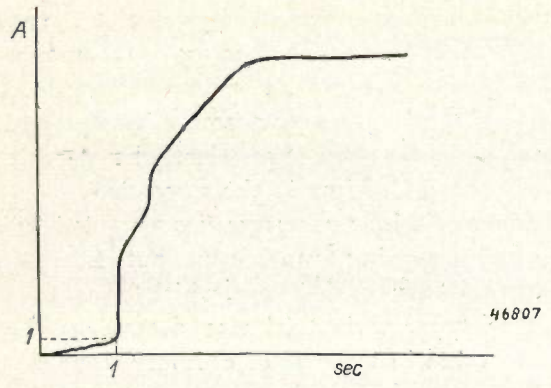


Fig. 2. Oscillogram of the current trend when the arc of Contact 15 is being ignited. The arc is struck at approximately 1 amp. and 1 sec.

may take place accidentally and result in the extinction of the arc, when a welding transformer is used with a low open voltage. The conductivity of the coating of the new electrodes then ensures, that the arc is immediately reignited. As a result with electrodes of the type Contact 15, which have a core diameter of less than 5 mm, it is possible even to touch-weld with open voltages of the welding transformer lower than for the corresponding electrodes Ph 55, *i.e.* they need not be so far above the arc voltages. The contact electrodes 15-3.25 and 15-4, which have core diameters respectively of 3.25 and 4 mm, have for example an arc voltage of 25-30 V, and it has actually been found that touch-welding can, very succesfully, be done with them at an open voltage of 60 V. In some cases, it is found possible to work succesfully at lower open voltages (to 50 V) at which, if the corresponding normal electrodes were used, there would naturally be trouble from extinction. For the new electrodes with core diameters of at least 5 mm (thus corresponding to normal electrodes of at least 7 mm core diameter), however, higher voltages are necessary.

The new electrode Contact 15-5, for instance, has an arc voltage of 40 V. If a welding transformer is used with an open voltage, which lies only slightly above this, the transformer works too much in the flat part of its characteristic, which means, that only small variations in the voltage will already result in large current fluctuations.

The best results are obtained, when the open voltage of the welding transformer lies about 30 V higher than the arc voltages, thus, for the electrodes Contact 15-5, this will be 70 V. Welding then takes place in the steep part of the characteristic, *i.e.* in the part, where the voltage depends closely on the current, so that any variation in the arc voltage scarcely affects the current.

The advantage of contact arc-welding therefore lies in the possibility of easy touch-welding and the automatic ignition and re-ignition of the arc. Moreover, with a given current the amount of metal deposited per unit of time is considerably greater, as will be discussed in a subsequent article.

It has been found, that these new contact electrodes are not only suitable for contact arc-welding, for which purpose they were developed, but that they also offer important advantages for welding with the free arc. The contact electrode gives much less extinguishing of the arc, even when the open voltage of the welding transformer is lower than normal. This is probably the result of the great conductivity of the coating, which now and then touches the pool when the cup is so long.

The use of contact electrodes

It is remarkable, how easy it is to weld with the contact electrodes. This is due not only to the ease of touch-welding, but also to the self-starting. If one must start at a precisely fixed point it is a great convenience, if the starting is automatic, especially when using holders with a push-button mechanism. The welder places his electrode tip on the predetermined spot, puts the shield before his eyes and presses the button (see

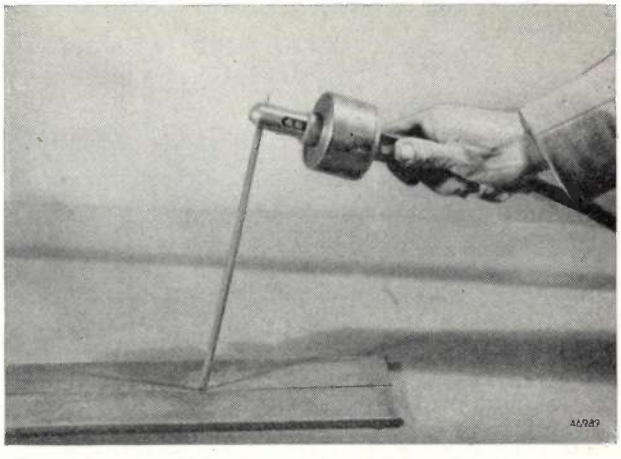


Fig. 3. Photograph of electrode holder specially designed for contact arc-welding. Starting is by pressing the button under the thumb, which closes the circuit.

fig. 3). It is a great convenience for beginners too that there is no longer the trouble of tapping and striking the arc.

As we have already stated, practice is necessary in order to maintain the proper distance between

[4]) These two new electrodes would correspond in weight to normal electrodes of the type Ph 55 with core diameters of 4·7 and 5·6 mm respectively, which, however, are not manufactured.

core and workpiece when welding with the free arc, and it is found, that this even presents difficulties for experienced welders, when welding has to be done in difficult positions, so that they too may then have trouble with freezing. With the newly developed electrodes, however, the coating has been made so thick that the cup of the rod is always deep and strong enough to keep the core in touch-welding, at such a distance from the workpiece, that it cannot freeze [5]).

The mechanical properties of the material, deposited [6]) by these new electrodes Contact 15 and Contact 18, are found to be similar to those of the corresponding normal electrodes Ph 55 and Ph 48. The fact that, in the case of the contact rods, the arc is for the greater part surrounded by the deep cup is a favourable factor, diminishing the absorption of oxygen and nitrogen from the air.

While it is usual to weld from left to right, welding from right to left also often occurs in practice, and left-handed welders were desired for such work. With contact electrodes this work, which is no more difficult than melting a candle at a uniform rate against a hot plate, can also be done by right-handed welders, namely with the back of the hand turned towards the workpiece, called "backhand" welding, or by holding the holder in the left hand, which, in the case of the more difficult normal welding can, only be done succesfully by left-handed welders.

Just as with normal electrodes, with contact rods care must also be taken to prevent the slag running ahead of the arc. To accomplish this, in contact arc-welding, the angle between electrode and workpiece must not be too large. In the down-hand position an angle of 40 to 45° is best. If a smaller angle is chosen there is more sputter. Another disadvantage of a very small angle is, that the arc does not "dig" so well into the root of the weld, which is often of importance, when using these rods with deep cups. In practice, it sometimes happens that it is impossible to maintain the desired angle of 40 to 45°; for instance because of the proximity of a transverse partition. It is then possible, to weld from right to left or to use the free arc. One can then weld in such a way that the slag remains behind the arc, although the rod makes a large angle with the workpiece.

The most convenient position for making a welded joint is the flat position, where the surface

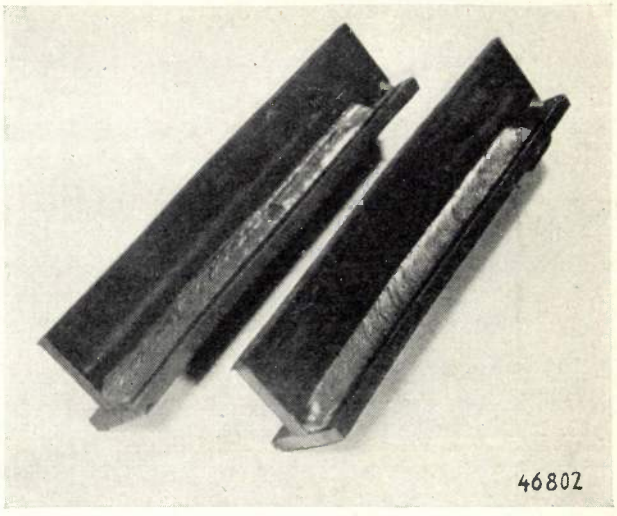


Fig. 4. This photograph shows two fillet welds made in the flat position with Contact 15-5; that on the right, made with D.C., has a smoother surface than that on the left made with A.C.

of the bead is in the horizontal plane (*fig. 4*). In this position the heaviest electrodes can be used without difficulty. This is true, not only for the methods of welding so far commonly applied, but also for contact arc-welding, with which very fine results can be obtained. *Fig. 4* shows two fillet welds made with Contact 15-5, on the left with 360 amp. A.C. on the right with 310 amp. D.C.

For fillet welds, made in the horizontal position, where one plate is horizontal and the other perpendicular to it, Contacts 15-3·25 and 15-4 can be used. With heavier electrodes the weld is strongly convex in form.

A most remarkable achievement is the ease with which overhead welding can be done [7]), with Contacts 15-3·25 and 15-4. This is the most difficult position for welding, and, as far as is known, it was never possible to apply touch-welding with success in this position. Overhead contact arc-welding, however, is found to be very easy. *Fig. 5* is a photograph of a weld made in this position with Contact 15-3·25 and using 140 amp. D.C. In overhead welding experienced welders sometimes prefer to use the free arc, for instance in order to make the weld

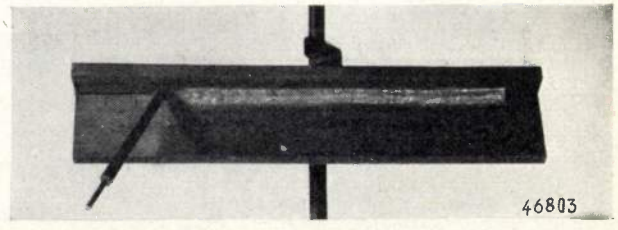


Fig. 5. Overhead weld made with Contact 15-3·25.

[5]) In the special case, where the electrode is cooled after being partly used, it may happen that non-conducting slag is formed on the edge of the cup and prevents the arc from starting. This slag can, however, very easily be broken off by tapping.

[6]) See the article cited in footnotes [1]) and [2]).

[7]) *Cf.* for example Philips Techn. Rev. 4, 9, 1939.

somewhat wider by means of a weaving motion. Also in this respect Contact 15 offers advantages over the normal coated electrodes, the most important of which results from the deep cup of Contact 15, which gives better direction to the drops; consequently very little material falls to the ground. When, however, Contact 15-5 and larger sizes are used, the pool becomes too large and the molten iron falls out of it, so that Contact 15-4 is the heaviest electrode that can be used for overhead welding.

In work on ships it is often necessary to make a horizontal weld in the vertical plane (the ship's side). This work can be done very well with Contacts 15-3·25 and 15-4. In vertical-up positions the electrode should, in general, be moved sideways (weaving) perpendicular to the direction of the weld, in order to obtain a sufficiently wide bead, the fused material being left to solidify at the desired spot; thus the great advantage of contact arc welding, namely that no precisely regulated motions need be made, does not apply in this case. In vertical-up welding, therefore the free arc will always be used, but the electrodes already mentioned can be used here to advantage. *Fig. 6* shows a good vertically-up weld made with the free arc using Contact 15-3.25 and a welding transformer with an open voltage of only 50 V. In vertical-down welding Contact 15 generally gives rise to an excessive amount of slag. The welder must try to keep the slag above the welding arc, because, if it is allowed to run down, it pushes the arc and the molten metal aside, resulting in faults in the surface of the weld. For these reasons a contact electrode has been developed with a thinner coating, namely Contact 18. Taking as basis Ph 48, which is very much used for vertical-down welding, the same principle has been applied here as in Contact 15. Thanks to the small amount of rapidly solidifying slag, it is easy with Contact 18 to keep the slag above the welding arc. Welds made in this manner have a particularly good appearance. Since the surface of the bead is very smooth, the slag often loosens of itself. This contact electrode, like Contact 15, is suitable not only for D.C. but also for A.C. The weld of *fig. 7* was made with the help of Contact 18-4.

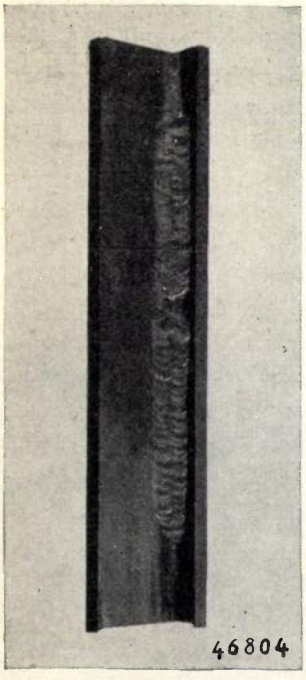


Fig. 6

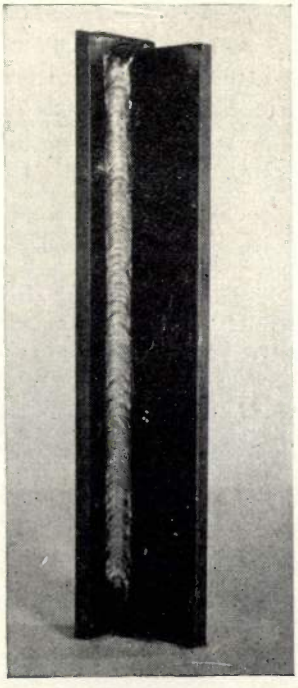


Fig. 7

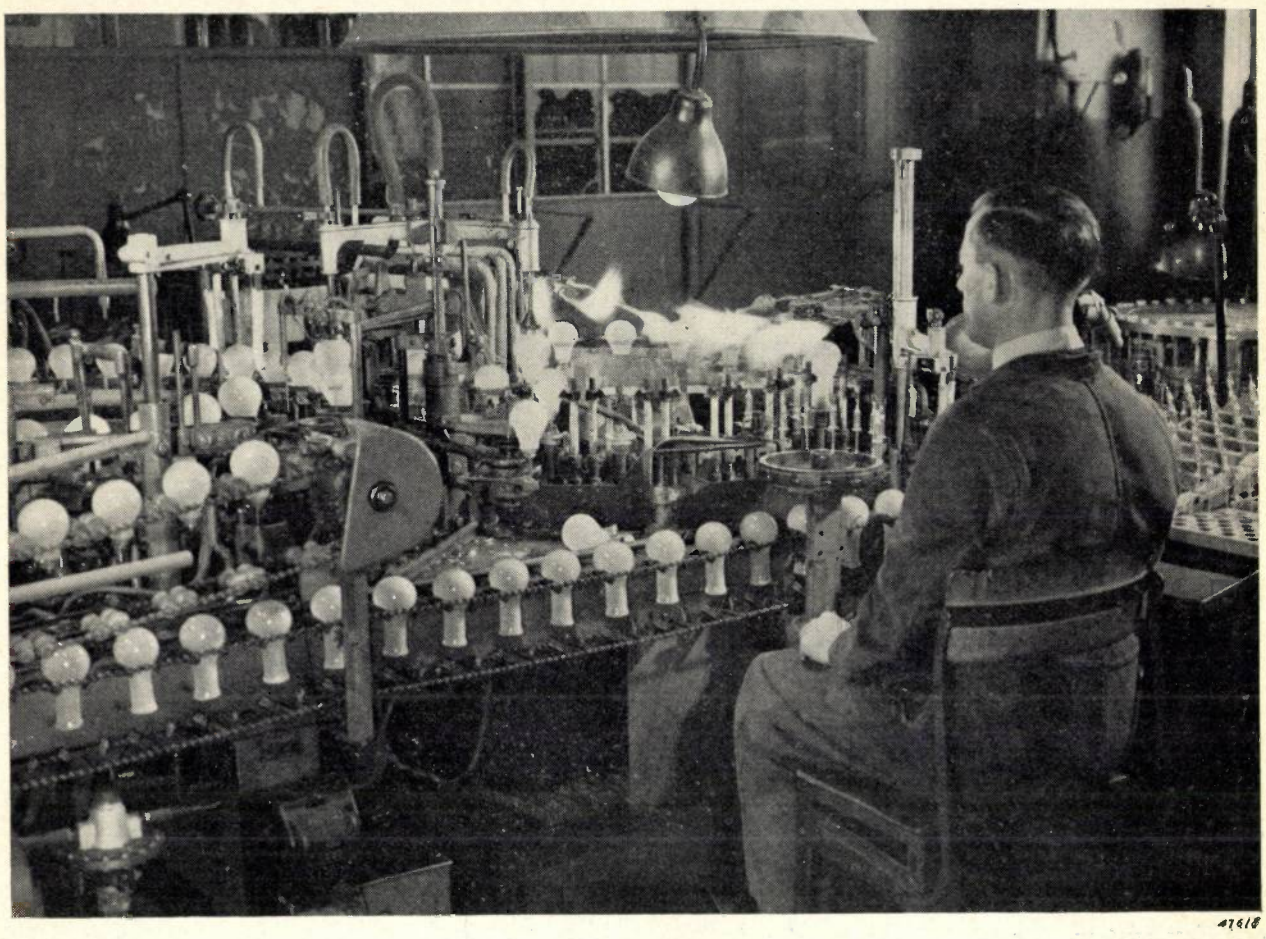
Fig. 6. Vertical-up weld made with the electrode Contact 15-3·25. In this case the free arc was used, since contact arc-welding offers no advantage here, where the welding has to be done with a weaving motion. The open voltage of the welding transformer was only 50 V.
Fig. 7. Vertical-down weld made with Contact 18-4, which has a thinner coating.

In conclusion, we shall make a few remarks about welding with high currents, where heavy welding rods are used, and which in the United States is called "hot welding". The currents used are higher than 600 amp. and the diameters of the electrodes greater than about 7 mm. As yet, however, this method has not met with much success, for the following reasons:

a) Handling heavy electrodes becomes fatiguing for the welder.
b) There is much spatter with these rods.
c) Heat radiation becomes troublesome, especially in hot weather.
d) Difficulty is very often experienced from undercut, the plate material being burnt away along the toe of the weld.

If Contact 15-6, 15-8 or 15-10 is used for this purpose (the last corresponds to Ph 55-14 and is used with 900-1200 amp.), the first difficulty is overcome, since in contact arc-welding the electrode rests on the workpiece. Due to the deep cup shielding the arc, also spatter is considerably reduced; tests showed spatter losses only one-third of those obtained with the corresponding electrode Ph 55. For the same reason, heat radiation from the arc is lowered, especially in the direction of the weld, since in that direction the cup almost completely shields the arc. Finally with Contact 15 undercut is eliminated as a result of the special form of the penetration, to be discussed in a later article.

SERIAL GROUP FOR THE MANUFACTURE OF INCANDESCENT LAMPS



An incandescent lamp factory originally comprised a number of departments housed in large workshops, where the various stages of the manufacturing process were entirely separate from each other. The stages were: the construction and mounting of the stem, *i.e.* the internal glass part with current leads, pump tube and filament; the sealing of the base in the bulb along the lower flanged rim; the lamp thus built up then has to be evacuated or filled with gas (Argon), during which process the lamp was heated to degas the component parts; finally in a fourth department the base was cemented to the lamp.

Obviously, in course of time these processes became highly mechanised. There were certain advantages attaching to the separation of manufacture among different workshops, owing to the specialization of certain parts of the processes. The drawbacks, however, among which was the transport problem, were much weightier; for instance a great deal of administrative work and care was required in order to keep together the various parts of a certain type of lamp in passing from one shop to another.

The photograph reproduced here shows how all this has been changed. All the processes mentioned are combined in this one single unit, called the serial group. The conveyor, clearly seen in the foreground, brings in the bulbs. Mechanical hands automatically pick up a bulb and a mounted stem and place them together on the sealing machine. This machine rotates and takes up a number of positions, in each of which a certain part of the sealing of stem to bulb takes place on the turntable immediately underneath the lamp. The part of the machinery on the left of the photograph is where the pumping, gas-filling and further finishing is done.

Such a serial group turns out a complete lamp of a certain type with the minimum of manual labour, all the component parts being assembled together automatically, while the various operations are arranged in the proper sequence by a very carefully studied timing.

SIGNALLING IN CARRIER TELEPHONY

by F. A. de GROOT.

621.395.63:621.395.44

The establishing of a connection between two subscribers of a telephone network requires the transmission of different signals from the subscribers' instruments to the exchange and *vice versa*. In local networks this signalling is done partly with direct current. Since, however, in the case of long trunk lines, and especially in the case of carrier connections, this is impossible, other methods had to be developed. After a short discussion of signalling with alternating current of audio-frequency, the Philips system of signalling with the carrier waves is considered. In the description of the practical execution of this method of signalling special attention is devoted to the manner, in which interferences with the signal and by the signal are avoided, and how the distortions, which the signals may experience upon transmission, can be counteracted.

In order to carry on a telephone conversation, it is not enough to have available an installation which is capable of transmitting speech clearly and without interferences. Each subscriber must also be able to ring up any other subscriber. To do this, he must be able to inform the telephone exchange, that he wishes a call to be put through; the exchange must prepare the connection, warn the subscriber called, inform the subscriber calling, whether or not the desired line is already engaged, etc.; finally after the conclusion of the conversation the exchange must be warned, so that everything can be returned to a position of rest.

All these and any other warnings and communications are accomplished by certain signals, and the complex system of aids, which serve for the excitation and transmission of these signals, is called the signalling system.

A signalling system is necessary, not only for local, but also for trunk connections. The problems thereby involved, especially when carrier telephony is used for the trunk connection, will be discussed in this article. We shall deal especially with a signalling system, developed by Philips, in which the carrier waves themselves are used for the signalling. We shall begin with a simplified description of what happens, in a local telephone conversation, between two subscribers on an automatic exchange.

Signalling in a local connection with automatic exchange

In *fig. 1* the circuit diagram is given of the connection between two subscribers *a* and *b*, omitting all non-essential details.

In the state of rest, the telephones of the two subscribers hang on their hooks, the contacts H_a and H_b are then open, the lines carry no current, the relays R_a and R_b are released. Upon *a* lifting up the receiver, the contact H_a is closed and a direct current I_a, furnished by the battery E, begins to flow through the line and the apparatus. This direct current (which, moreover, also serves to feed the carbon microphone M) energizes the relay R_a and the exchange is by this means informed, that *a* wishes to make a call. The exchange, by means of a switching process, which may be

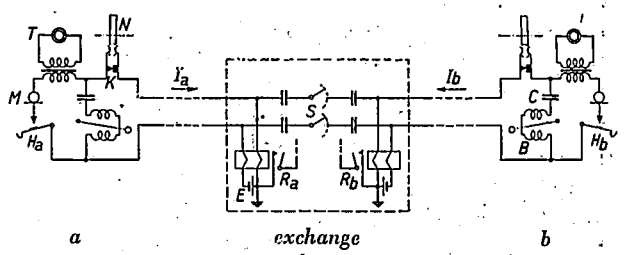


Fig. 1. Simplified diagram of the signalling system in a local telephone connection with automatic exchange. M microphone, T telephone, H_a and H_b hook contacts, B bell, C condenser, N cam disc, K contact of dial, I_a and I_b line currents, R_a and R_b relays, S selectors.

disregarded here, now makes itself ready to "hear", with whom *a* wishes to speak. The fact, that the exchange is ready for this, is communicated to *a* by a so-called dialling tone, which is sent over the line by the exchange. Upon hearing this, the subscriber begins to dial, *i.e.* he turns his dial and lets it run back; during this running back the line current is interrupted (by means of the cam N and the contact K), this being repeated as many times as there are figures in the number dialled. Upon each interruption, the relay R_a releases and thus passes the impulses through to the automatic selectors set up at the exchange, which thus select the line of the subscriber being called. As soon as this line has been found, it is ascertained, whether it is free (by whether its relay R_b is "up" or not). If it is not free, the exchange sends over the line to *a* the "engaged" signal. If the line is free, the exchange sends an alternating current with a fundamental frequency of 16 c/sec to *b*, the so-called ringing current. The hook contact of this subscriber is of

course open, but the alternating current passes the condenser C, which is tuned with the self-induction of the bell to 16 c/sec and causes the bell to ring. At the same time the exchange sends a ringing tone over the line to subscriber a, as a sign, that b's bell is ringing. As soon as b removes the receiver direct current also begins to flow in his line, the ringing is stopped by the operation of the relay R_b and the conversation can take place. When the conversation is ended, the subscribers hang up their receivers again, the direct current in the two lines is thereby interrupted, the relays R_a and R_b are released and as a consequence the selectors and the various relays in the exchange return to their position of rest.

In *fig. 2* all these happenings are represented graphically, by the variation with time of the

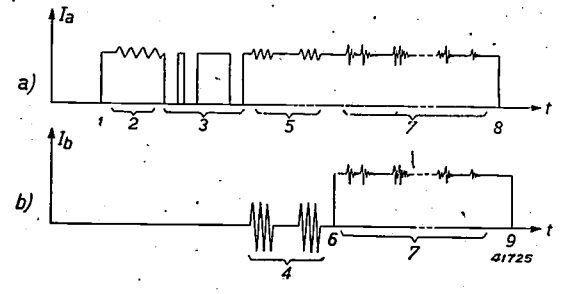


Fig. 2. Behaviour of the line currents I_a and I_b during a telephone call. *a*) Line of the subscriber making the call, *b*) line of the subscriber called. *1* Calling signal, *2* dialling signal, *3* dialling impulses, *4* ringing current, *5* ringing tone, *6* answering signal, *7* speech currents, *8* and *9* clearing signals.

currents I_a and I_b on the lines of the subscriber calling and the subscriber called respectively. It is seen that the signalling comprises the following signals [1]):

The calling signal (*1*) switching on the direct current I_a).

The dialling tone (*2*).

The dialling impulses (*3*) interrupting the direct current I_a.

The ringing current of 16 c/sec (*4*).

The ringing tone or engaged tone (5).

The answering signal (*6*) switching on the direct current I_b.

The clearing signal (*8-9*) switching off I_a and I_b.

Signalling in trunk traffic

More and more use is being made of automatic dialling, also for trunk traffic. In the Netherlands, for example, the telephone network in important parts of the country has already been made automatic. In principle, the manner of establishing the connection is the same, as described above, for a local call. Suppose, that the subscriber a in A wishes to speak with subscriber d in D. He lifts up the receiver, waits until he hears the dialling tone of his local exchange A and then dials a certain number, upon which the local exchange connects him with the trunk exchange B (see *fig. 3*).

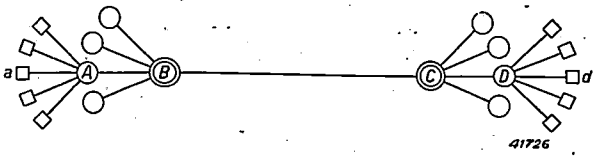


Fig. 3. Diagram of a trunk connection between subscriber a connected with the local exchange A and subscriber d connected with the local exchange D. Various neighbouring local exchanges are connected with a trunk exchange (B and C). The line, between local exchange and trunk exchange, is still considered as "local" line. Between the trunk exchanges B and C, which are far apart, lies the trunk line proper.

In this exchange, as a result of new dialling impulses, the selectors find a line to the trunk exchange C. As soon as this line is found, a warning is sent to C (a new calling signal). The subscriber now continues to dial and the impulses, then following, have to be transmitted over the trunk line to C, to cause the selectors there to transmit to a line, connecting C with the local exchange D. This exchange now sends the dialling tone back to subscriber a, who now with his dial sends impulses over the whole connection to the selectors in D, which find the line of subscriber d. The exchange D then sends the ringing current to d in the manner already described and at the same time the ringing tone or the engaged tone, as the case may be, is sent back to a. After the conclusion of the call, a clearing signal from B to C (or *vice versa*) is again necessary to break the connection and bring the apparatus back to the position of rest [2]).

It is thus clear, that practically all of the signals mentioned for local connections have to be transmitted also over the trunk line: calling signal, dialling impulses and clearing signal as signals intended for the apparatus, further dialling tone and ringing or engaged tone as signals intended

[1]) In addition to those mentioned here, there are other signals, such as those for counting the calls, measuring the duration of a call, etc. Since these do not involve anything new in the problems to be considered here, they will be left out of consideration.

[2]) The course of events sketched here is characteristic of the so-called direct system. There exist also other dialling systems (registering systems), but the signals they require are in principle the same as those with the direct system.

for the ear of the caller, to let him know what the situation is [3]).

Different methods of signalling

We have seen that for signalling on the local line, as far as the signals intended for the apparatus are concerned, use is made of direct current, which is in any case required for the microphone. Direct current is also used for short-distance trunk connections. For long-distance connections, however, it is difficult to obtain a satisfactory solution in this way: the transformers and any necessary repeaters in the line cannot transmit the direct current directly, so that every transformer and amplifier would have to be shunted by a relay system. Through coil-loaded cables, which are often used for long distances, no direct current at all may be transmitted, because the magnetization thereby caused is detrimental to the properties of the loading coils.

In carrier telephony too, direct current cannot be used for signalling, because here a number of calls are sent through one pair of conductors and only one direct current can be sent through a pair of conductors. The signals therefore have to be transmitted in some other way, and, for every speech channel, there has to be a signal path independent of the other channels.

The method so far commonly used in these cases is voice-frequency signalling. The D.C. signals occurring on the local line are converted in the trunk exchange into A.C. signals with a frequency in the audible region, between 300 and 2800 c/sec. Since the trunk line must in any case transmit this frequency region, either directly or modulated on a carrier wave, the A.C. signals can be immediately passed on over the trunk line. At the receiving end the signals are converted back into D.C. signals by filtering out the signal frequency and activating a relay with the signal voltage.

In voice-frequency signalling the calling and clearing signals, which occur on the local line, when the direct current is switched on and off respectively, cannot be transmitted in the form in which they occur as the result of the switching on and off of the alternating current, because this alternating current would then have to be maintained during the whole conversation and would be heard by the subscribers as a disturbing whistle. Calling and clearing signals are therefore now given by A.C. impulses of a certain length. One speaks in this case of impulse signalling in contrast to continuous signalling with a direct current, where the signal current continues to flow during

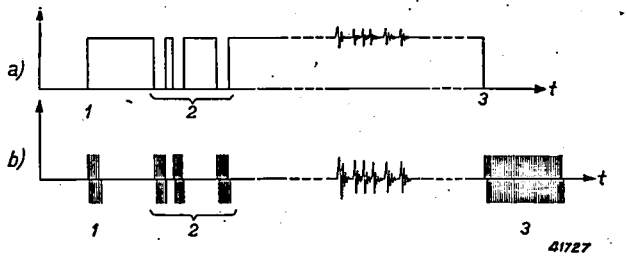


Fig. 4. Transition from continuous signalling with direct current (*a*) to impulse signalling with voice-frequency alternating current (*b*). *1* Calling signal, *2* dialling impulses, *3* clearing signal.

the whole call. The dialling impulses, which in continuous signalling consist of interruptions of the signal current, are passed on in impulse signalling by A.C. impulses, like the calling and clearing signals. The transition from continuous signalling to impulse signalling is illustrated in *fig. 4*. In practice, it is realized by a combination of a number of relays, a relay group.

The relay groups make the signalling more complicated and increase the chance of interferences. Besides this complication, however, there is also a fundamental difficulty in the method of voice-frequency signalling. Owing to the fact, that the signalling frequency is chosen in the frequency region of speech, it is possible that the signalling

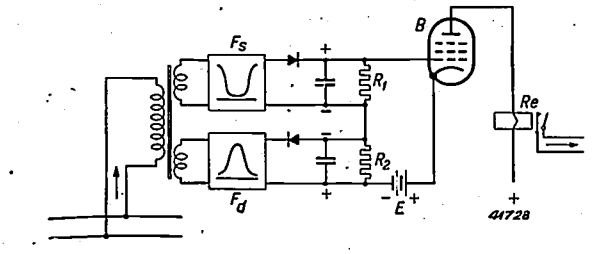


Fig. 5. Signalling receiver for voice-frequency signalling. The instrument has to distinguish, whether an input voltage with the signalling frequency originates from a signal, from speech or from interferences. For this purpose the input voltage is fed to two filters F_s and F_d connected in parallel. F_s passes the signalling frequency through and damps the other frequencies, F_d damping the signalling frequency and passing the others through. The output voltages of the two filters are rectified and fed with opposite polarity to the control grid of the amplifier valve B. Due to the negative bias from the battery E, the amplifier valve does not transmit any anode current, so that the relay Re is released. When a current with the signalling frequency arrives, a positive voltage occurs over R_1, the valve begins to carry current and Re reacts. If, however, other frequencies arrive, simultaneously with the signalling frequency, a negative voltage will occur on R_2, and if this is equal to or greater than the voltage on R_1 the valve B remains "overbiased" and Re does not operate.

[3]) A connection between two manual exchanges, as far as the signalling is concerned, is often simpler than one between two automatic exchanges, since there are no dialling impulses nor signals intended for the ear of the subscriber calling. The calling and clearing signals are still necessary, since it is impossible to have an operator listening on each line all the time. It is the dialling impulses, however, that cause the greatest technical difficulties, as we shall see.

receiver also reacts to the speech voltages themselves: if the signalling frequency occurs in any intensity in a speech sound the receiver may, for instance, construe this as a clearing signal and break the connection prematurely. In order to prevent this, the signalling receiver must be so constructed, that it can distinguish, whether a voltage with the signalling frequency originates from signal or from some other source. Use is often made here of the fact that in speech different frequencies always occur at the same time, whereas, in signalling only the signalling frequency occurs. The voltages with other frequencies are then used to put the signalling receiver out of action with the help of connections, which are shown in *fig. 5* and explained in the text underneath. This arrangement works satisfactorily, but it has the great drawback, that current surges or loud speech may cause the signalling receiver to be blocked, as a result of which impulses are lost or distorted.

Carrier wave signalling

While the method of voice-frequency signalling described, can be used for normal (low-frequency) as well as for carrier telephone connections, the method, about to be discussed, is especially intended for carrier telephony. With this method, which was developed by Philips several years ago, reaction to speech is out of the question.

In carrier telephony the speech vibrations of each call are fed to a modulator, together with a carrier wave of higher frequency. The modulator converts the speech vibrations into side bands of the carrier wave, but, thanks to suitable connections, does not allow the carrier wave itself to pass, or at least not to any extent worth mentioning [4]). The carrier wave itself is thus not transmitted over the line. In order to be able to bring the transmitted side band — only one is transmitted since the other is suppressed in the transmission band filter — back into the original frequency region, the carrier wave is added again in the demodulator at the receiving end.

In the new method of signalling, the carrier wave frequency of each telephone channel is used for signalling in that channel. Separate oscillators for the excitation of the carrier wave frequencies are not required, as the carrier waves are also necessary for the modulators (and demodulators) and are therefore already excited in the exchange [5]). By means of a transmission relay, the carrier wave can be applied in each channel to the line behind the modulator. At the receiving end a selective receiver is connected in parallel with the demodulator, which receiver amplifies and rectifies the carrier wave. A receiving relay is operated by the D.C. voltage obtained, and this relay passes the signals to the respective apparatus in the automatic exchange.

For this method, therefore, a frequency is used, which does not occur in the modulated speech spectrum, and which gives no audible tone after demodulation. In the first place, this makes it possible to use continuous signalling as is done in local telephony by means of direct current (the carrier wave itself may in fact also be considered as "displaced direct current" of the corresponding channel). As a result the above-mentioned relay groups are made much simpler. On the other hand, continuous signalling, as we shall see later, is also accompanied by drawbacks connected with the necessary limitation of the permissible signal intensity. If, for this reason, impulse signalling is after all preferred, there still remains the important advantage, that the reaction of the signalling receiver to speech need not be feared in principle, since the carrier wave frequency does not occur in the modulated speech. (The speech currents of the neighbouring channels, which, in principle, may contain the carrier frequency of the channel first considered, are already kept out of the channel by the filters in order to prevent cross talk, and so they cannot disturb the signalling receiver either.)

In the practical realization of carrier wave signalling, the following requirements must be kept in mind:

1) The signalling must not cause any disturbance of the speech (or the signals) of other channels.
2) The signals must be transmitted undistorted.

The second requirement is especially important in connection with the dialling impulses. These impulses are given, for instance, at a rate of 10 per second, each impulse lasting twice as long as the interval between two impulses (see *fig. 4*). In order to ensure reliable functioning of the selectors in the automatic exchange, only certain deviations from the nominal duration of the impulses are permissible. These tolerances, however, are for the greater part already consumed by the rest of the apparatus, for example by the variations in the turning of the dial of the subscriber's instrument. Thus, for trunk connection, it is essential that the length of the dialling impulses supplied at the beginning should be reproduced in the receiving

[4]) See: F. A. de Groot and P. J. den Haan, Modulators for carrier wave telephony, Philips Techn. Rev. 7, 83, 1942.
[5]) See: D. Goedhart and G. Hepp, Philips Techn. Rev. 8, 137, 1946.

station, with not more than several thousandths of a second deviation.

On the basis of the requirements mentioned, we shall now discuss the practical realization of carrier signalling.

lower channels after demodulation a tone of 1 500 c/sec, lies about 50 dB below the speech level. This of itself is not very disturbing, but, due to the fact, that several hundred of these harmonics fall within the speech band, there occurs in the aggregate an

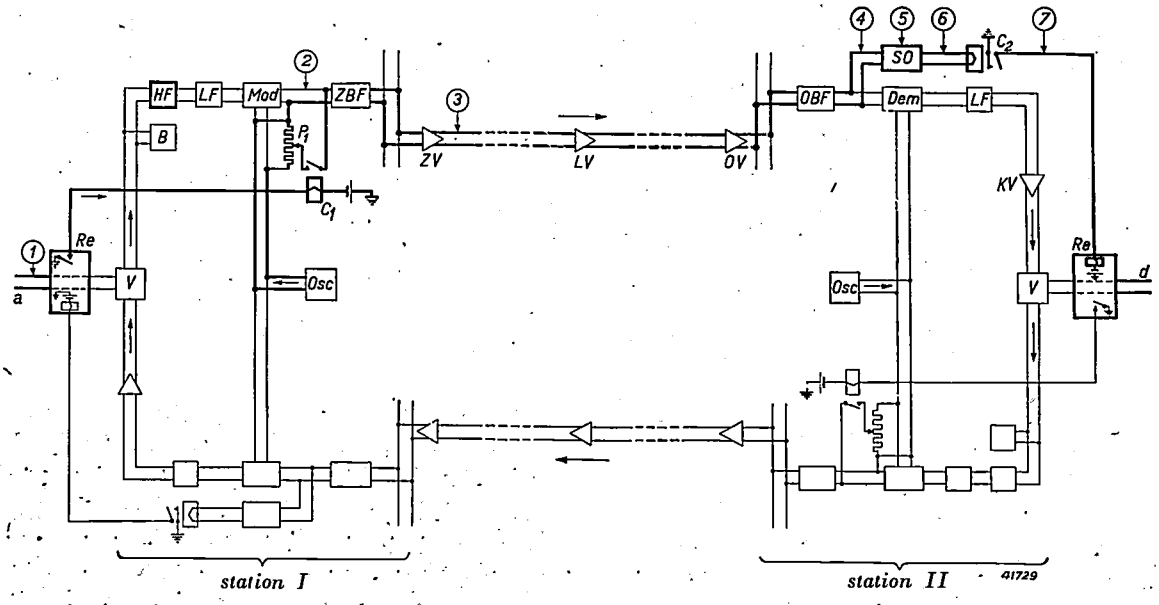


Fig. 6. Simplified diagram of one channel of a carrier telephone connection between two stations *I* and *II*. The components, added for signalling, and the path of the signals are indicated with heavy lines. At *a* (or *d*) the "local" line comes in. *Re* relay group. *V* four-wire termination, *B* limiter, *HF* high-pass filter, *LF* low-pass filter, *Mod* modulator, *Osc* generator for the carrier wave, C_1 transmitting relay, *P* potentiometer for regulating the intensity of the carrier signal, *ZBF* transmitting band filter, *ZV* transmitting amplifier, *LV* line repeaters, *OV* receiving amplifier, *OBF* receiving band filter, *SO* signal receiver, C_2 receiving relay, *Dem* demodulator, *KV* channel amplifier.

Practical execution of carrier signalling

Avoidance of interferences

Fig. 6 represents, in the form of a block diagram, the apparatus for one channel of a carrier telephone connection, the heavy lines indicating the path of the signals and the components added for the signalling.

As may be seen in the diagram, the carrier signals are not fed directly to the common line but (with the help of the relay C_1) to the input of the transmission band filter. This is necessary, in order to avoid interferences in other channels. The dialling impulses, which are sent through the relay, may be considered as a carrier wave, modulated with a sort of block-shaped curve with a fundamental frequency of 10 c/sec. The harmonics of this frequency, of which the block-shaped Fourier curve is built up, occur as side-band frequencies of the carrier wave and these side-band frequencies fall partially in neighbouring channels (*fig. 7*). A closer examination shows that, for instance the 250th harmonic (2 500 c/sec) which, with a carrier spacing of 4 000 c/sec between two adjacent channels, gives in the intolerably strong interference. All the side-band frequencies, which would fall in other channels, are now damped by the transmission band filter, as may be seen from the filter curve in fig. 7.

At the receiving end may be seen in fig. 6 the

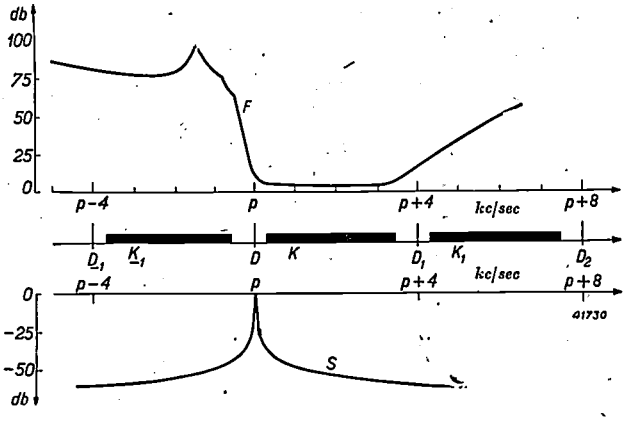


Fig. 7. Damping curve *F* of the transmission band filter of a channel, *D* carrier wave, *K* frequency region, that has to be transmitted for the speech in this channel; in addition the corresponding frequency regions and the carrier waves of several neighbouring channels are indicated. *S* amplitude of the harmonics of the dialling impulses for the channel in question *K*; the harmonics have a spacing of 10 c/sec. Each third harmonic has the intensity zero, but this fine structure cannot be seen in the drawing.

signalling receiver (*SO*), which upon reception of a signal activates the relay C_2. In order to prevent the receiver from reacting also to the speech voltages, it must contain a sharp filter for the signalling frequency. So as to avoid having to construct a receiver with a different filter for every channel, which is not very well feasible either for

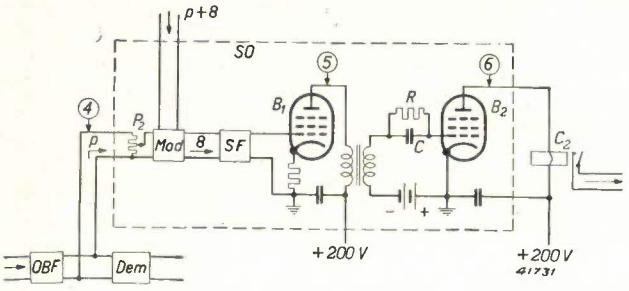


Fig. 8. Connections of the signalling receiver *SO*. P_2 potentiometer for regulating the intensity of the input signal, *Mod* modulator, *SF* filter for 8 kc/sec, B_1 amplifier valve, B_2 detector valve.

the designing, for the manufacture or the operation of the installation, the incoming signal (frequency p) is modulated in the receiver with an 8 kc/sec higher or lower frequency ($p \pm 8$), this giving rise, *i.a.*, to a difference frequency of 8 kc/sec. This is filtered out and fed to an amplifier and rectifier (*fig. 8*). In this way the same filter, namely for 8 kc/sec, can be used for all channels. This filter has the damping curve reproduced in *fig. 9*. It is seen, that already at a distance of 300 c/sec from the signalling frequency the damping amounts to more than 30 dB, while the speech band proper only begins here.

The reason why a frequency of 8 kc/sec is used is easy to understand. In the first place the frequencies $p \pm 8$, necessary for modulation, coincide with the carrier waves of other channels and are thus available in terminal stations of the installation without further auxiliary apparatus. In the second place it is an advantage to choose the difference frequency

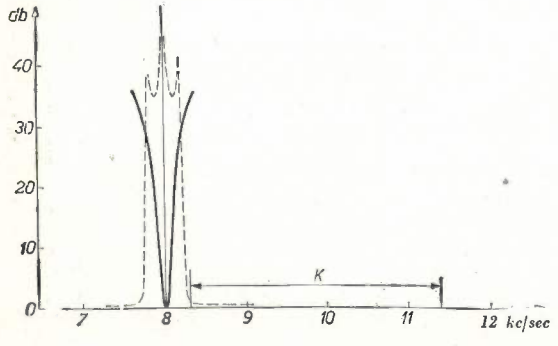


Fig. 9. Damping curve of the filter for 8 kc/sec in the signalling receiver. *K* is the frequency band of the corresponding speech channel. The broken line curve is the damping characteristic of the high-pass filter to be described later, shifted to 8 kc/sec.

low, as the filter curve can then be made relatively less narrow: an absolute width of the transmission region of 600 c/sec at a signalling frequency of 60 kc/sec means a relative width of 1 percent, whereas at 8 kc/sec it means a relative width of 7.4 percent, which is much easier to realize.

Since the carrier spacing normally amounts to 4 kc/sec, it might really have been better to use a difference frequency of 4 kc/sec, but this was not done for the following reason. The carrier frequency p, after modulation with $p + 8$, gives in the signalling receiver the difference frequency 8 kc/sec. But the carrier frequency $p + 16$ ("mirror frequency"), which belongs to the speech channel lying four channels higher and which transmits the signals for that channel, also gives in the receiver, first considered, the difference frequency 8 kc/sec, so that the receiver would also react to that. In order to prevent this, the signalling receiver is not connected directly with the line, but is placed behind the receiving band filter of its channel (see the diagram of fig. 6) or, more precisely, behind the first sections of the receiving band filter, in order not to expose the signals to the full damping, which that filter produces for the carrier frequency. The filter has, in the main, the same damping curve as the transmission band filter (fig. 7). The damping of the sections in question at $p + 8$ is not yet so great as to preclude entirely reaction of the receiving relay to a signal with this frequency, which danger threatens, when a difference frequency of 4 kc/sec is used. At $p + 16$, however, the damping is so great, that there is no danger of this.

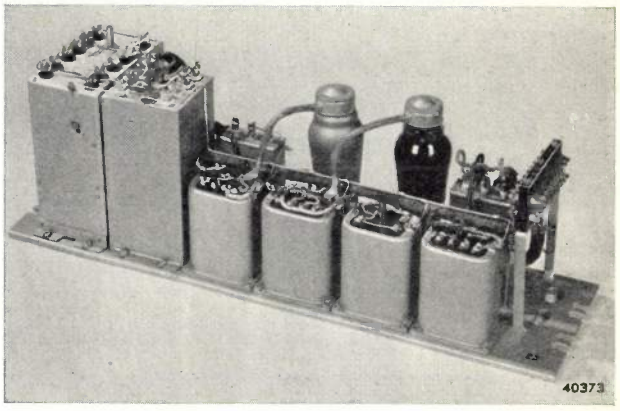


Fig. 10. Signalling receiver. On the left the modulator and the filter, on the right at the rear the two valves.

In *fig. 10* a photograph is shown of the signalling receiver described, while *fig. 11* shows a bay with signalling receivers and appurtenances, for a carrier installation for 17 channels.

Distortion of the signals

We have seen that the carrier signals must pass through the transmission band filter and the receiving band filter, among others. These filters suppress the side-band frequencies, which fall in their damping region, and also cause a relative attenuation of those frequencies, which fall in the transition region between damping and transmission, namely the carrier wave itself and the 30 lowest harmonics of the dialling impulses. This results in a certain (slight) distortion of the dialling impulses, as may be seen in the oscillograms of *fig. 12*. The first oscillogram shows the dialling impulses, as given with direct current (point *1* in fig. 6). The second oscillogram gives the voltage in front of the transmission band filter; during the time that the incoming line is not carrying current the carrier wave is transmitted (transition from the continuous to impulse signalling). These carrier impulses are practically rectangular. The third picture shows the impulses behind the transmission band filter, amplified in the transmission amplifier. Some rounding off has taken place. In the case of the input signal of the signalling receiver, oscillogram no. 4, the rounding off has become somewhat greater, owing to the influence of the receiving band filter.

The most important distortion, however, occurs in the following step: the filtering out of the 8 kc/sec frequency by the sharp filter in the signalling receiver. Here, since it is a question of eliminating the speech frequencies, the higher harmonics of the dialling impulses are attenuated. At 300 c/sec, *i.e.* the 30th harmonic, the damping of this filter, as already mentioned, is 30 dB; but also the 5th harmonic, for instance, is already attenuated by 6 dB. The resultant considerable rounding off of the impulse may be seen in the fifth oscillogram of fig. 12.

What is in fact of importance with dialling impulses is the need to keep to their correct length. A closer examination of the functioning of the signalling receiver (fig. 8) shows, to what extent errors may occur in this length, as a result of the rounding off of the impulses. After passing the filter the signals are amplified and fed to a valve (B_2) acting as anode detector. The I_a-V_g characteristic of the valve (anode direct current as a function of the grid A.C. voltage amplitude) is drawn in *fig. 13*. The fixed negative bias $-V_0$ of the grid is chosen so large, that plate current only begins to flow at several volts signal voltage. (This threshold voltage, below which the receiver does not work at all, gives an extra security against slight interferences, such as the carrier leak of the modulator at the transmitting end.)

With a sufficiently intense signal, plate current begins to flow; the relay C_2 will react as soon as the plate direct current has reached a value of 3 mA. The impulse transmitted thus becomes rectangular again in any case, but what about its length? In fig. 13 the variation of the grid A.C. voltage amplitude is drawn for the case, where the rounded off dialling impulses, reaching the grid, have the form of oscillogram no. 5, dialling impulses of three different intensities being assumed. The length of the original impulse is AB in all three cases. In the case of the weak sig-

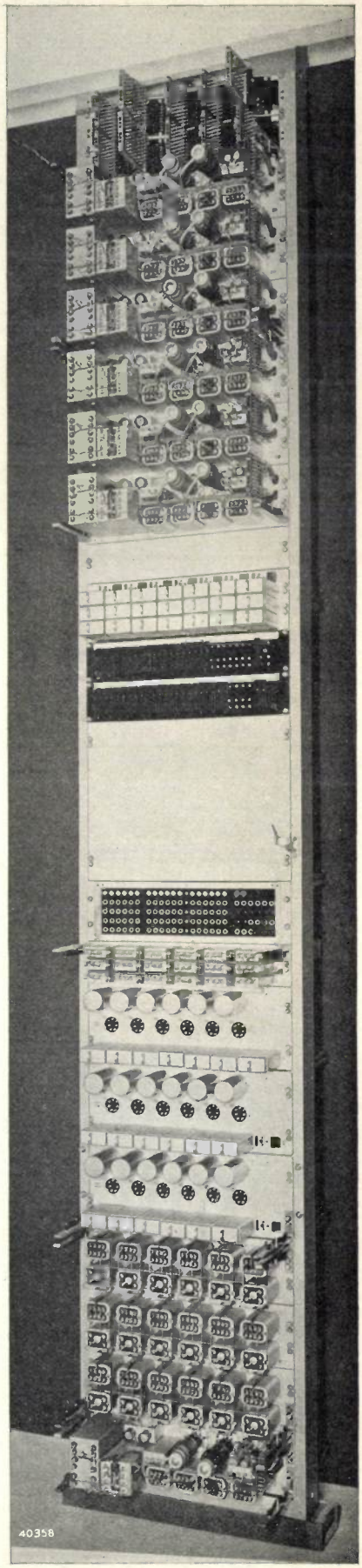


Fig. 11. Bay with signalling receivers and relays of a carrier installation for 17 channels (11 of the 18 receivers are mounted on the back of the bay; one receiver serves as reserve).

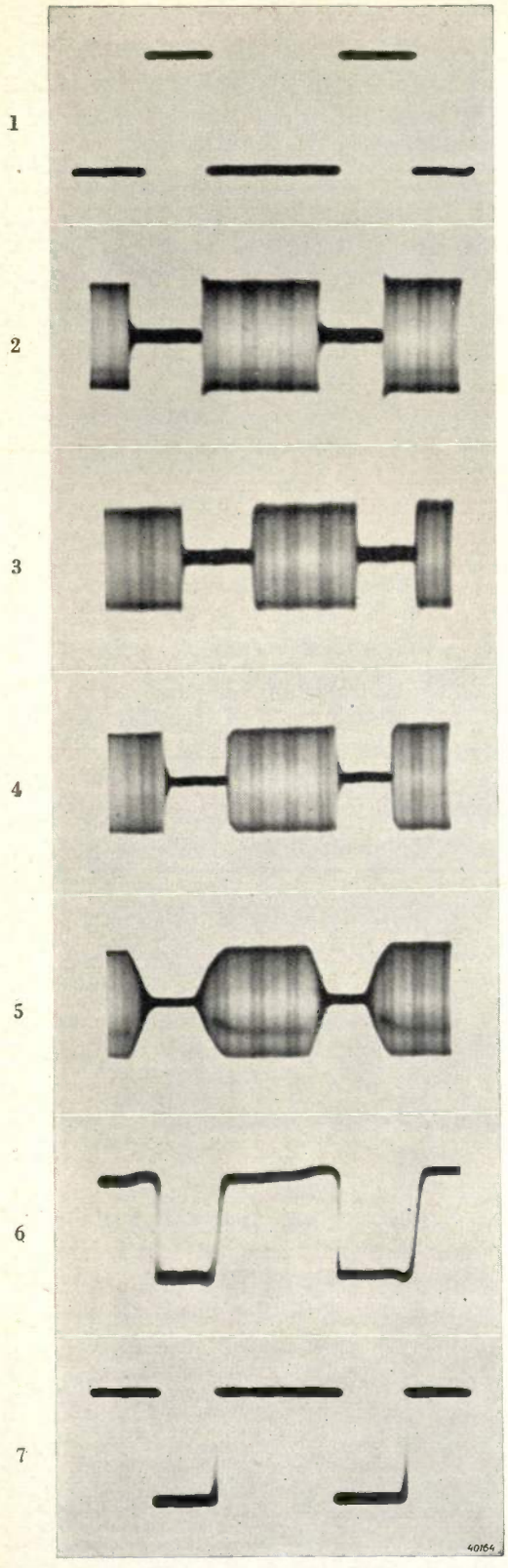


Fig. 12. Oscillograms of the dialling impulses recorded at points 1-7 of the diagram of fig. 6.

nal *1* the relay will react at *C* and release at *D*; the impulse, passed on by the relay C_2, is thus considerably shorter (length *CD*) than the original impulse. With the stronger signal *2* the relay operates from *E* to *F*, which is equal in length to the time *AB*. In the case of the very strong signal *3*, the relay operates from *G* to *H*, and the impulse passed on is thus longer than the original one.

If the signal intensity received were always the same, then, by a suitable choice of the bias V_0, it could be ensured, that the impulses, passed on, have the correct length. Actually, however, one must count on signals of different intensities. The cause of this lies for a large part in the dependence on temperature of the damping of the transmitting and receiving band filters. Temperature changes cause the values of the capacities and self-inductions in the filter to vary somewhat, so that the whole damping characteristic is shifted slightly in frequency. Since the carrier wave lies just in the region, where the damping curve mounts quite steeply, a slight shift in the damping curve quickly causes a considerable change in the damping for the carrier wave.

By means of a simple device, it is possible to neutralize almost completely this distortion caused by the variation in intensity. Provision is made that the weakest signal occurring has the shape of curve *2*, with which no grid currents yet occur. With larger signals, like that of curve *3*, grid current begins to flow. By now including in the grid circuit, as may be seen in fig. 8, a grid condenser *C* and a leakage resistance *R* of suitable value, the grid current is made to increase the negative bias, by such an amount (V_0' in fig. 13), that the operation and release of C_2 takes place at points *I* and *K* (instead of *G* and *H*), which with the signal in the form of curve *3* have exactly the desired time interval. At a different signal intensity the correction will not be quite exact, but a good approximation is obtained in the whole practically important range of variation of the signal intensity.

In the sixth oscillogram of fig. 12, the variation of the anode current of the detector valve is shown, in the seventh oscillogram, the impulses set up by the contact of the relay C_2. It is seen that here, apart from the inversion, which is cancelled in the relay group, a faithful image of the original impulses is obtained.

In addition to the distortion by the various filters discussed here, there is still another sort of distortion, requiring special measures. Let us return to the diagram of fig. 6. The relay C_1, that puts the carrier wave on the line, is controlled, *via* the relay group *Re*, by the D.C. impulses arriving at *a*. The direct current is now passed on to the modulator by the transformers situated in the four-wire termination *V*. The D.C. impulses, however, are built up not only from the direct current, but also from a series of A.C. components, which are allowed to pass and which are modulated on the carrier wave in the modulator, like the ordinary speech frequencies. This takes place practically without any retardation, calculated from *a* to the output of the modulator. The relays, however, which transmit the signals proper, work with a certain lag. The relay in the

[6]) See the article referred to in footnote [4]).

relay group *Re*, for instance, needs 0.01 sec to react, while C_1, an especially high-speed relay of special construction [7]), still needs 0.002 sec. Therefore, the A.C. components of the D.C. impulses reach the signalling receiver 0.012 sec before the

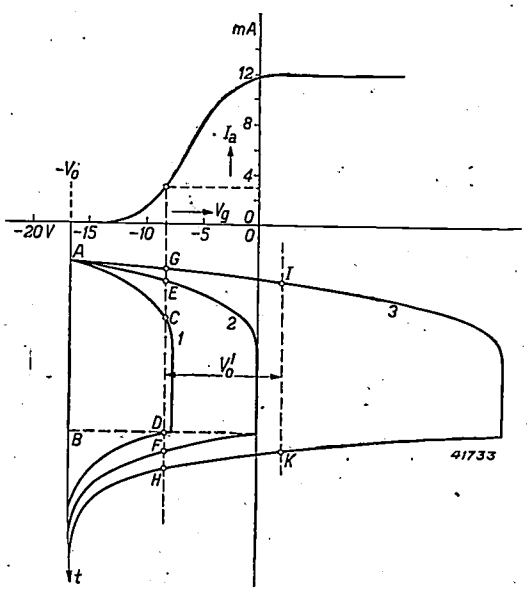


Fig. 13. I_a-V_g diagram of the detector valve in the signalling receiver. *1*, *2*, *3* behaviour with time of the grid A.C. voltage amplitude at three different intensities of the incoming dialling impulses.

signal proper; in as far as the frequencies of these components lie below about 300 c/sec, they can pass the filter in the signalling receiver without too great attenuation and cause relay C_2 to operate. As a result, a short surge precedes the impulse proper and the grid voltage of the detector valve may become so strongly negative, that the succeeding impulse is not received well.

The effect is counteracted, to some extent, by the limiter connected behind the four-wire termination, cutting off all peak voltages above a certain value (level for example 6 mW). This is not enough however. In order to dissipate the effect entirely, another filter is connected here in the speech channel, which causes a reasonable attenuation, namely at least 35 dB for frequencies below 300 c/sec (*i.e.* below the frequency band necessary for speech). The damping curve of this high-pass filter is shown as a broken line in fig. 9 for the modulated vibrations at the input of the 8 kc/sec filter in the signalling receiver.

The high-pass filter is also important for rendering the lowest frequencies in the speech currents absolutely harmless, and also to prevent any other signals, such as dialling tone or ringing tone, sent over the speech channel and also having components in the low-frequency region from acting upon the signalling receiver.

Finally a few words must be added about the intensity of the signals. The higher the intensity, the less the trouble experienced from all kinds of interferences, such as the carrier leak of the modulators. On the other hand, the signalling intensity may not be so high that the amplifiers common to all channels are overloaded. When only impulses are transmitted, which have a very short duration compared with the total duration of the call, this danger is only slight. Thus in the case of the system for carrier telephony with 17 channels, formerly developed by Philips, a level can be used for signalling, which corresponds to 1·5 mW, measured at the input of the four-wire termination. This corresponds, approximately, to the average level of the conversation itself and lies 30 dB, or more, above the possible carrier leak. If, however, it is desired to signal continuously and not with impulses, either the signal intensity must be chosen considerably lower or the amplifiers must have larger dimensions. It will depend upon the circumstances, whether the above-mentioned advantages of continuous signalling carry more weight than these disadvantages.

[7]) This is done also because of the distortion. Of itself an operating time and a releasing time of for instance 0.01 sec. is no objection, since it means only a slight retardation of all the impulses. If, however, the operating time and the releasing time of the relay C_1 are not exactly the same, the impulses become too long or too short. The unavoidable differences between the two times, taken absolutely, will be smaller, the shorter the two are taken.

THE "INFRAPHIL" AN APPARATUS FOR INFRA-RED THERAPY

by Th. J. J. A. MANDERS.

615.831.7:621.384

The therapeutic action of infra-red rays lies mainly in the temperature increase in the irradiated tissue. It is important, that the radiation should penetrate into the body as deeply as possible.
A study has therefore been made of the spectral distribution of the radiation necessary, in order that, with the given reflection and absorption of the skin, this requirement may be satisfied. An irradiator is then described, which emits rays with a spectral distribution, corresponding quite satisfactorily with the optimum distribution.

For several years already the therapeutic action of infra-red rays has been a subject of interest to the medical profession. At first, little was known about the way, in which the rays act, and very diverse sources of radiation were employed. On the one hand neon tubes were used, which emit the greater part of their radiation in the long-wave visible region; on the other hand electric heaters were used, whose radiation is mainly far in the infra-red. Research in recent years, however, has shown, that the curative effect of the radiation is, for the most part, to be accounted for by the heating of the tissue. From this it follows, that the wave length of the radiation is not of primary importance, the heating effect being entirely determined by the amount of radiant energy absorbed. But the spectral distribution of the radiation is by no means unimportant. From the investigations referred to, it has also been found, that the most favourable effect is obtained, when the radiation penetrates, as deeply as possible, into the tissues. A radiation must therefore be used, which is absorbed as little as possible by the skin.

Further it is obvious, that the rays emitted should be reflected as little as possible by the skin, thereby increasing the efficiency of the irradiation. By these requirements the spectral distribution of the radiation, to be used, is to a large extent fixed.

In this article we will describe the "Infraphil" irradiator, an apparatus of such simple construction, that under a doctor's directions, it can be used by laymen, for instance for alleviating rheumatic pains, for the treatmant of inflammations, etc. By a suitable construction of the radiator proper, it has been ensured, that a high percentage of the energy emitted, is converted into heat in the deeper-lying layers of the tissue.

The properties of the skin

In the foregoing it has been shown, that the properties of the skin almost completely determine the spectral distribution of the radiation. Those properties will therefore be examined in some detail.

Reflection

The reflectivity of the skin has been determined by various investigators, among whom are Schultze, Bode, Hardy and Muschenheim and Büttner. It has been found, that the reflectivity for infra-red rays, contrary to that for visible light, depends little on the intensity of the pigmentation of the skin. The reflection has a maximum between 0.6 and 0.7 μ and decreases rapidly towards the infra-red. In *fig. 1* the variation of the reflectivity is shown for a moderately pigmented skin, as found by the investigators mentioned.

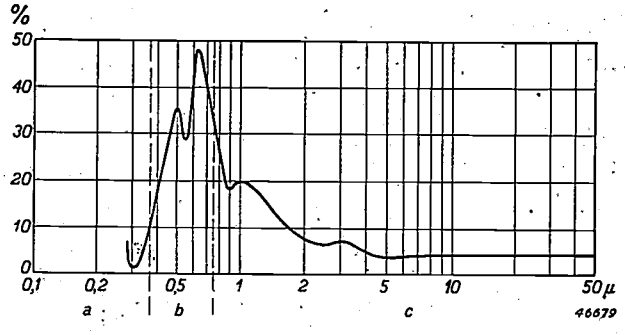


Fig. 1. Reflectivity of the skin as a function of the wave length in μ. *a* Ultra-violet region, *b* Visible region, *c* Infra-red region.

Translucency of the skin

As already noted, the red and especially the infra-red rays are relatively little reflected. The greater part of this radiation, therefore, penetrates into the skin. The depth of penetration will depend upon the scattering and absorption by the skin.

Little is known of the scattering power of the skin. Plotnikow found, that it is dependent on the wave length of the radiation and further on the thickness of the different skin layers. The investigators, mentioned below, have attempted to take this into account in the determination of the absorption.

The absorption of the skin, for different wave lengths, has been measured by various investigators, among whom Hardy and Muschenheim and Henschke. Their results are reproduced in *fig. 2*,

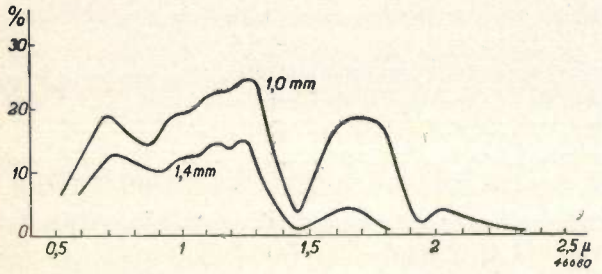


Fig. 2. Translucency of the skin as a function of the wave length in μ, for a skin thickness of 1.0 and 1.4 mm respectively.

from which it may be seen that, especially with a rather thick skin, practically only the wave length region from 0.6 to 1.4 μ is to be considered. The maximum translucency lies at about 1.2 μ. The shape of the translucency curve can be explained satisfactorily from the corresponding curves for the substances mainly responsible for the absorption in the skin, namely water and oxyhaemoglobin. Its shape above 1.2 μ especially is mainly determined by the water, while the sharp limit at 0.5 μ is to be ascribed to the action of the oxyhaemoglobin.

Furthermore Bachem has ascertained, that the region of greatest translucency for the more deeply lying parts of the skin is shifted towards long wave lengths. This phenomenon too is to be ascribed to the oxyhaemoglobin. The more deeply lying parts of the skin (*i.e.* the corium) are richer in blood than the epidermis and therefore contain relatively more oxyhaemoglobin.

For the rest no exaggerated ideas of the depth, to which the radiation penetrates, can be entertained, even for the most favourable wave lengths. From experiments by Henschke, Heald and Hofmann it has been found, that the intensity of radiation, in the region between 0.7 and 1.1 μ, is reduced to 0.1 percent of the original intensity, already at a depth of 2 cm. This is confirmed by measurements of the temperature distribution in the skin by Henschke. His results from experiments carried out with radiation in the region of 0.6—1.3 μ are reproduced in *fig. 3*, from which it may be seen, that considerable increases of temperature may occur at the surface of the skin, whereas at depths greater than 1 cm practically no increase can be detected.

From the foregoing it may be deduced, that a radiation in the wave-length region of 0.6 μ—1.4 μ is most suitable for irradiation. Radiation, in the region outside this interval, must be suppressed as much as possible, since it is mainly absorbed in the outer layers of the skin and heats it intensely, thereby limiting the amount of energy that can be irradiated on the skin per cm². This maximum energy per unit of area, the so-called loading capacity, is determined by the pain limit. If the intensity of the radiation is too high, a strong heating effect is first experienced, then an irritation, which finally becomes an intolerable pain. With all irradiations, therefore, there is a limit to the intensity that can be tolerated for a long time. This limit depends naturally on the spectral composition of the rays emitted. In choosing the dimensions of the irradiation apparatus, to be described farther on, it is therefore important, as

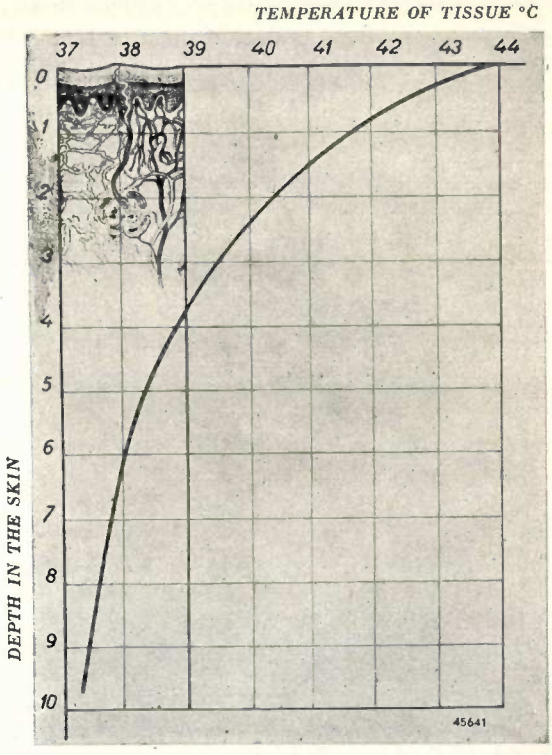


Fig. 3. Temperature of the tissue at different depths in the skin, when the surface of the skin is kept at 44 °C by irradiation (in the region between 0.6 and 1.3 μ).

Henschke found, that the limit in question should vary but little (about 5%) for different people. Van Wijk[1]) determined the loading capacity for heat irradiators of different temperatures mounted in glass bulbs (wall thickness about 1 mm). His results are shown in *fig. 4*. It is found, that the loading capacity increases considerably with rising temperature of the radiating body.

[1]) Philips techn. Rev. 6, 202, 1941.

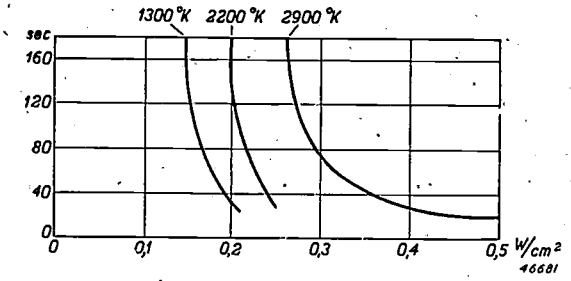


Fig. 4. Time, during which an irradiation can be tolerated, as a function of the intensity of irradiation for different temperatures of the incandescent body of the irradiator. The time is measured in seconds, the irradiation intensity in W/cm² and the temperature of the incandescent body in degrees K.

Physical properties of the apparatus

From the foregoing, it has become evident, that an apparatus for irradiation with infra-red rays must satisfy the requirement, that the largest possible part of the rays should lie in the wavelength region 0.6 μ—1.4 μ. Confining ourselves to heat irradiators, the maximum of the radiation curve should therefore lie in this interval, and since these curves fall off much more rapidly towards the short wave side than towards the other side, the maximum must lie quite close to 0.6 μ. With the help of Wien's displacement law, it may already be deduced that the temperature of the irradiator has to be very high. If we express the wave-length λ in μ and the temperature T in degrees K, Wien's law is as follows:

$$\lambda_{max} \cdot T = 2880.$$

For $\lambda_{max} = 6$ μ, it follows, that $T = 2880$ °K.

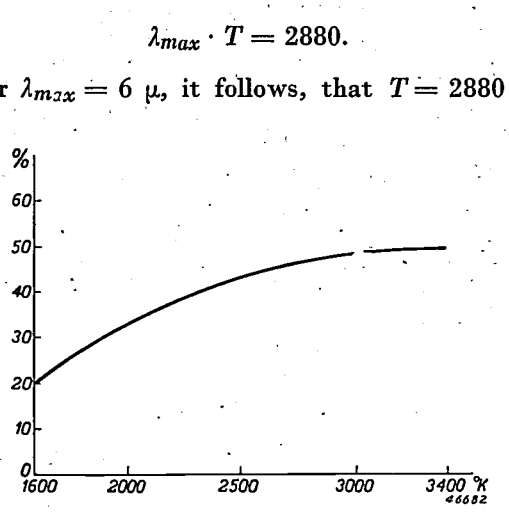


Fig. 5. The part of the total radiation of a tungsten coil, which falls in the region 0.7 μ—1.4, μ at different temperatures expressed in degrees K.

In *fig. 5* the part of the energy from a tungsten coil, which is emitted in the region in question, is indicated for different temperatures. It may be seen that, in good agreement with the above, the part, falling in the region from 0.6 μ to 1.4 μ, increases up to about 3400 °K.

In the case of the "Infraphil" apparatus, a tungsten coil is actually used, which, when in operation, consumes 150 W and reaches a temperature of 2800 °K. Although this is about the same temperature as is reached in an ordinary electric lamp of the same power, it is impossible to go much higher, because the tungsten would then evaporate at such a rate, that the lifetime of the irradiator would become far too short, to be of any practical use. At the temperature mentioned, therefore, the maximum yield of radiation in the region 0.6 μ to 1.4 μ is not obtained. Since, however, as is shown in figs. 1 and 2, the shortwave radiation in that region is more strongly absorbed by the skin, and especially more strongly reflected, a very satisfactory result is, nevertheless, ultimately obtained.

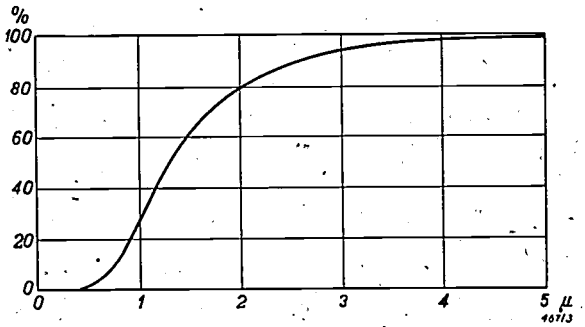


Fig. 6. Spectral distribution of the radiation of a tungsten coil at 2800 °K, represented in accumulated form.

In *fig. 6* the spectral distribution is given of the radiation of a tungsten coil at a temperature of 2800 °K (curve A). The diagram is given in an accumulated form, so that it is possible, to read off directly, what part of the radiation is emitted in the region from zero to a given wave length (and by subtraction, therefore, also the part in a region between any two wave lengths; with a non-accumulated distribution curve these percentages would have to be determined, for instance, by planimetry).

As tungsten readily oxidizes, especially at high temperatures, the incandescent body has to be fused into a glass bulb. The socket end of this bulb has the shape of a paraboloid and is mirrored on the inside, thus concentrating the energy. The greatest diameter of the bulb is 125 mm (see *fig. 7*).

The internally deposited aluminium mirror offers many advantages over an externally applied reflector. In the first place the rays have to pass only once through the wall of the bulb, whereas with an external reflector at least the central portion of the beam has to pass through the glass three times. The absorption is therefore less. Further, the internal mirror is less exposed to damage and is protec-

Fig. 7. The radiator of the "Infraphil" apparatus, showing the frosted front and the mirrored back of the bulb.

ted against contamination; finally it occupies no extra space.

The mirror is applied by evaporation of aluminium on the inside of the bulb. In *fig. 8* the reflection of aluminium, deposited by evaporation, is given as a

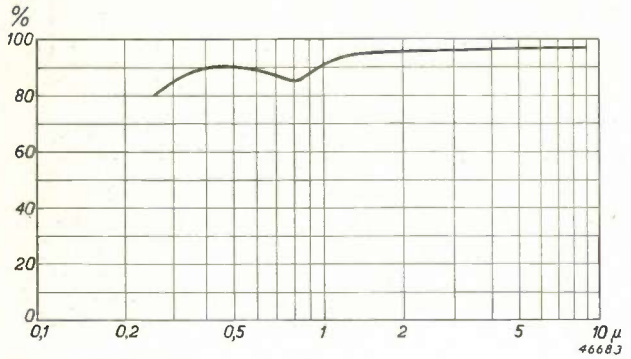


Fig. 8. Reflectivity of aluminium deposited by evaporation for different wave lengths (in μ).

function of the wavelength. In the whole region in question, aluminium is found to be a good reflector, so that the mirror causes practically no changes in the spectral composition of the ray beam emitted.

A second factor of influence on the radiation emitted through the bulb is the translucency of the glass. This is represented graphically in *fig. 9*. The transmission is very good in the region that counts. Only at 2.5 μ does absorption clearly begin to occur. This, however, is an advantage, because part of the undesired radiation with a wave length greater than 1.4 μ is thereby removed.

After what has been stated in the foregoing, it will not be surprizing that in the region with which we are concerned, the spectral distribution of the beam leaving the bulb differs little from that of the radiation from the filament, shown in fig. 7. Notwithstanding the above-mentioned absorption effect, a not inconsiderable part of the radiation emitted has a wave length greater than 1.4 μ. By applying a water filter (see the article by van Wijk already referred to) this radiation could practically be eliminated. It would thereby be possible to increase the loading capacity from 0.27 W/cm² (see fig. 4) to about 0.45 W/cm². For an apparatus mainly intended to be used by laymen, however, this would be too complicated and also make it too expensive.

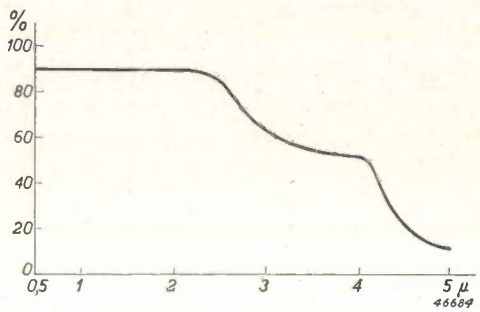


Fig. 9. Translucency of calcium glass for different wave lengths (in μ). Thickness of glass 1 mm.

The front of the bulb is frosted, in order to render the intensity at the centre of the beam equal at all points. If this were not done, local irradiation maxima would occur on the part of the skin irradiated. This has an unfavourable effect, because the total radiation, that can be tolerated, is finally determined by the loading capacity of the most intensely irradiated area of the skin.

Finally in *fig. 10*, the irradiation intensity is shown on a plane perpendicular to the axis of the irradiator, at different distances from the axis.

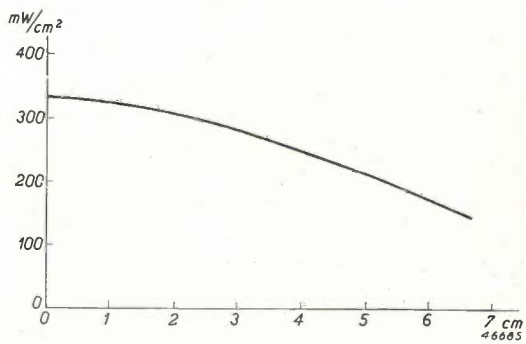


Fig. 10. Intensity of the irradiation as a function of the distance from the centre of the beam, measured in a flat plane perpendicular to the axis at 25 cm distance from the front of the irradiator.

Details of the apparatus

In *fig. 11* the complete "Infraphil" apparatus is

Fig. 11. The complete "Infraphil" apparatus. The bracket, in which the irradiator is hung, can be translated and rotated with respect to the base. The irradiator can therefore be placed in any desired position.

shown. The irradiator, already discussed, is contained in a metal housing held in a bracket.

The bracket is made movable by means of a grooved clamp, so that the beam is adjustable within the plane of the bracket. Moreover, the clamp can be rotated about the axis of the base, so that when the apparatus is hung on the wall — an eyelet is provided for this purpose underneath the base — the beam can be adjusted in every desired direction. For use by a medical practitioner, the apparatus is supplied with a stand, providings till wider possibilities of adjustment (*fig. 12*). Here the radiator with bracket is attached to a tube, instead of to the base. This tube passes through a clamp, in which it can be rotated, and also moved in a horizontal direction. The clamp itself can, moreover, be moved along a vertical tube and rotated about it. A switch is mounted between bracket and tube.

In certain cases it is desirable to be able to administer very local irradiation. For this purpose,

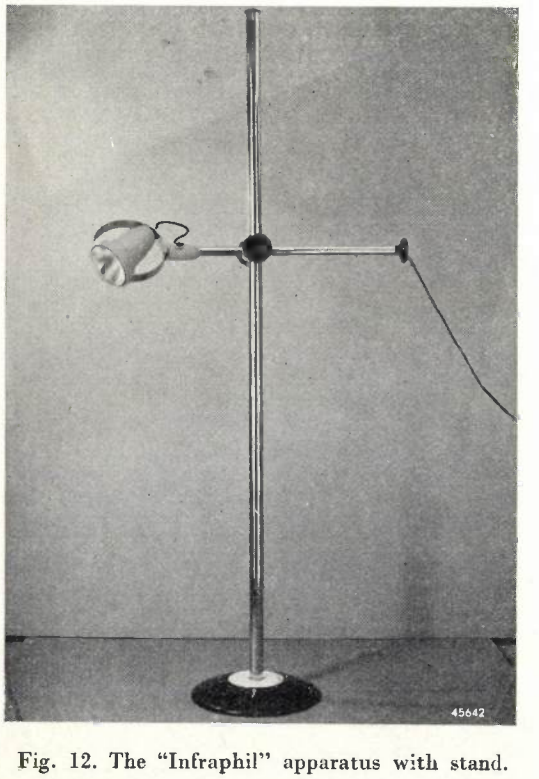


Fig. 12. The "Infraphil" apparatus with stand.

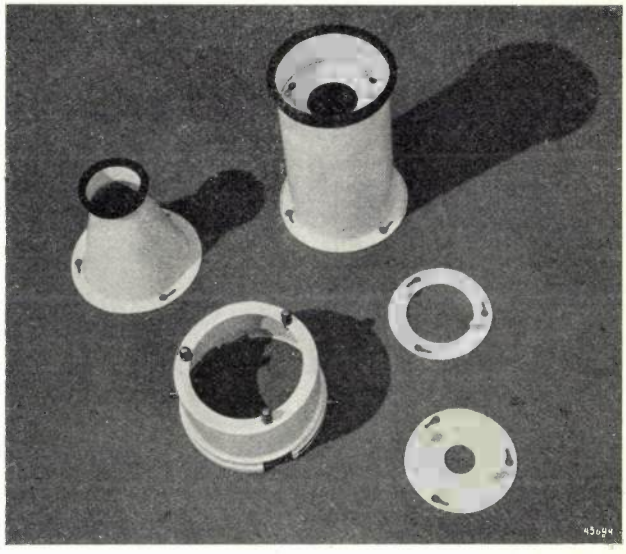


Fig. 13. Various forms of localizers, designed for very local irradiation. On the left in the foreground, is the ring, with which the localizers are attached to the housing.

a number of localizers have been designed (*fig. 13*), which can be attached to the apparatus very simply, with the help of a ring. The apparatus with a localizer is shown in *fig. 14*.

Applying the apparatus

It can be taken that infra-red rays have no injurious effect upon the eyes, during the short time, that irradiation is usually applied. In fact, the visible radiation is so intense, that one automatically shuts the eyes, if the beam falls upon them. Nevertheless, in the latter case it is desirable to protect the eyes, as much as possible. Care must be taken, when using goggles, because they may absorb so much heat that, they may burn the skin,

where the frame touches it. Readily inflammable objects, such as combs, etc., must be kept out of the field of radiation.

In general, the intensity of irradiation should be as high as possible. The distance from the skin must therefore be as short as possible, but not so short as to be painful. The heat should be comfortably bearable. From this, it follows, that care must be taken, when irradiating places which for some reason or other have become insensitive, for then there is no feeling of pain to give warning, if the beam becomes too intensive, and burns may result.

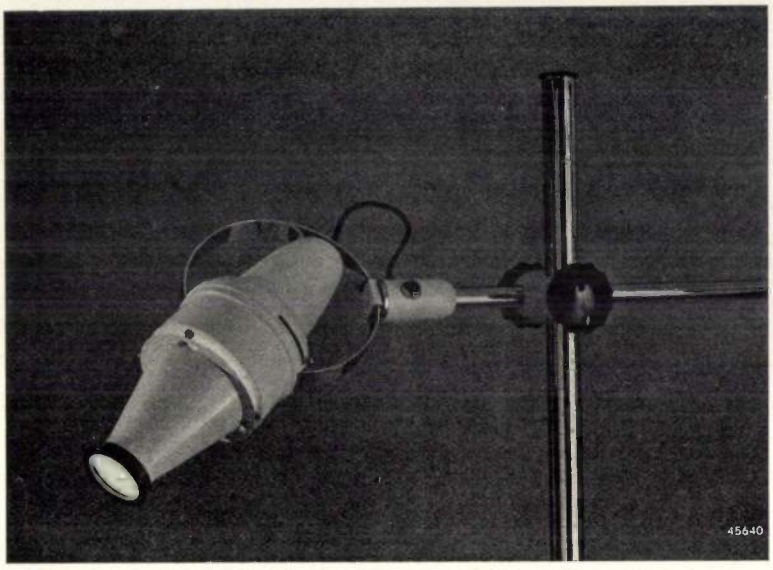


Fig. 14. The "Infraphil" apparatus on a stand, with a localizer in place.

ELIMINATING SCATTERED RADIATION IN MEDICAL X-RAY PHOTOGRAPHS

by W. J. OOSTERKAMP.

778.33:535.361.2

The scattered radiation, occurring in the diagnostical X-ray examination of the human body, produces a uniform secondary exposure of the X-ray film (or secondary illumination of the fluoroscope), thereby considerably reducing the sharpness of the contrasts. Various methods have therefore been developed, to eliminate scattered radiation or to reduce its effect on the film or screen. The criterion for the value of each method is the "selectivity", *i.e.* the ratio between the attenuation factor for the scattered rays and that for the primary radiation, producing the desired X-ray shadow picture. The present article first gives a survey of the usual methods of eliminating scattered rays and then proceeds to deal in more detail with one of the most important of them, the so-called Potter-Bucky diaphragm. Various forms of such diaphragms are compared as to selectivity and other properties, from which indications are deduced for their most efficient construction. In most of the methods of counteracting scattered radiation, there is some disadvantage or other, that has to be taken into the bargain, such as a distortion, further lack of definition, etc. The doctor must therefore decide for himself, which method or combination of methods is most suitable for a certain examination.

Introduction

The familiar shadow image of the internal organs and skeletal parts, obtained in X-ray photography (or fluoroscopy) of the human body is brought about by the difference in degree of attenuation of the X-rays, while passing through the various parts of the body. According as the intensity of the primary radiation, propagated rectilinearly from the focus of the X-ray tube, varies from place to place, so certain contrasts are formed in the blackening of the X-ray film.

Actually, however, the contrasts are always less defined, than would be expected from the above. The reason for this lies in the scattering of X-rays. The attenuation of the primary X-rays in matter is caused partly by absorption and partly by scattering, *i.e.* a change in the direction of the rays (accompanied by a certain loss of energy, which, however, with the tube voltage used in X-ray diagnostics is only slight). Owing to the fact, that the scattered rays from every particle of the body exposed to the primary radiation travel in relatively random directions, they cause a more or less diffuse exposure of the film, which leads to fogging and, consequently, diminishes the contrasts.

It will be seen from the following, that this effect is very important in X-ray photographing of the human body. In the case of elements with a high atomic weight, the attenuation of the X-rays is mainly due to absorption, for example in the bones, which consist for a large part of calcareous matter (with the relatively heavy element calcium). The other human tissues, however, are composed almost entirely of elements with a low atomic weight (hydrogen, carbon, nitrogen, oxygen), and in them, with the usual tube voltages employed, the attenuation due to scattering is of the same order of magnitude, as that due to absorption. Since in diagnostics, it is in many cases a question of distinguishing very slight contrasts in the X-ray picture, various methods have been developed for neutralizing or eliminating the scattered rays. These methods will be described in this article and as far as possible considered quantitatively. The striking improvement in contrast, sometimes obtained by counteracting scattered rays, is demonstrated by *fig. 1*, in which two X-ray photographs of the pelvic region are reproduced. The first was taken without any special measures for the elimination of scattered rays, whereas, in the second a large part of the scattered radiation was intercepted by means of a so-called Potter-Bucky diaphragm.

We shall begin with a somewhat more detailed consideration of the influence of scattered rays on the formation of the X-ray picture.

Influence of scattered rays on the X-ray picture

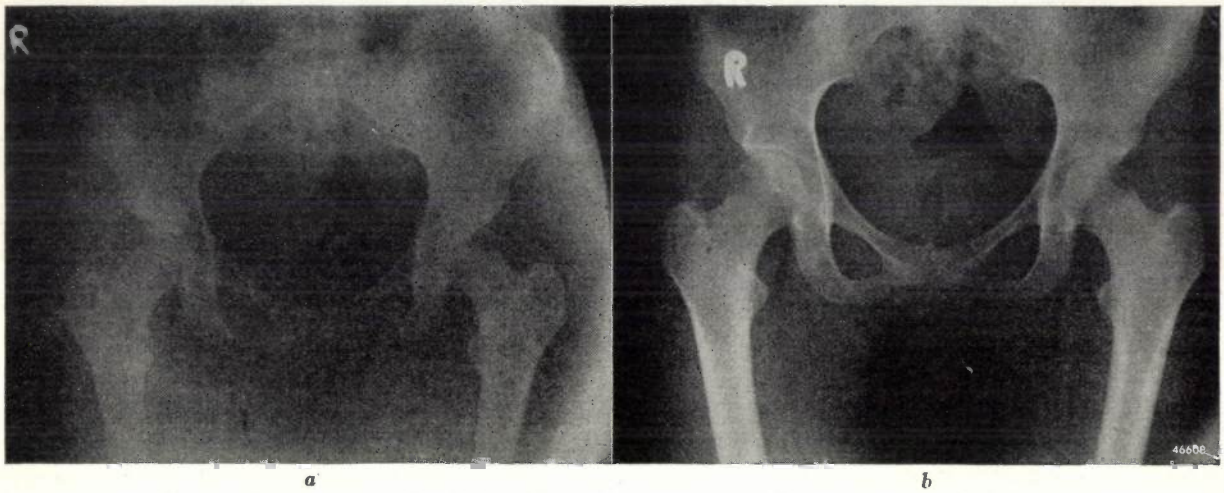
Radiation contrast

We assume, for the sake of simplicity, that the object irradiated is of a homogeneous composition with the exception of a small part M, in which the attenuation of the primary rays is slightly greater, than in the remaining part (see *fig. 2*).

This part M then appears on the negative as a spot M', which is slightly lighter than the surroundings. However, M' is struck not only by the direct rays but also by scattered rays, coming from

all the points of the object irradiated. The intensity of these scattered rays is practically the same, all over the film.

What effect does this have on the contrast? To be precise, we must here distinguish between radiation contrast, *i.e.* the relative difference in radiation intensity, and the resultant photographic contrast, *i.e.* the relative difference in blackening of the film. As to the first, we assume, that in the example mentioned, the primary radiation has



a *b*

Fig. 1. X-ray photograph of the pelvic region, *a*) taken without measures to eliminate scattered radiation, *b*) taken with a diaphragm to reduce the scattered radiation (Potter-Bucky diaphragm).

the intensity I_p all over the film, except for the spot M', where the intensity is $I_p - \Delta I_p$. The radiation contrast K_p between M' and the rest of the

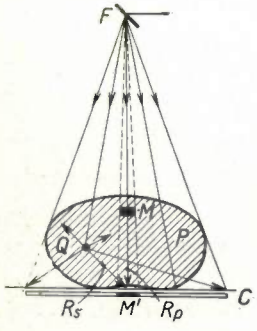


Fig. 2. In the body P there is an inhomogeniety M. This causes a somewhat greater attenuation of the primary X-rays R_p, coming from the focus F, and thus throws a shadow M' on the film C. Scattered rays R_s, which originate at every point Q of the body P, also strike the film at M'.

film, which would be obtained, if there were no scattered radiation, then amounts to:

$$K_p = \frac{I_p - (I_p - \Delta I_p)}{I_p} = \frac{\Delta I_p}{I_p}.$$

Actually, there is a scattered radiation all over the film, the intensity of which we shall call I_s. As a consequence, the total intensity on M' is $I_p - \Delta I_p + I_s$, and on the rest of the film $I_p + I_s$. Thus, under the influence of the scattered rays, the radiation contrast is changed and becomes:

$$K_s = \frac{(I_p + I_s) - (I_p - \Delta I_p + I_s)}{I_p + I_s} = \frac{\Delta I_p}{I_p + I_s}.$$

It is thus clear, that the radiation contrast is diminished by the scattered radiation, the decrease being in the ratio of:

$$\frac{K_p}{K_s} = \frac{I_p + I_s}{I_p}.$$

For the sake of orientation, it may be stated that, in a photograph of the lung, the intensity of the scattered radiation is approximately equal to that of the primary radiation, while, in photographs of the abdominal organs, I_s may be three or four times I_p, or even more. In lung photographs therefore $K_p/K_s = 2$, thus the radiation contrast is reduced to one half by the scattered radiation, and in abdomen photographs $K_p/K_s = 4$ or 5, the contrast thus being reduced to 1/4 or 1/5 of the value, which would be obtained without scattered radiation.

Photographic contrast

With the total radiation $I_p + I_s$ acting upon the photographic film, one could imagine the situation being such, that the film is first exposed, for the prescribed time, only with the direct radiation I_p, which forms the picture, and then re-exposed once more for the same time, without the patient, to a uniform radiation I_s/I_p times as intense. The disastrous effect of such a uniform after-exposure is obvious. Upon closer examination, however, it is seen, that in reality the situation is more compli-

cated. The reduction in contrast on the film, due to the scattered rays, will only be of the same order as that of the radiation contrast, when one is working on the linear part of the blackening curve, both with and without scattered rays. This condition will certainly not be satisfied, if by applying some means or other the scattered radiation were entirely eliminated and a photograph were then made with the same exposure time, as was first used with scattered radiation present.

Especially, in the case of moving objects (heart, lungs, etc), having regard to the lack of definition of the movement, the exposure time is chosen as short as is compatible with the requirement of sufficient blackening density, and if that density is only just sufficient with scattered radiation, it will certainly be insufficient without it. Since this means, that one works in a less favourable part of the blackening density curve and consequently with a given radiation contrast, one obtains a smaller photographic contrast, it is to be concluded that in an ordinary X-ray photograph the uniformly distributed exposure by the scattered rays does indeed spoil the radiation contrast, but in the photographic contrast partially makes up for that effect, by shifting all the amounts of radiation to a more favourable part of the density curve.

This need not, however, prevent us from eliminating the scattered rays, because, in order to reach the most favourable part of the density curve, there are other means available, which have no such prohibitive drawbacks. In principle, the essence of these methods lies in an increasing of the quantity of the primary radiation itself, so as to obtain the original blackening.

Increasing the quantity of primary radiation

The quantity of primary radiation can be most simply increased by lengthening the exposure time. If the scattered radiation I_s is entirely eliminated and the intensity I_p of the primary radiation remains unchanged, the total radiation is decreased by a factor $I_p/(I_p + I_s)$. In order, therefore, to obtain the same blackening density, the exposure time must be increased by the factor [1]).

In an abdomen photograph, therefore, the exposure time would have to be four to five times as long.

Actually, in almost all methods of eliminating scattered radiation, on the one hand also I_p is decreased and on the other I_s is not reduced exactly to zero. If we denote the intensities, to which the primary and the scattered radiations are reduced, by I_p' and I_s' respectively, the required factor of increase of exposure time becomes:

$$B = \frac{I_p' + I_s'}{I_p + I_s}.$$

When so-called Potter-Bucky diaphragms are used to eliminate scattered radiation, which we shall presently discuss in detail, B is called the "Bucky factor".

We have already alluded to the fact that, increasing the exposure time in the case of moving objects increases the lack of definition, due to the movement. It is possible to retain the original exposure time and instead: 1) increase the intensity of the primary radiation by expanding the area of the focus of the X-ray tube, or 2) use more sensitive intensifying screens, at the front and back of the X-ray film, in order to convert the primary X-radiation into a form of radiation more favourable for the blackening. By the first method, the geometrical blurring (width of the half-shadow caused by the finite focus width) is increased, whilst the second is accompanied by a greater blurring due to the screens, since more sensitive screens give in general a poorer definition [2]). Thus in any case part of the definition must be sacrificed. The best course to follow is determined by the so-called uniformity condition, *i.e.* the three components of the blurring (that caused by the moving of the object, that due to the screen and the geometrical blurring) should be approximately equal.

There is also another possibility of compensating for the loss of intensity, caused by eliminating the scattered radiation, namely by increasing the voltage on the X-ray tube, thereby obtaining a more intensive but, also particularly, a harder radiation, which is less attenuated in the patient's body, so that much more radiation reaches the film. Although it is true that, in consequence of this reduced attenuation of the primary rays, a somewhat reduced radiation contrast is obtained, usually the gain in contrast for the total radiation, resulting from the removal of the scattered rays, more than compensates for that.

Review of the methods of eliminating scattered radiation

When discussing the different methods of counter-

[1]) If we assume, for the sake of simplicity, that the blackening is determined entirely by the quantity of radiation. This is a slight deviation from the actual fact. (Schwarzschild effect.)

[2]) For an explanation of this see for example Philips Techn. Rev. 5, 266, 1940. It is in any case assumed, that the highest possible value permissible in connection with the focus temperature is chosen for the tube voltage, which is one of the factors determining the X-ray intensity.

acting scattered radiation and comparing them with each other, we must have a measure, by which to express the value of a method quantitatively. The attenuation factor for the scattered rays, which we shall call S ($S = I_s'/I_s$), is not suitable for the purpose, because if S is small, but at the same time the primary radiation is also very much attenuated (attenuation factor $P = I_p'/I_p$), the method has little or no value. The requirement is rather that the scattered rays should be much more strongly attenuated than the primary radiation, in other words, that the method should have a selective attenuating effect on the scattered rays. The value of a method can therefore be judged according to the ratio $P/S = \Sigma$, the so-called selectivity, which was first introduced by de Waard [3]). The larger Σ, the more closely the ultimate object is approached. At the same time, however, there is the condition, that P itself may not be very much smaller than unity, since a large loss of direct radiation necessitates a further increase of the exposure time (in addition to the already mentioned increase necessary as a result of the elimination of the scattered radiation).

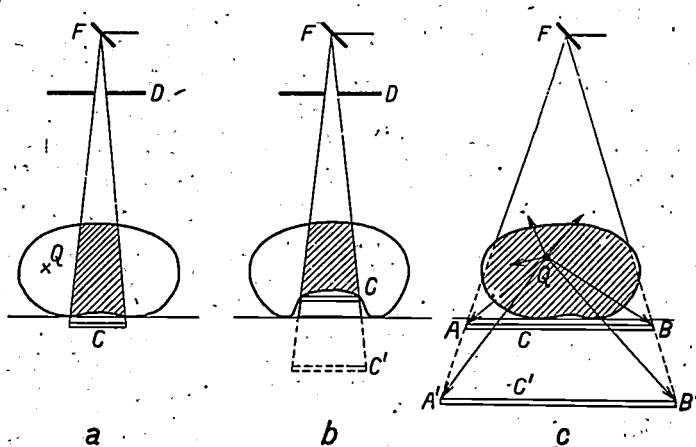


Fig. 3. Three simple methods of limiting the scattered rays.
a) Reducing the size of the field by means of a diaphragm D. There are then many points Q in the body, which contribute less to the scattered radiation.
b) Pushing aside as much body tissue as possible by compression. The method can best be combined with that of *c*).
c) Increasing the distance between patient and film. In the position C of the film it is struck by all the scattered radiation from Q within the solid angle AQB; in position C' only by the scattered radiation within the angle $A'QB'$.

Apart from the selectivity, there are also other factors, which are not so easily expressed in numbers and which are likewise of importance in estimating the usefulness of a method for counteracting the scattered radiation. These factors will be dealt with, as they arise in the discussion of the different methods.

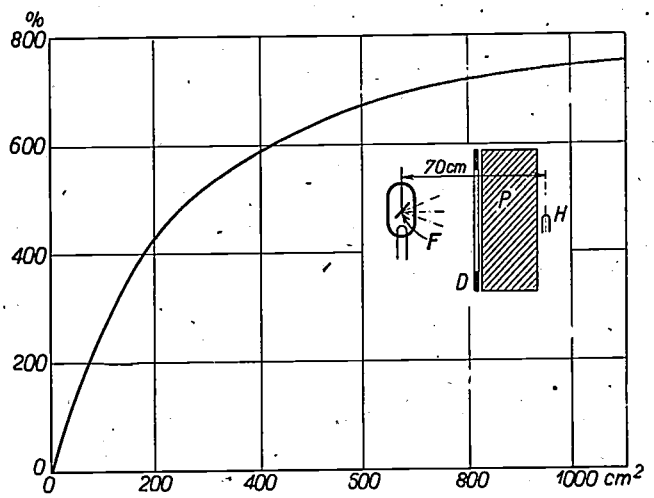


Fig. 4. Influence of the size of the field on the scattered radiation. With the help of a phantom P, consisting of "Philite", with a total thickness of 16 cm and intermediate layers of air totally 5 cm in thickness, a stomach examination was imitated, while the intensity of the scattered radiation was determined with a Hammer thimble dosemeter H for different openings of the square diaphragm D, at 100 kV tube voltage (DC voltage with a ripple of 30 kV) and 30 mA tube current. The scattered radiation, in percent of the direct radiation, is plotted, as a function of the area of the diaphragm in cm².

1) *Limitation of the size of the field*

The amount of scattered radiation increases with the size of the mass of tissue, exposed to the primary beam, since each particle of this mass plays its part. Sometimes, a very considerable diminution of the scattered radiation can be obtained, by chosing the primary beam of rays no wider, *i.e.* the field irradiated no larger, than is absolutely necessary for the examination (*fig. 3a*). *Fig. 4* shows, how the scattered radiation depends upon the size of the field in the case of a stomach examination (imitated for the experiments by the irradiation of a so-called "phantom" made of "Philite"). The reduction of the field has no effect on the intensity of the primary radiation, thus $P = 1$. The selectivity Σ is therefore in this case equal to $1/S$, the reciprocal of the attenuation factor for the scattered rays. *Fig. 5* indicates the selectivity of the field-reduction method, relative to a field of 35 × 35 cm, where Σ is set equal to 1.

2) *Compression*

The mass of tissue, through which the primary radiation has to pass, can sometimes be further reduced, by local compression of the patient's body, part of the tissue being pushed to one side (*fig. 3b*). This method is particularly important in the examination of the abdomen. It only helps of course when, according to method 1), the field

[3]) R. H. de Waard, Fortschr. Röntgenstr. **49**, 415, 1934 and **50**, 606, 1934.

has been limited to the compressed part and when actual parts of the tissue are pushed aside out of the field irradiated; pressing the air out of organs filled with air has no effect, since it is not the volume, but

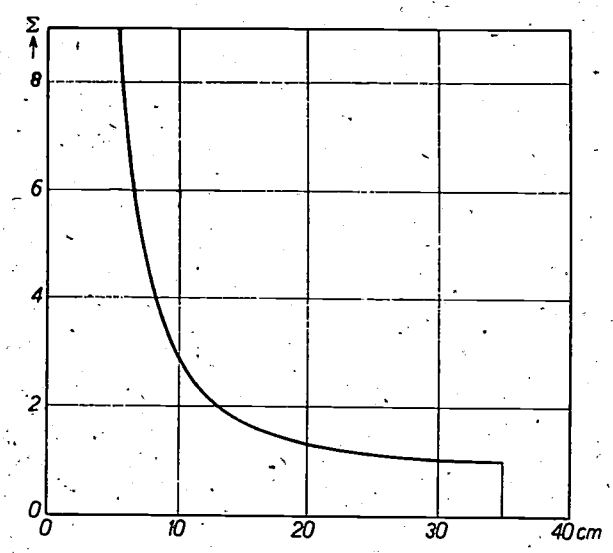


Fig. 5. The selectivity Σ (ratio between the attenuation factor for the direct radiation and that for the scattered radiation) of the method of reduction of the field. The reference field was 35 cm square. Same set-up as indicated in fig. 4.

the mass radiated, which determines the amount of scattered radiation. The advantage of this method is only fully realized, when it is combined with the following method.

3) *Increasing the distance between film and object*

If the film is placed not directly behind the object, but at some distance from it, it is withdrawn from the influence of a part of the scattered radiation, as may be seen from *fig. 3c*[4]). If the film is at C, all the scattered radiation from the point Q, within the angle contained between QA and QB, reaches the film; with the film at C' the angle, now between QA' and QB', is much smaller. The selectivity of this method is better, the smaller the field chosen. This may be seen from *fig. 6*, where the selectivity, measured in experiments with the same phantom as used for fig. 4, is plotted as a function of the distance between film and object, for a large and for a small field.

A great objection to this method is, that an increase of the distance between film and object results in a rapid increase of the geometrical blurring. If, for example, the patient's body is 20 cm thick, then with a distance 10 cm between film and patient the geometrical blurring u_g of a point in the centre of the patient's body is twice as great, as that with immediate contact between film and patient. Since the other two components of blurring, mentioned in the foregoing, remain the same, when we disregard for the present the increase in exposure time necessary, due to the decrease in the total intensity of radiation, the increase in the total blurring is smaller; in the example given, it amounts to about 30%. In *fig. 7* the geometrical and the total blurring are plotted as functions of the distance between film and object, for a normal stomach examination, with a focus width of 2.0 mm (full-line curves). The broken lines indicate analogous curves for the smaller focus width 1.2 mm and the necessarily longer exposure time. It may be seen, that with large distances between film and object the latter curves become more favourable, because the geometrical blurring, which then constitutes the major part of the total blurring, is less.

It may therefore be concluded, that the method discussed here is only suitable, when an X-ray tube with a small focus but with a high permissible load is used.

4) *Filters*

A certain attenuation of the scattered radiation can be obtained, by placing between patient and film a plane foil of some kind of highly absorbent metal, for instance copper or tin. The scattered rays pass through the foil for the most part obliquely, and thus cover longer distances in it, than the primary radiation. Moreover, the scattered radiation is somewhat softer, but with the normal tube voltages, used in diagnostics, this is not of much importance. In order to obtain any effect, the filter must be made

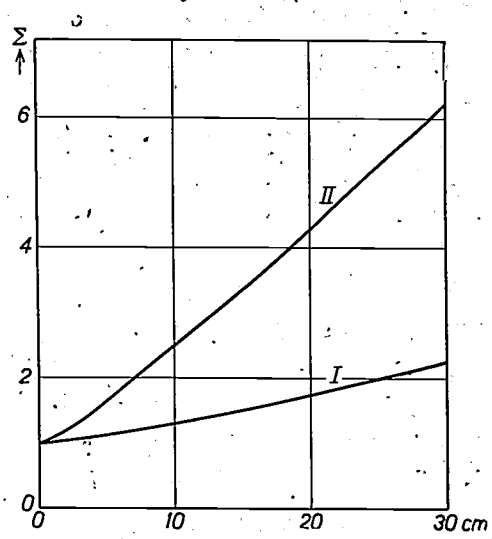


Fig. 6. The selectivity Σ of the method of increasing the distance between patient and film. The measurements were carried out with the same arrangement as that of fig. 4. Along the abscissa is plotted the distance from the phantom to the dose-meter. Curve I is for a field of 30 × 30 cm², curve II for 8 × 8 cm², on the tube side of the phantom.

[4]) H. Laurell, Acta Radiologica, **12**, 574, 1931.

rather thick and this weakens also the primary radiation very much. For a selectivity of only 2, for example, the filter has to be so thick, that the

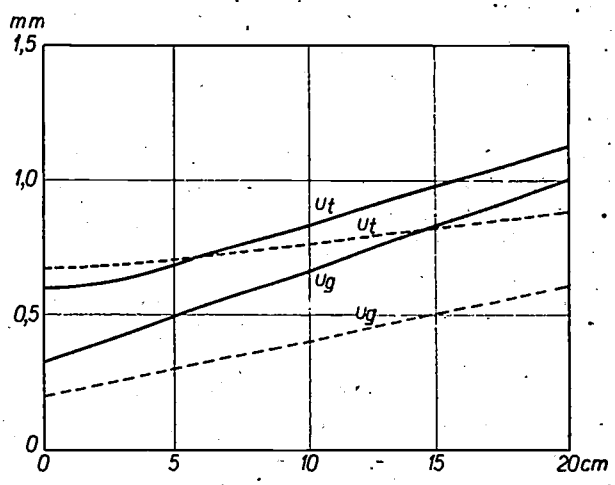


Fig. 7. The geometrical blurring u_g and the total blurring u_t, as functions of the distance in cm between patient and film, for a point in the middle of patient's body 20 cm thick. It is assumed, that there is a blurring due to the intensifying screens of 0.4 mm and a blurring due to movement of 0.3 mm for the full-line curves and 0.5 mm for the broken-line curves. The full-line curves are for a focus of 2.0 mm width, the broken-line curves for one of 1.2 mm. (The fact, that the broken-line curve u_t crosses the analogous full-line curve, is due to the uniformity condition on the left of the diagram, being approximately satisfied in the case of the full-line curves, and not on the right, while it is just the reverse in the case of the broken-line curves.)

intensity of the primary beam is reduced to 20%. In general, therefore, the employment of this method offers little advantage in diagnostics [5]).

5) *Method of slit diaphragms*

A very effective method is that sketched in *fig. 8* [6]). A slit diaphragm, with a slit of at most a few centimeters in width, is placed above and below the patient. The slits are moved along the patient, in a direction perpendicular to the slit, in such a way, that the lower slit allows the beam to pass, which has just passed through the upper slit. The adjacent parts of the object are thus exposed successively and the effect is the same, as if a number of photographs were made with a very narrow field, according to the principle of *fig. 3a*.

By making the slits narrow enough, the amount of scattered radiation can be reduced to any desired low value. A serious objection to the method, however, is the fact that the slits must move so slowly, that each strip of the film receives just as much radiation as in an ordinary exposure. The total exposure time is thus increased in the ratio of the width of the film to the width of the slit. This does not, of course, in itself lead to greater lack of definition due to movement, but the different strips of the object are photographed at appreciably different moments, so that the contours of rapidly moving organs, such as the heart or stomach, are seriously distorted [7]). Moreover, owing to the longer total time of exposure, the X-ray tube may not be loaded so heavily (for instance only by one half the power, which can be used for very short times), so that the exposure time for each strip has to be longer than for an ordinary photograph, which thus does, indeed, lead to an increase in the blurring.

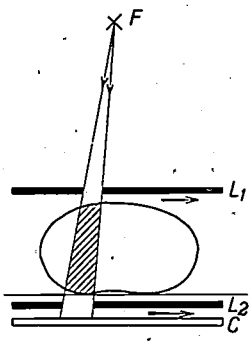


Fig. 8. Method of slit diaphragms. The two slits in the diaphragm L_1 and L_2 are moved along the object in such a way, that all parts of the latter are successively projected on to the film C. Thanks to the strict limitation of the field, each part of the film receives only little scattered radiation.

6) *The Potter-Bucky diaphragm*

One of the most important aids in counteracting scattered radiation is the Potter-Bucky diaphragm. In its simplest form it consists of a number of thin strips of lead, standing on their narrow edges, and placed in such a way between patient and film, that, seen from the position of the focus, each strip is as narrow as possible, see *fig. 9*. The scattered rays, in as far as their direction is not at too small an angle to that of the primary rays, strike the lead lamellae and are thereby very much attenuated. The primary rays on the other hand, thanks to the focussing effect of the lamellae, are allowed to pass quite freely except, where they strike the narrow edge of the lamellae, and except for a slight attenuation by the material between the lamellae. In order to preserve the distances between the lamellae and keep them in the correct position, they are placed in a filling material, like wood, organic synthethic material or in some cases a light metal, which is readily permeable for X-rays.

The value of this method is illustrated in *fig. 10*

[5]) To a small extent the filter effect is also obtained by the absorption of the intensifying screen placed in front of the film (see above). The filter method shows to better advantage and is therefore frequently applied in the technical examination of materials (rather than for diagnostic) where usually much higher voltages are used.

[6]) Chantraine, Röntgenpraxis 11, 37, 1939.

[7]) This effect is entirely analogous to the familiar distortion effect, which can be obtained upon photographing rapidly moving objects, such as race-cars, with a camera with slit shutter.

by four photographs of a special object with and without a Potter-Bucky diaphragm.

Because of its practical importance and the not uninteresting particulars of the technique of this method, we shall now go somewhat deeper into the subject of the Potter-Bucky diaphragm.

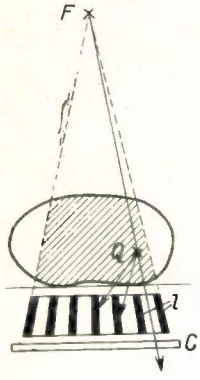


Fig. 9. Diagrammatic representation of a Potter-Bucky diaphragm. The lead lamellae *l* standing on their narrow edges and embedded in a filler, which absorbs only little X-radiation are so placed with respect to the focal spot *F* that the rays from the latter in their passage to the film *C* strike as much as possible of the open space between the lammelae. The scattered radiation from any point *Q* on the other hand is in most directions opposed by one or more lamellae.

Closer consideration of the Potter-Bucky diaphragm

Movement and focussing of the lamellae

In order to prevent the lead lamellae from producing disturbing shadow stripes in the X-ray photograph, the diaphragm is moved during the exposure in the direction perpendicular to the lamellae. The distance, over which the diaphragm must be moved, to render the shadows of the lamellae so faint as to be invisible is smaller the finer the diaphragm. In fluoroscopy, a continuous rapid movement of the diaphragm would be necessary to prevent shadow stripes on the image. Since, however, the fluoroscope screen itself causes considerable blurring of the image, diaphragms with sufficiently fine lamellae (the so-called fine-grid diaphragms with a lamellae separation of 0.5 mm or less) can be used stationary for fluoroscopy, without the stripes being disturbing.

It should be noted, that even with a moving diaphragm stripes may still appear on the X-ray photograph, namely when an X-ray apparatus with pulsating high-voltage is used, and when the frequency of the pulsations of the high voltage and the frequency with which a point of the film is shielded, owing to lead lamellae being passed at regular intervals, occur, in respect to each other, approximately in the order of whole numbers. By a suitable choice of the velocity of the motion this "stroboscopic" effect can be avoided.

In the case of a plane grid of lamellae, it will be moved in its own plane perpendicular to the direction of the lamellae. The diaphragm is then accurately "focussed" at its centre position, see *fig. 11a*. If the movement amounts to only a few centimeters to the right and left, the deviation of the focussing at the two extreme positions is not serious.

There are two methods of construction of diaphragms in which the lamellae always remain focussed, even during motion. In the one case, the lamellae are not placed on a flat plane, but along

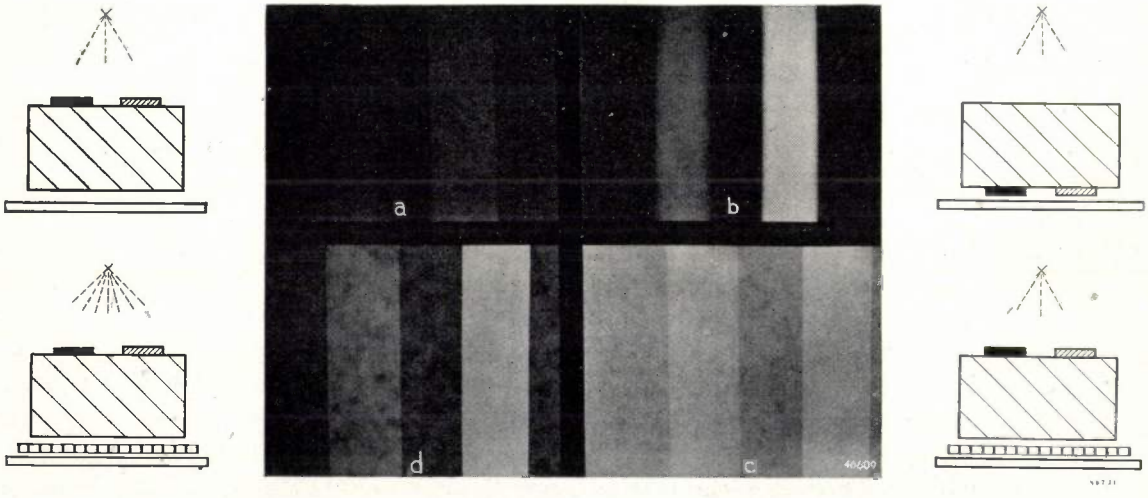


Fig. 10. X-ray photographs of a lead and an aluminium strip on a block of paraffin. *a*) The strips lay on the side of the block closest to the X-ray tube; the shadow picture is quite spoiled by the scattered rays. *b*) The strips lay between the block and the film; the shadows of the strips are now sharp, since the scattered rays, which arise in the paraffin help, as it were, to cast the shadows. The structure of the paraffin block itself, however, remains practically invisible, owing to the scattered radiation. *c*) The same as (*a*) but now with a moving Potter-Bucky diaphragm and the same exposure time. The two strips are sharp, but the picture is under-exposed. *d*) The same as (*c*) with an exposure time three times as long. Here, due to the limitation of the scattered radiation, the structure of the paraffin, characterized by numerous air bubbles, is also visible.

describing lines on the surface of a horizontal cylinder, whose axis passes through the focus (*fig. 11b*) and which can rotate about its axis. The disadvantage of this cylindrical diaphragm is, that a fairly large distance is necessary between patient and film, which increases the geometrical lack of definition, a disadvantage which also occurs with the plane diaphragms, although to a smaller extent, and which necessitates making the diaphragms as thin as possible. In the case of the second method of construction with permanent focussing, the lamellae are placed in the same plane as in *fig. 11a*, but radially, with the axis of symmetry through the focus, instead of parallel to each other (*fig. 11c*). During the exposure, the diaphragm rotates about that axis. A disadvantage, in this case, is that a white spot appears on the photograph at the position of the axis of rotation. An advantage, on the other hand, is that the diaphragm is suitable for every distance from the focus. The cylindrical diaphragm is, of course, only to be used for one distance, and the same is true of the simple plane diaphragms with parallel lamellae (even in the stationary condition). This distance is usually from 70 to 100 cm. In the case of the plane diaphragm, where a slight defocussing occurs in any case upon movement, small deviations from this distance are permissible; the transmission for the primary rays is then smaller at the edges than in the middle.

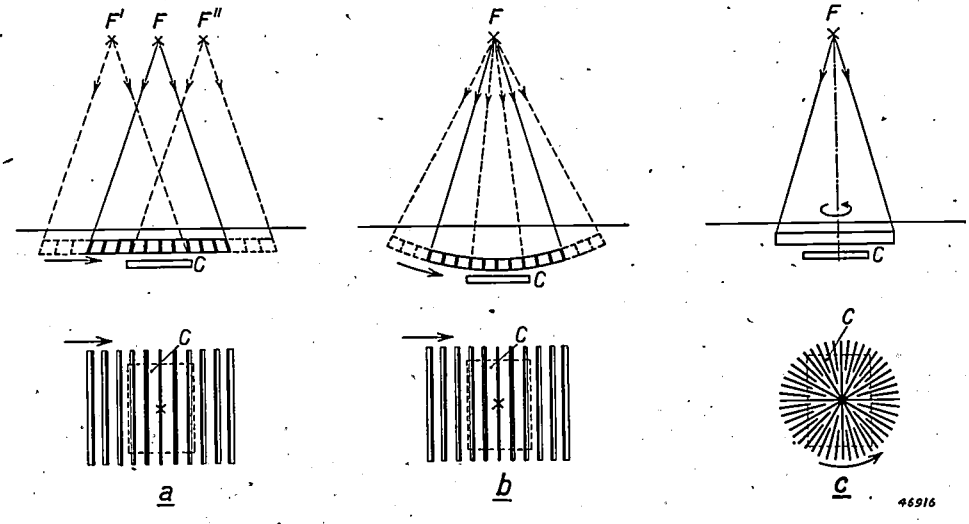


Fig. 11. Three forms of moving Potter-Bucky diaphragms. *a*) Lamellae in one plane with movement of the whole in that plane. In the position to the right or left the lamellae are focussed toward points F' and F'' respectively, instead of the focus F. The primary radiation from F is therefore more attenuated at those moments than when they are in the intermediate position. *b*) Lamellae mounted on the surface of a cylinder whose axis passes through the focus, with a swinging motion around this axis. The lamellae are always in the correct position. *c*) Lamellae placed radially in the same plane as in (*a*), about an axis of symmetry through the focus F about which the whole can rotate. The lamellae are always in the correct position.

Dimensions of the lamellae

In order to be able to give some guidance for the choice of thickness, height and distance between the lamellae, the action of the Potter-Bucky diaphragm must be considered in rather more detail. For the sake of simplicity, we shall confine ourselves to the consideration of a plane diaphragm with parallel, vertical lamellae, see *fig. 12*.

The part of the primary radiation R_p, falling between the lamellae, is transmitted almost without any weakening. But also a more or less large part of the oblique scattered rays R_s is passed unweakened, if the angle α with the primary rays is smaller, than the limiting angle α_g. With larger values of the angle α, the scattered rays would be entirely cut off, if the lead lamellae were absolutely impermeable for them. The transmission of the diaphragm, as a function of the angle of incidence of the rays, would then be as represented by curve I in *fig. 13a*. Actually, the lead lamellae may not be considered as entirely impermeable, especially as they may be very thin. With angles $\alpha > \alpha_g$, therefore, scattered radiation is indeed transmitted, though very much attenuated. The larger the angle of incidence, the more lamellae are struck by the radiation; the path in each lamellae, however, becomes shorter. The result is, that the scattered radiation transmitted at $\alpha > \alpha_g$ first decreases slowly with increasing value of α and then more rapidly. This can be represented

diagrammatically by curve *II* in fig. 13*a*. This curve will, of course, also depend very much upon the tube voltage used, since the absorption in the lead decreases rapidly with increasing hard-

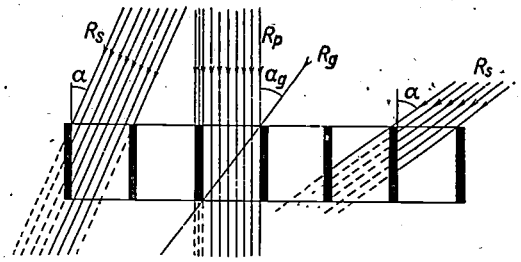


Fig. 12. Simplified Potter-Bucky diaphragm. Beams of rays striking the diaphragm at different angles are attenuated to different degrees. At angles larger than the limiting angle α_g no ray can pass through the diaphragm without attenuation.

ness of the rays. In *fig. 13b* the measured transmission of a fine-grid diaphragm is given for two different tube voltages. It may be seen, that these curves are not actually so smooth as the theoretical curve in fig. 13a, and, that they show slight maxima and minima.

From the curves of fig. 13, it is possible to deduce the selectivity Σ, to be obtained with a Potter-Bucky diaphragm. In this case, the selectivity can be set equal to the ratio between the transmission for the primary rays ($\alpha = 0$) and the average transmission for the scattered rays ($\alpha \neq 0$). A measure of the latter is simply the area of the surface below the transmission curve (crosshatched in the case of curve II in fig. 13a). Thus the smaller this surface the greater the selectivity [8]).

The surface below the curve becomes smaller, according as the limiting angle α_g is smaller and the more or less horizontal part of the transmission curve lies lower. Calculation shows, that the height of the flat part is determined by the amount of lead used per square cm of surface of the diaphragm, regardless, whether that lead is used in the form of thick or thin, high or low, many or few lamellae. Much lead and small limiting angle (high lamellae with narrow spaces between) are thus favourable for selectivity. But, as already stated, the transmission for the primary rays may not be reduced too much. In order to make more precise conclusions possible, we have calculated the transmission curves for several configurations of lamellae at a tube voltage of 105 kV, and they may be seen in *fig. 14*. *Fig. 14a* refers to the same type of diaphragms, as that with which the measurements of fig. 13 were made. The transmission curve is identical with curve *II* in fig. 13a. The selectivity calculated from this is $\Sigma = 6.3$ and the transmission for the primary radiation 72%. In *fig. 14b* the thickness of and distance between the lamellae are halved. The total amount of lead has thus remained the same, but the limiting angle is half as large. The selectivity has now risen to 7.8, while the transmission for the primary rays remains unchanged at 72 %. The greater the subdivision of the lead lamellae the greater the selectivity. In the case of the diaphragm of *fig. 14c* the distance between the lamellae and their thickness are the same as in *a*, but their height has been doubled. The limiting angle is thus equal to that in *b* but, thanks to the doubling of the amount of lead, the selectivity is considerably improved, to $\Sigma = 17$. The transmission for the direct radiation is slightly decreased, to 68%, due to the greater attenuation in the material between the lamellae (thicker layer). From the above it follows, that diaphragms *a* and *b* do not by any means represent the most favourable possibility for 105 kV, but that more lead is desirable for that voltage; for lower voltages, however, *a* and *b* may be quite suitable. In this way, for every tube voltage, the minimum amount of

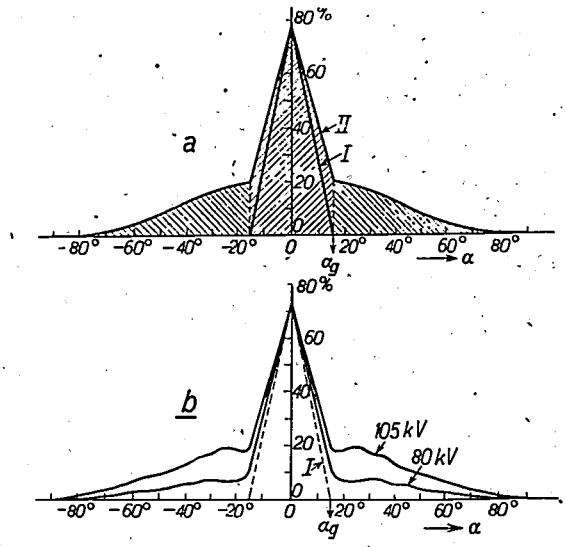


Fig. 13. The transmission of a Potter-Bucky diaphragm according to fig. 12, in percent, as a function of the angle α between the beam in question and the primary beam. *a*) Calculated for an extremely soft radiation ("zero" kV, complete absorption in the lead, curve *I*) and a hard radiation (105 kV, curve *II*). *b*) Measured, for a diaphragm with the same limiting angle α_g (ideal curve *I*), at 80 kV and at 105 kV. The total area, below each curve, is a measure of the amount of scattered radiation.

[8]) This is not strictly correct. It would only be true, if the scattered radiation were distributed spherically symmetrically, and thus the intensity of the scattered radiation falling on the diaphragm were the same from all directions. Actually, the intensity of the scattered radiation in the forward direction (small value of α) is the greatest, so that the attenuation of the scattered rays at larger angles α is of less relative importance and the actual selectivity becomes smaller than that calculated on the assumption, that all parts of the crosshatched area are of the same value. For simplicity's sake, we will adhere to this assumption; the discrepancies have no effect on the validity of the conclusions to be drawn.

lead can be calculated, which a scattered ray diaphragm should contain. The condition, thus deduced, is always easily satisfied in the case of coarse diaphragms, where the lamellae must be rather high in order to obtain a sufficiently small limiting angle. In the case of fine diaphragms, however, this condition must certainly be taken into account. Given a certain amount then, according to the above, the rule for increasing selectivity is, that the lamellae should be as thin as possible. The extra advantage is then obtained, that the disturbances due to stripes on the picture are less troublesome and easier to avoid. A limit is set to the reduction in the thickness of the lamellae mainly by difficulties in constructing the diaphragm.

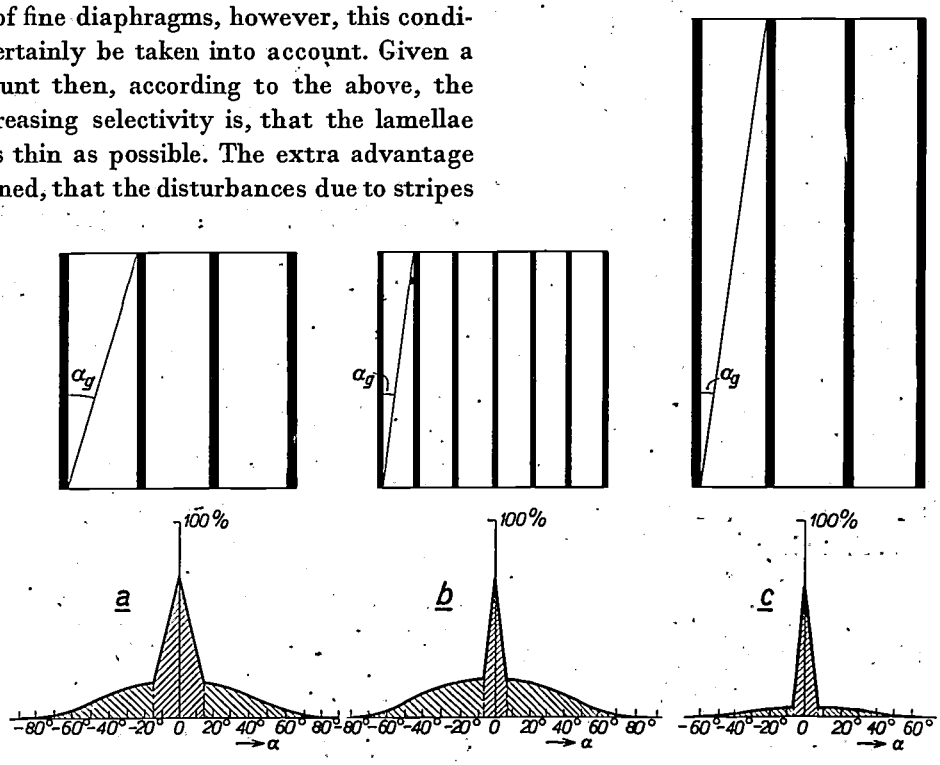


Fig. 14. Diagram of the form and spacing of the lamellae for three different Potter-Bucky diaphragms, and corresponding transmission curves for 105 kV. While *a*) and *b*) do not differ very much in quality, diaphragm *c*) proves to be considerably better. The calculated selectivity Σ is for *a*) 6.3, for *b*) 7.8 and for *c*) 17, while the primary radiation is attenuated to 72, 72 and 68 percent respectively.

Concluding remark

It is of importance for practical application to note, that none of the methods considered, is able to eliminate the scattered radiation entirely, but that, on the other hand, most of the methods can be used in combination with each other to obtain better results than with one alone. Thus, for example, even when a Potter-Bucky diaphragm is used, the field irradiated can always be made as small as possible. In connection with the method of compressing the tissue, it has already been stated, that a combination with the method of increased distance between object and film is desired. The method of the two moving slits can very well be combined with the application of the Potter-Bucky diaphragm, with the lamellae perpendicular to the direction of the slits, etc. Having regard to the drawbacks attaching to some of the methods, such as possible stripes with the Potter-Bucky diaphragm, increased blurring, when increasing the distance between object and film, distortion with the moving slits, etc., the doctor must decide from case to case, which combination of methods for eliminating scattered radiation is most suitable.

VOL. 8 No. 7, p. 193-224 JULY 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

AN EXPERIMENTAL RECEIVER FOR ULTRA-SHORT-WAVE RADIO-TELEPHONY WITH FREQUENCY MODULATION

by A. van WEEL.

621.396.5: 621.396.621

A description is given of a superheterodyne receiver for frequency-modulated signals, which forms part of an experimental ultra-short-wave radio telephony link on wave lengths in the neighbourhood of 1 m. This link was designed for the simultaneous transmission of 48 calls ("channels") and is therefore able to handle modulation frequencies up to about 200 kc/sec. The most interesting detail of the receiver is the push-pull mixing stage, equipped with triodes, which is made self-oscillating by introducing, in addition to the normal symmetric push-pull input circuit tuned to the signal frequency, an "asymmetric" input circuit tuned to the local oscillator frequency, which is generated in the same circuit as a consequence of the coupling between anode and grid circuits. As no separate oscillator valve is necessary, an appreciable part of the fluctuation noise is eliminated. As a consequence of the favourable properties of the mixing stage, as far as the noise is concerned, a high-frequency amplifier stage would give only little improvement in the ratio between intensity of signal and noise. The receiver, therefore, does not possess a high-frequency amplifier stage. The provisions for automatic volume control, usually found in receivers for amplitude-modulated signals, could also be omitted in this frequency-modulated receiver, thanks to the large suppression factor of the limiter.

A new transmitting-receiving apparatus has been developed for the experimental ultra-short-wave radio-telephonic link between the Philips factories in Eindhoven and those in Tilburg, which link has been in existence for a number of years already. After having described the transmitter in a previous article [1]), we shall now discuss the receiver. For the convenience of the reader we shall repeat briefly the most important facts about the installation. For the connection in one direction a wave length of 90.5 cm is used, in the other direction 99 cm. The transmitter and receiver function as link in a carrier telephony system with which 48 channels can be transmitted at the same time on one pair of conductors (in this case, on one radio wave). For this purpose both the transmitter and the receiver must be able to handle modulation frequencies up to about 200 kc/sec. For this link frequency modulation is employed. The maximum frequency swing, *i.e.* the largest deviation of the frequency emitted, compared with the average transmitter frequency (332.1 Mc/sec for one direction, 303.0 Mc/sec for the other), amounts to 0.6 Mc/sec.

General construction of the receiver

The receiver works on the superheterodyne principle. A block diagram of the most important parts is given in *fig. 1*. The signal received by the

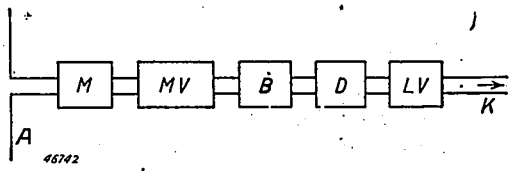


Fig. 1. Block diagram of the receiver; *A* aerial, *M* mixing stage, *MV* intermediate-frequency amplifier, *B* limiter, *D* frequency detector, *LV* low-frequency amplifier, *K* cable to the carrier-telephone apparatus.

aerial is applied to a mixing stage where the signal frequency is converted into an intermediate frequency of 18 Mc/sec. The converted signal is amplified and then passed through a limiter, the function of which is to suppress any amplitude modulation which may be present, especially the noise which

[1]) A. van Weel, Philips Techn. Rev. 8, 121, 1946.

occurs in that form [2]). The output of the limiter is detected and the low-frequency signal obtained is amplified and then passed through a cable to the apparatus for carrier telephony in the telephone exchange, where the 48 channels are split up and can be connected with the corresponding subscribers.

In principle this construction of the receiver does not differ from the normal receiver for frequency-modulated signals. However, it will be noted that there is no high-frequency amplifier stage preceding the mixing valve. Such a stage is in general desirable to improve the ratio between the intensity of the signal and that of the noise. In our case, however, the mixing stage has such good properties, as far as noise is concerned, that the addition of a high-frequency amplifier stage could have produced only a very small improvement. That stage could therefore be omitted, which meant a considerable simplification of the receiver.

In the following we shall discuss mainly the design of the mixing stage, while the other parts of the receiver will only be considered briefly.

Principle of the mixing stage

Mixing in triodes

The familiar multigrid mixing valves, generally employed for the conversion of signals on wavelengths longer than about 10 m, are no longer suitable for wave lengths of about 1 m, their special advantages being lost in this case because: 1) the shielding effect of the various screen grids has practically no result, since, due to the self-inductions and mutual inductions of the internal leads to these screen grids at these high frequencies, it is practically impossible to keep the grids free of high-frequency voltage variations; 2) the high internal resistance of the multigrid mixing valves, seen from the intermediate-frequency circuit, means no advantage because, to amplify the large frequency bandwidths required, heavily damped circuits have to be used. On the other hand the disadvantage of all multigrid valves — the occurrence of the so-called distribution noise caused by fluctuations in the distribution of the cathode current over the various current carrying grids — is the same with short waves as with longer waves. In the ultra short wave region, therefore, diodes or triodes are used as mixing valves.

The advantage of a diode mixing stage lies in the high input impedance, which makes a high step-up ratio of the signal delivered by the previous stage. However, the conversion amplification, *i.e.* the ratio between the intermediate output voltage and the high-frequency input voltage, cannot exceed unity.

When a triode is used as a mixer the input impedance is smaller than with a diode. On the other hand the conversion amplification is in general considerably larger than unity.

Further, it has been found from theoretical investigations [3]) that here a good ratio can be obtained between signal intensity and noise. These facts, together with another important advantage which will be discussed farther on, led us to employ triode mixing in the receiver in question.

Mixing in a triode is in principle accomplished by applying to the grid of the triode, together with the high-frequency signal voltage, a second voltage from a local oscillator (usually indicated briefly as the "local oscillator voltage").

The anode current then contains an intermediate-frequency component (difference of the frequencies of the two voltages mentioned), and by introducing in the anode circuit a circuit tuned to the intermediate frequency the desired intermediate-frequency output voltage is obtained.

At the high frequencies with which we are dealing here it is in general an advantage to construct amplifier and mixing stages on the push-pull principle [4]). The following general rule is then valid: of the three voltages occurring in a mixing stage, namely high-frequency signal voltage, local oscillator voltage and intermediate output voltage, two must always be in push-pull, while the third has to be in the same phase for the two mixing systems. Because of the symmetrical dipole aerial it is reasonable to apply the high-frequency signal in balance to the convertor valves. We now apply the equal-phased local oscillator voltage in the same sense to both valves and then obtain the intermediate frequency voltage in balance again. In that way we arrive at the diagram sketched in *fig. 2*, showing the principle of a push-pull mixing stage with two triodes. The connection of the anodes with the intermediate-frequency circuit and the source of anode voltage are omitted temporarily for the sake of simplicity. Between the control grids there is a circuit L_1-C_1 tuned to the signal frequency, which is inductively coupled with the aerial. In this way an e.m.f. in the aerial gives rise to grid a.c. voltages which are equal in magnitude

[2]) See for example Th. J. Weyers, Philips Techn. Rev. 8, 42 and 89 1946.

[3]) See A. van Weel, thesis, Delft 1943.

[4]) See in connection with this and the following: M. J. O. Strutt and A. van der Ziel, The diode as a frequency-changing valve, especially with decimeter waves, Philips Techn. Rev. 6, 285, 1941. In that article will be found explanations and the literature about the properties of valves on decimeter and meter waves, which have merely been mentioned in the above.

but opposite in phase for the two triodes (thus balanced). Between the middle of the circuit self-induction L_1 and the earth point (*i.e.* the chassis) there is an oscillator for the auxiliary voltage, so

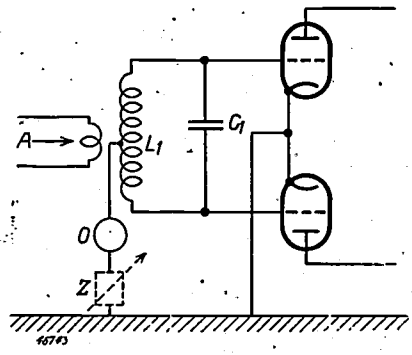


Fig. 2. The signal from the aerial is applied to the grids of the two triodes of the push-pull mixing stage by inductive coupling with a tuned push-pull input circuit (L_1-C_1). The auxiliary voltage (oscillator voltage) delivered by the local oscillator Q is applied to the grids in equal phase. The "asymmetric" circuit can be tuned to the oscillator frequency by an impedance L. The intermediate-frequency circuit to be connected between the anodes is here omitted for the sake of simplicity.

that the latter comes in equal phase between the grids of the two triodes and the cathode, which are connected with the earth point.

Separate tuning of the oscillator circuit

With such connections it is often found to be difficult to obtain a sufficiently high auxiliary voltage on the mixing valves, due to the fact that the asymmetrical circuit (earth point — oscillator — circuit self-induction — grid-cathode capacities of the valves — earth point) is not in general tuned to the oscillator frequency. This tuning can, however, easily be realized by including a tuning element somewhere in the circuit. Since the push-pull circuit, tuned to the signal frequency, may not be affected by this second tuning, the tuning element should be included in a part of the circuit through which no balanced currents flow. In fig. 2 the element in question is shown with a dotted line.

It is perhaps advisable to place some emphasis on the principle applied here: *to the input electrodes of two valves in push-pull connection two independent circuits tuned to different frequencies can be connected, viz. as push-pull circuit and an asymmetric circuit.* The separate tuning of each circuit can be varied with the help of circuit elements through which flow, respectively, only balanced currents or only equal-phased currents.

Excitation of the auxiliary voltage by the mixing connections themselves

In order to eliminate the objection of the above-mentioned low input impedance of the triodes at high frequencies, back-coupling is employed in the asymmetric circuit. This causes a reversal of damping in that circuit or, in other words, an increase in the input impedance of the valves and with it an increase of the auxiliary voltage obtained on the input electrodes.

If the reversal of the damping is made large enough (the back-coupling strong enough) the circuit itself begins to oscillate, and it oscillates at the frequency to which the asymmetric input circuit is tuned, *i.e.* the frequency of the auxiliary voltage.

A separate oscillator is then not needed for the excitation of this auxiliary voltage. As far as the conversion effect of the triodes is concerned it makes no difference whether the auxiliary voltage between grid and cathode comes from a separate oscillator or whether it is excited by the triodes themselves. The mixing action is determined exclusively by the magnitude of the auxiliary voltage and the non-linearity of the i_a-v_g characteristic.

The back-coupling mentioned in the case of a triode can be obtained in the familiar way by conconecting a self-induction between the anode and the cathode.

This can be understood most easily by writing the equations for the scheme shown in *fig. 3*. If for the sake of simplicity we disregard the internal resistance of the valve these equations are as follows:

$$v_g = \frac{i_{g1}}{j\omega C_{ag}} - (i_a - i_{g1})\, j\omega L = \frac{i_{g2}}{j\omega C_{gk}},$$

$$i_a = S \cdot v_g.$$

From this it follows that:

$$\frac{i_g}{v_g} = \frac{i_{g1}}{v_g} + \frac{i_{g2}}{v_g} = j\omega C_{gk} + \frac{1}{j\omega L + \frac{1}{j\omega C_{ag}}} - \frac{\omega^2 SLC_{ag}}{1 - \omega^2 LC_{ag}}.$$

While with short-circuited self-induction L the triode, as an element of the input circuit, can be replaced by the capacities C_{gk} and C_{ag} connected in parallel, it may be seen from the last equation that by the intermediate connection of the self-induction the equivalent connections become a connection in parallel of C_{gk} with L and C_{ag} in series and a negative resistance. The latter causes the desired increase in the input impedance of the valve, since the admittance i_g/v_g is lowered by that term.

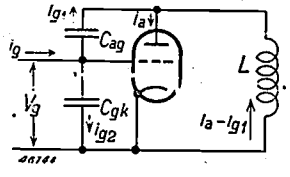


Fig. 3. Back-coupling of a triode *via* the grid-anode capacity C_{ag} and a self-induction L in the anode circuit of the triode. The total grid current i_g is composed of the two parts i_{g1} and i_{g2}, of which only i_{g1} is important for the back-coupling.

In the push-pull mixing stage the self-induction is introduced in the manner shown in principle in *fig. 4*. The equal-phased circuit, which is tuned to the desired oscillator frequency by means of the variable impedance Z, is now so strongly regenerated by the self-induction L_2 as to cause oscillation.

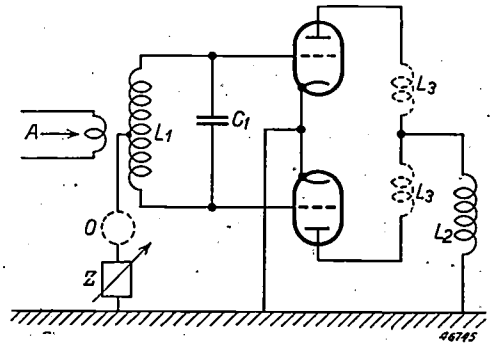


Fig. 4. The "asymmetric" input circuit of the push-pull mixing stage is regenerated by the self-induction L_2. The push-pull input circuit can be regenerated quite independently by the self-induction L_3.

It must be pointed out that the push-pull circuit can also be regenerated in a similar manner. For that purpose the two self-inductions L_3 shown with dotted lines in fig. 4 have to be placed in the anode circuit. In this way it is possible to obtain a stronger excitation of the signal delivered by the aerial, and thus a better relation between signal and noise [5]). The degree of back-coupling can be chosen different for the push-pull circuit and for the asymmetric circuit, but L_2 has no effect on the push-pull circuit. This is important because the back-coupling for the push-pull circuit may not, of course, be made so great that oscillation (with signal frequency) occurs in that circuit.

The ratio of signal intensity to noise

The fact that a separate oscillator valve can be omitted constitutes of itself a welcome simplification of the connections. Still more important is the fact that an improvement in the relation between signal intensity and noise is thereby obtained. This can easily be explained. In the anode current of an oscillator valve, as is the case with every valve, there is a noise with a continuous frequency spectrum. The circuit connected with the anode circuit of the oscillator valve, which is tuned to the oscillator frequency, acts as a kind of filter which passes chiefly the oscillator frequency but to a certain extent also the neighbouring signal frequency and, to a smaller extent, the much lower intermediate frequency. The components of the noise with those frequencies are thus also applied more or less intensely to the mixing valve as input voltage together with the oscillator voltage proper. The noise at signal frequency is then also converted with the signal, the noise at intermediate frequency is directly amplified, and both therefore give rise to an increased level of noise in the intermediate-frequency output voltage of the mixing valve.

It is clear that with the omission of the separate oscillator valve this extra contribution to the noise is eliminated.

Choice of the intermediate frequency

The above-described effect of the noise of an oscillator valve is stronger according as the intermediate frequency chosen is lower. This means that the signal frequency lies very close to the oscillator frequency, so that the component of the noise at the signal frequency is passed by the tuned circuit of the oscillator almost in full strength. From this point of view, therefore, the intermediate frequency should be as high as possible.

Although in our case, due to the omission of the oscillator valve and the noises inherent therein, this consideration is no longer applicable, as far as the choice of the intermediate frequency is concerned one reaches the same conclusion. In the above considerations it was tacitly assumed that the symmetry of the push-pull stage is complete, so that the balanced and asymmetric circuits do not affect each other at all. In practice there will always be a certain asymmetry and consequently a coupling between the two circuits. If, now, the characteristic frequencies of the two circuits lie close together (*i.e.* if the intermediate frequency is low), the oscillation of one circuit causes resonances in the other, with the further consequence of unequal voltages on the two valves, etc. The intermediate frequency may not, therefore, be chosen too low. Upon closer consideration it is found that it should amount at least to 1/20 of the signal frequency in order to render the coupling in question harmless. The receiver described here works with an intermediate frequency of 18 Mc/sec; the condition just given is therefore satisfied up to a signal frequency of 360 Mc/sec, *i.e.* a wave length of 83 cm.

Complete connections of the mixing stage

In *fig. 5* a diagram is given of the complete connections of the mixing stage, while *fig. 6* shows a photograph of the stage when constructed.

To the grids of the two button triodes used as mixing valves a Lecher system is connected, which together with a condenser (C_1) forms the

[5]) A. van Weel, loc. cit.

push-pull input circuit of the mixing stage. The capacities C_2 serve only for separating the D.C. voltage situation of the two grids. The desired regeneration of the push-pull input circuit is obtained by the self-induction L_3 introduced between the anodes.

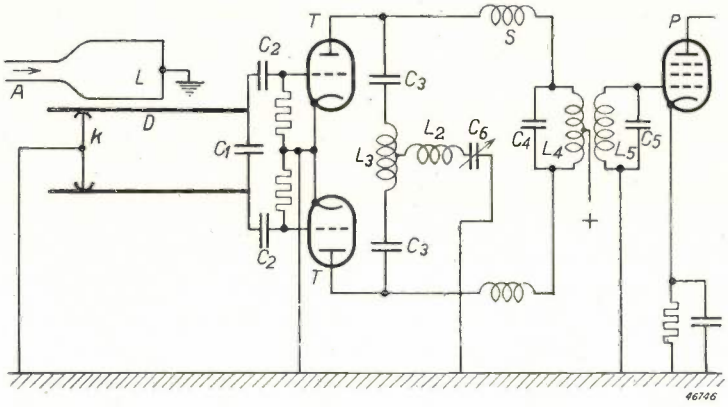


Fig. 5. Diagram of complete connections of the mixing stage. T triodes, D Lecher system with sliding short-circuiting bridge K and tuning condenser C_1, A aerial connection with coupling loop L; L_2 and C_6 back-coupling elements of the "asymmetric" input circuit, in which C_6 also serves for the fine tuning to the local oscillator frequency, L_3 back-coupling element of the push-pull input circuit, C_4-L_4-L_5-C_5 band filter for the intermediate frequency amplifier, S high-frequency choke coils, C_2, C_3 coupling condensers.

The asymmetric input circuit of the mixing stage is formed by the two conductors of the Lecher system, which for this purpose are to be considered simply as two separate impedances through which rectified currents pass. This input circuit is closed by the chassis, since the sliding short-circuiting bridge of the Lecher system is connected directly with the chassis. The back-coupling for the equiphased oscillation is accomplished by the self-induction L_2 connected to the middle of L_3 (to which has to be added the two halves of L_3 to be considered as connected in parallel). The magnitude of this self-induction, however, not only determines the degree of back-coupling but also has an influence on the frequency generated. This may be seen from the equation derived in connection with fig. 3, which indicates that L_2 introduces not only a negative resistance but also a reactive component in the equivalent connections for the valve. Since the adjustment of the back-coupling is not very critical, advantage can be taken of the influence mentioned to regulate the oscillator frequency more exactly by varying L_2. Since, however, a continuously variable self-induction is less easy to realize, for this purpose a rotating condenser (C_6) is introduced in series with L_2.

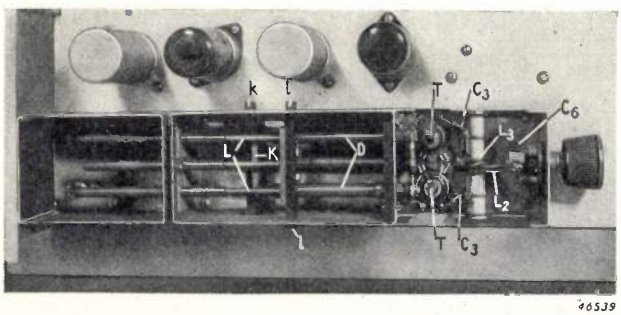


Fig. 6. The push-pull mixing stage with shielding cap removed, showing the two button triodes T and the input Lecher system D, the short-circuiting bridge K of which can be moved by means of the projections k. In front of that system may be seen the "loop" L, which can be varied with the projections l and which makes the coupling with the aerial cable and whose middle point remains in connection with the earth point (the chassis) by way of the middle conductor. To the right of the button triodes may be seen the capacities and self-inductions indicated in fig. 5, C_3, C_6, L_3, L_2.

The tuning of the push-pull and asymmetric circuits is now accomplished as follows. After L_2 and C_6 have been so chosen that sufficient regeneration has been obtained for the occurrence of oscillation in the asymmetric circuit, that circuit is tuned approximately to the oscillator frequency, by moving the short-circuiting bridge of the Lecher system. By a correct choice of C_1 the push-pull circuit is then tuned approximately to the signal frequency, while the fine tuning is obtained again by a slight movement of the short-circuiting bridge, and the fine tuning to the oscillator frequency is finally obtained with C_6. This rather laborious manipulation constitutes no objection in our case, since the receiver always remains set on the same wave length. Finally, there is the intermediate-frequency circuit. This consists of the self-induction L_4 and the capacity C_4. Between the triodes and this circuit the high-frequency choke coils S are connected in order to prevent the high-frequency circuits from being affected by L_4-C_4. Conversely, the capacities C_3,

which have a high impedance for the intermediate frequency and a low impedance for the signal and oscillator frequency, provide that the circuit L_4-C_4 is not practically short-circuited by the small self-induction L_s.

Remaining parts of the receiver

The intermediate-frequency amplifier consists of a number of stages with EF 51 valves. The valves are coupled by band filters, but in one stage connections are used like those described in the article about the transmitter of this installation [1]). By this means it was possible (as in the case of the transmitter) to divide the receiver into two parts in this stage, thus at the middle point of the intermediate-frequency amplifier, and to house the two parts in two separate panels without it being necessary to shield the connections between the panels. The two panels may be seen in the photograph of *fig. 7*.

After the intermediate-frequency amplification a limiter is applied to the signal, as is customary in receivers for frequency-modulated signals. This consists of two valves provided with grid condenser and leak resistance. When the amplitude of the A.C. voltage on the grid of a valve in these connections increases, the grid becomes more negative due to the occurrence of grid detection. This reduces the anode current and, in spite of the increased input voltage, the output voltage of the valve shows little or no increase. In this way any amplitude modulation which may still be present in the frequency-modulated signals, and especially the noise present in that form, is suppressed to a large extent. The output voltage of the first limiter valve, which is already quite constant, is chosen so high that the second valve is adjusted at the point of the characteristic most favourable for the limiting effect. A total suppression factor of 100 is thereby attained, *i.e.* the depth of the amplitude modulation of the signal at the output of the limiter is 100 times smaller than at the input. Receivers for amplitude-modulated signals are usually equipped with an automatic volume control, with which a part of the output voltage of the last intermediate-frequency amplifier valve is rectified and used to regulate the grid bias of one or more of the preceding amplifier valves. In the case of a frequency-modulation receiver, when the limiter has a sufficiently large suppression factor such a volume control may be omitted, since the limiter already provides for a practically constant output voltage. In our case, therefore, all the valves of the intermediate-frequency amplifier work with constant negative grid bias. Account must then be taken, however, of the possibility that the last valve of the amplifier may be overloaded and damaged by a strong signal. In order to prevent this, that valve, as well as the valves of the limiter, is provided with a grid condenser and leak resistance in such a way that the valve simply begins of itself to act as a limiter as soon as the signal exceeds the permissible strength. The only result of an increase in the signal then is that there is a stronger suppression of any amplitude modulation which may be present, which at the most can but improve the quality of reception.

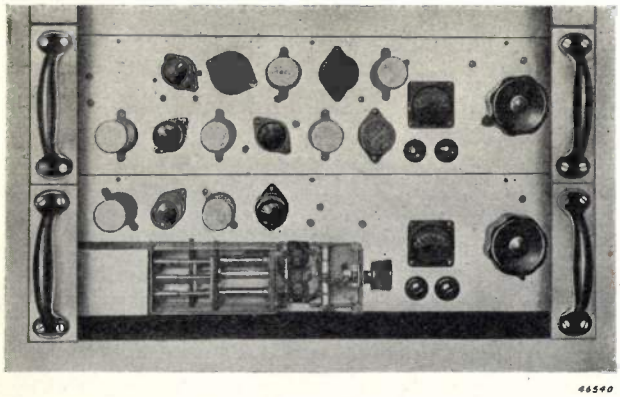


Fig. 7. The receiver is assembled in two panels, which can be slid in and out of the bay containing the whole transmitter-receiver installation by means of handles. Upon sliding them in, all the necessary connections are made automatically by means of plugs on the panels and sockets on the bay. The meters with the adjacent switching knobs on each panel serve for checking the cathode current of all the valves.

Following the limiter there comes finally a normal frequency detector and a low-frequency amplifier valve. The low-frequency signal obtained is conducted *via* a transformer to the cable leading to the carrier-telephony apparatus.

After the previous description of the construction of the transmitter it is unnecessary to go further into that of the receiver. For several particulars reference is made to figures 6 and 7 and the annotation thereby.

HIGH-VOLTAGE RECTIFIER VALVES FOR X-RAY DIAGNOSTICS

by J. H. van der TUUK. 621.314.671:621.386.1:616.073

The high-voltage rectifier valves used in X-ray installations, especially in diagnostic apparatus, have to satisfy in the first place the following requirements: they should be resistant to high voltages, furnish the necessary large peak currents with a reproducible and not too large voltage loss, and should not develop too much heat. In this article it is explained in what way two fundamental types of valves, namely vacuum and gas-filled valves, can satisfy the requirements mentioned. Gas-filled valves are found to possess several properties which are very favourable for the object in view. However, also in the field of vacuum valves constructions have recently been developed which give interesting results, thanks to the use of thoriated tungsten cathodes, and which may supplant entirely or for a large part gas-filled valves because of their universal applicability.

In principle, after being stepped up, the alternating voltage from the mains can be applied directly to an X-ray tube. In the half cycle in which the hot cathode is negative and the anode positive X-rays are excited, while in the other half nothing happens. The fact that in many cases it is, nevertheless, found preferable to apply direct voltage or at least pulsating direct voltage to the X-ray tube has two reasons. In the first place when X-rays are excited in both halves of the cycle, with a given peak value of the tube current, double the X-ray intensity is obtained and, moreover, upon smoothing the rectified voltage X-ray output becomes larger, compared to the heat developed. In the second place, when an X-ray tube works on A.C. sometimes something does happen in the "other" half of the cycle, and what happens is extremely undesirable, *viz.* the so-called back-lash. This may occur in the half cycle in which the anode is negative and starts to emit electrons, either by becoming too hot (thermal emission) or because the electric field at its surface exceeds a certain value (cold emission) or it is struck by positive ions of gas. Since such a back-lash often leads to the destruction of the tube it must be prevented under all circumstances. When rectified voltage is applied the problem of back-lash, which presents special difficulties in the case of the X-ray tube because of the great heat development on the relatively small focus [1]), is transferred to the high-voltage valves, which serve for the rectification. These valves are very similar to the X-ray tube itself: they, too, possess a cathode and an anode in a vacuum-tight envelope. Since, however, in the case of valves the electrons carrying the current in the direction of transmission need not be strongly accelerated and focussed, local intense heating of the anode can be avoided by suitable construction, so that this important source of back-lash is eliminated.

Let us now consider somewhat more closely the requirements demanded of high-voltage rectifier valves. From the above follows the general requirement of voltage reliability, *i.e.* resistance to the high-voltage in the negative phase without danger of back-lash. In the positive phase, of course, the rectifier valve must be able to pass the current necessary for the X-ray tube. In the case of X-ray tubes for diagnostics it is, at present, a question of short-lived peak currents of 1 to 1.5 A, while maximum voltages up to 125 kV occur. In the case of therapy tubes much smaller currents are used, for instance maximum 40 mA continuously, with, however, considerably higher voltages, *viz.* 200 to 400 kV.

We shall devote ourselves here especially to the high-voltage rectifier valves used for diagnostics. In this case there are several additional requirements. For easy operation of an apparatus for diagnostics it is desirable that a certain tube voltage always corresponds to a definite position of the voltage regulator; for that purpose the voltage loss depending on the current should be small in all the parts of the high-voltage generator (transformer, valves, connections) and it should be as independent as possible of the circumstances of operation. The valves may, therefore, even at the highest currents, only require a low and reproducible voltage in the direction of transmission. This is especially true when the apparatus is provided with an automatic adjustment for the tube current [2]).

In the case of modern apparatus for diagnostics the valves are often housed in the same oil-filled

[1]) See the discussion of the problem of back-lash in the article J. H. van der Tuuk, Hard-glass X-ray tubes in oil, Philips Techn. Rev. 6, 309, 1941.

[2]) Such an arrangement, which serves to load up the focus of the X-ray tube automatically to the permissible temperature at every adjustment of tube voltage and loading time, is discussed in detail in: H. A. G. Hazeu and J. M. Ledeboer, A universal apparatus for X-ray diagnosis, Philips Techn. Rev. 6, 12, 1941.

container with oil as the high-voltage transformer, etc. Since, in order to keep the voltage loss (*i.e.* the copper and iron losses) constant, the whole generator may not become too warm, the additional requirement should be made of the valves that they develop little heat. On the one hand this is already ensured if the valve only takes up a low voltage in the direction of transmission, since the anode dissipation is proportional to it; on the other hand it also makes desirable a high specific emissivity of the cathode, since then only a small filament power is used for the emission of the peak currents.

We shall now examine several old and new valve constructions and ascertain how, in each case, the requirements mentioned can be satisfied.

Vacuum rectifier valves

Fig. 1 shows diagrammatically the construction of the old vacuum rectifiers as they were used for many years in all kinds of high-voltage installations. The envelope of the valve is so far evacuated

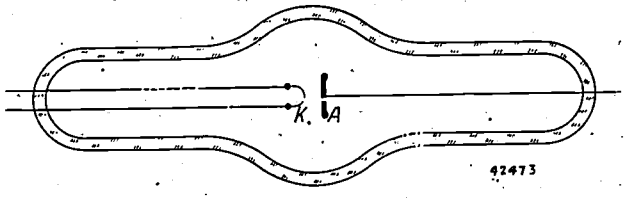


Fig. 1. Diagram of the cross-section of a vacuum rectifier valve as used formerly in high-voltage installations. An anode plate A is situated at a short distance opposite a tungsten cathode K in an evacuated glass tube.

that no breakdown can be caused by gas ions. The most important factor which should be taken into account to prevent back-lash is therefore the cold emission of the anode. In order to remove any such danger the field strength at all points on the surface of the anode should remain below a certain value, for instance 10^6 V/cm. For this purpose sufficient distance between the electrodes is necessary. Furthermore, the anode should be properly smooth-finished, in order to avoid local increases in the field strength, which are known to occur wherever there are slight irregularities in the surface. (In this respect also the problem of back-lash is easier to solve in the case of a rectifier valve than in that of an X-ray tube: in the latter case the anode surface at the position of the focus is always rendered slightly rough in the course of time by the intense heating (evaporation).)

As far as the distance between the electrodes is concerned, the larger it is made the greater the influence of the space charge on the variation of the field and the higher the voltages necessary in the direction of transmission to draw the desired current from the hot cathode. Therefore it is clear that the distance between the electrodes will not be made greater than absolutely necessary. For 100—125 kV, for example, a distance of 8—9 mm is sufficient. With a given valve according to fig. 1 the voltage loss amounted to about 2500 V for a current of 1A.

The small distance between electrodes, arrived at in this way, makes it impossible to use an oxide cathode, for there would be too great a danger of traces of barium from such a cathode striking the anode. Due to the increase in field strength at the unevenness thus formed (and due to the low work function of the barium) such a spot might act as a centre of cold emission in the counter-phase. Therefore in this type of valve tungsten cathodes are always used, where the danger mentioned cannot occur. Tungsten cathodes, however, require a much higher temperature for the same emission and a much higher cathode power than oxide cathodes. In order to supply momentary peak currents of 1—1.5 A, 125—150 W are needed with a tungsten cathode, compared with only about 8 W with an oxide cathode. In this case in valves of the type of fig. 1 temperatures of the tungsten wire of more than 2350 °C are reached. Since the evaporation of the tungsten is already very appreciable in this region of temperature (the vapour pressure lies in the neighbourhood of 10^{-6} mm of mercury) and as a consequence the filament has a lifetime of only a few hundred hours, the filament is often allowed to burn at a temperature of more than 100° lower during fluoroscopy (continuous operation), which already gives a lifetime of a couple of thousand hours, and the heating voltage is only increased to the necessary value for a moment just before each X-ray photographic exposure. Furthermore it is clear that fluctuations in the mains voltage have an unfavourable effect on the lifetime of the filaments when continually working so close to the limits. In order to improve this situation it is necessary to have recourse to the use of a stabilizer for the heating voltage.

The construction sketched in fig. 1 also presents difficulties as far as the reproducibility of the voltage loss is concerned. Secondary electrons from the anode strike the glass walls of the envelope, which are thereby charged and begin to exert a grid action on the current of electrons between cathode and anode. The potential to which the wall becomes charged depends upon all kinds of factors, for example on the occurrence of corona phenomena

outside the valve and the like. Although the relation between current transmitted and voltage loss of the valve at low measuring voltages is then often sufficiently reproducible, this is found to be no longer the case for operation under high-voltage. The voltage loss may then vary considerably and may sometimes be so high that the valve itself begins to emit X-rays to an appreciable extent and the anode becomes very hot.

Thus it is obvious that although the simple construction sketched might often answer well in installations for testing materials or for medical therapy and also for simple apparatus for diagnostics, where it is a question of maximum currents of a few hundred mA, it can certainly not be employed in every case in modern apparatus for diagnostics.

Gas-filled rectifier valves

An entirely different type of valve is found to satisfy better the requirements made of diagnostic apparatus. These valves are not highly evacuated but are filled with a gas to such a pressure that after ignition of the valve an arc discharge occurs. The rectifying action in this case is due to the fact that the ignition is initiated by the electron current emitted by the filament: in the half cycle in which the filament becomes positive normally there are not sufficient electrons and the valve does not ignite [3]).

Compared with vacuum valves we may conceive the situation during the arc discharge such that the space charge of the electrons emitted by the filament is compensated by positive ions of gas. Due to this, only extremely low voltages are necessary for passing even very large currents, for example 25 V at peak currents of 1.5 A. Such a small voltage loss can be entirely neglected with respect to the voltage on the X-ray tube.

The working voltage of the valve mentioned, in contrast to the case of the vacuum valves, depends little on the distance between the electrodes. Therefore, in constructing the valve there is no objection to have a rather large distance between the electrodes, which is of course desirable in order to avoid trouble from cold emission in the negative phase and to be able to use an oxide cathode.

Incidentally, in the case of gas-filled valves the cold emission is not the main cause of back-lash. Back-lash is more apt to occur due to a breakdown in the gas, resulting in an arc discharge in the wrong direction. The structural measures to be taken in order to avoid breakdown can be deduced from a consideration of the familiar Paschen curve, see *fig. 2*, which indicates the breakdown voltage V_d as a function of the product of gas pressure p and electrode distance d. If a given combination of gas

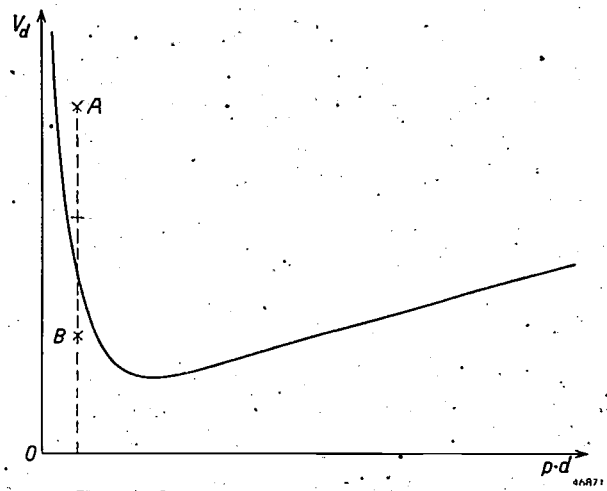


Fig. 2. Shape of Paschen's curve indicating for a given gas the breakdown voltage V_d as a function of $p \cdot d$ (p = gas pressure, d = distance between electrodes), for which, strictly speaking, two plane parallel electrodes are assumed. (Compare also Philips techn. Rev. 2, 123, 1937 where it may be seen from fig. 2 that for argon, for instance, the minimum breakdown voltage of more than 250 V is reached at a value of more than 1 mm of mercury times cm for $p \cdot d$.) The rise in V_d at very small values of $p \cdot d$ is due to the fact that at those values the free path of the electrons becomes greater than the distance between the electrodes, so that the electrons can pass through the potential difference without causing any appreciable ionization. The line AB, the significance of which is explained in the text, actually passes much closer to the ordinate than it is drawn here, and the point B lies in practice at about a hundred times the voltage of the minimum of the curve.

pressure, electrode distance and valve voltage V (highest voltage between the electrodes occurring in the negative phase) corresponds to a point A, which lies above the curve, breakdown will occur. In order to avoid breakdown, when p and V are given the electrode distance should obviously be chosen much larger or much smaller. In practice the first method is out of the question, since then one would arrive at enormous lengths of the tube for the high-voltages required. But the second method brings us into conflict with the requirement that the electrode distance should be large enough to exclude cold emission.

A solution of this dilemna is found in a simple way. Suppose that at point A (fig. 2) cold emission were precluded. The valve is then built up of a number of units in series, each with the same electrode distance as for point A, and to each of these compartments only a proportional part of the total voltage is applied [4]).

[3]) For the general principles of gas-filled rectifiers see: M. J. Druyvesteyn and J. G. W. Mulder, Philips Techn. Rev. 2, 122, 1937.

[4]) *Cf.* also J. G. W. Mulder, thesis Delft 1934.

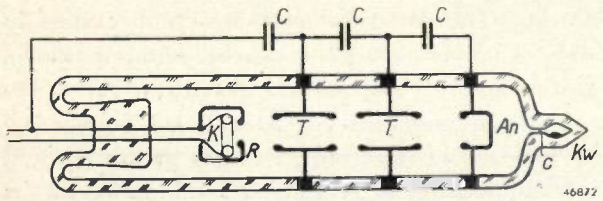


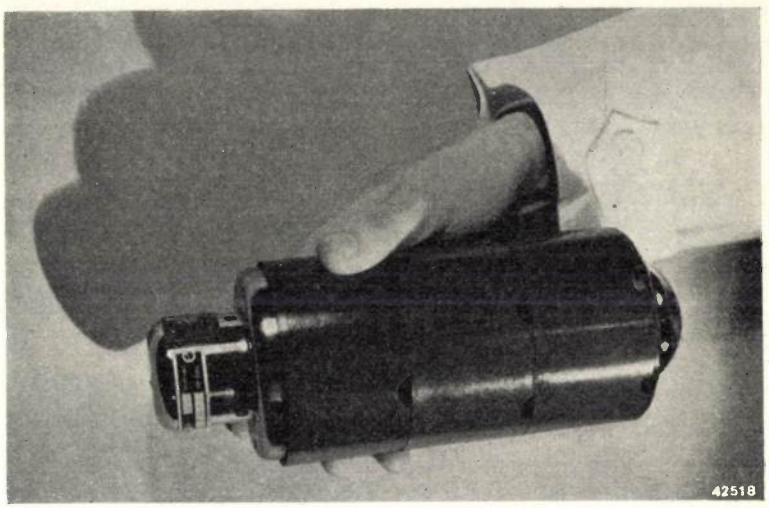
Fig. 3. Diagram of a cross-section of a gas-filled rectifier valve. The glass envelope is filled with saturated mercury vapour. The distance between the oxide cathode K and the anode An is divided into three stages by intermediate electrodes T, which are connected by the condensers C for the sake of a uniform voltage distribution. By a ring R placed in front of it the cathode is screened against bombardment by positive gas ions, while the ring at the same time lessens the chance of material from the oxide cathode striking the other electrodes. Kw is a mercury reservoir connected with the envelope by a capillary.

In this way one arrives at a work point for each separate compartment which, if the subdivision has been carried far enough, lies below the Paschen curve (B in fig. 2). Each compartment is then safe as far as breakdown is concerned, and since the field strength V/d (when the voltage is evenly distributed and the field homogeneous) is even smaller for point B than for point A[1], there is also no danger of cold emission. In practice, for example, a division of the valve into three stages is already sufficient. In *fig. 3* such a three-stage gas-filled valve is shown diagrammatically, while *fig. 4* shows two photographs of it.

We have just spoken of the assumption of a homogeneous field and a uniform distribution of the voltage over the electrodes. In the construction according to fig. 3 the field is not, of course, entirely homogeneous; the intermediate electrodes cannot be constructed as parallel plane plates but should have holes in order to allow the passage of the electrons from the cathode to the anode. In practice they are constructed as cylinders. The greatest field strength now depends also on the shape of these cylinders and it is possible to influence the chance of back-lash, for instance, by the choice of the size of the opening of the cylinders. In order to guarantee a sufficiently uniform distribution of the voltage over the different stages the impedance between all successive intermediate electrodes should be made as nearly equal as possible. To that end condensers of 80 μF are connected in parallel with the intermediate spaces, which condensers have the form of rings lying around the rectifying valve, as may be seen in the photograph of fig. 4*b*.

Let us now turn to the gas with which the valve is filled. A rare gas cannot be used because of the well-known phenomenon of gradual disappearance of the gas during operation of the valve at high voltage: it is taken up in the cathode and walls; the pressure in the valve falls [5]). Therefore the discharge is made to take place in saturated mercury vapour: a drop of mercury is placed in the valve and from this the mercury vapour which disappears is always supplemented by evaporation. The result, however, is that the vapour pressure in the valve depends upon the temperature of the surroundings,

[5]) The velocity of this process (called clean-up) increases rapidly with increasing voltage. It is not the same for different gases; for xenon, for instance, it is smaller than for argon.



a

b

Fig. 4. A Philips three-stage gas-filled rectifier valve for 125 kV; *a*) assembled and *b*) taken apart. The glass tube is surrounded by ring-shaped "Philite" condensers (see C in fig. 3). The tube is calculated for oil immersion of the rectifier. The length is more than 20 cm, the diameter of the condensers is almost 9 cm. When used in air, in order to avoid flashover along the outside of the glass wall, a greater insulation distance between the electrodes would be necessary, thus in general a longer length of tube.

see *fig. 5*. Gas-filled valves can therefore only be used in a limited temperature range, which in the case of the Philips high-voltage rectifiers for X-ray purposes lies between about 17 and 40 °C. At too low a temperature the pressure becomes so low that no

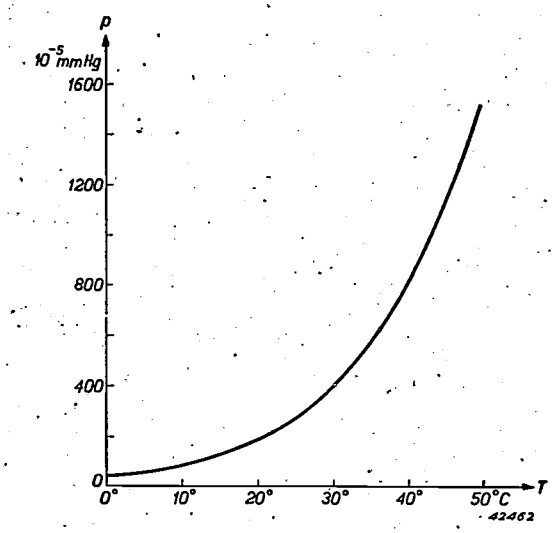


Fig. 5. Vapour pressure of mercury as a function of the temperature.

ignition occurs in the direction of transmission (the chance of formation of gaseous ions is too small); at too high a temperature the pressure becomes so high that breakdown occurs in the negative phase (the valve no longer rectifies). The latter can be seen directly from fig. 2: with increasing pressure the work point B is displaced horizontally to the right and after a certain pressure has been reached it comes to lie above the breakdown curve (unless B lies lower than the minimum of the curve, but in practice such a far-reaching subdivision of the voltage is not possible). The liquid mercury that has to be present in the valve should not come into contact with the anode or the intermediate electrodes, since local increases of the field strength would then occur and these might cause cold emission and thereby back-lash. Therefore the drop of mercury is placed in a special compartment, separated from the rest of the valve by a capillary (see fig. 3). Due to its surface tension the drop cannot flow through the capillary, while the mercury vapour is admitted freely to the valve.

Although, according to the above, we arrive at a distance between neighbouring electrodes in the gas-filled valves which is not greater than that in vacuum valves, it is, nevertheless, possible to employ an oxide cathode in gas-filled valves. Since there is no trouble with space charge here the cathode can be mounted behind a screen (the ring R in fig. 3) which prevents any particles shot off the cathode from reaching the neighbouring electrode. At the same time the screen prevents an ion bombardment of the hot cathode and the accompanying sputtering. As already stated, an oxide cathode requires a cathode power of only 8 W for the largest peak currents occurring; thereby the cathode temperature is about 900 °C. At this temperature it is found that even after some time no disturbing evaporation of barium from the oxide cathode occurs, although, remarkably enough, the vapour pressure of pure barium is quite considerable at the temperature mentioned.

Improved vacuum rectifiers

Let us return for a moment to the vacuum valves. The construction according to fig. 1 was not practicable for modern diagnostic apparatus. It proved, however, to be capable of improvement in various respects. An important improvement was the alteration of the construction as shown in *fig. 6*.

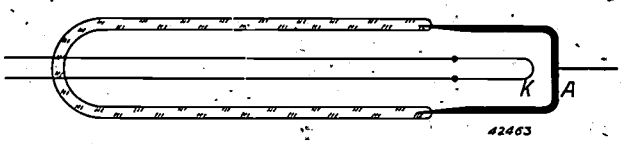


Fig. 6. The characteristic of the vacuum rectifiers can be considerably improved by the use of a cup-shaped anode A. In this case it forms part of the valve wall ("Metalix" rectifiers) which makes cooling easier.

Here the anode is not a flat plate at some distance from the cathode, but a cup surrounding the cathode. By this means the electrons having to pass from the filament to the anode are spread over a larger solid angle, the current density becomes smaller and therefore the space charge has less effect. In *fig. 7* the characteristic of a valve with

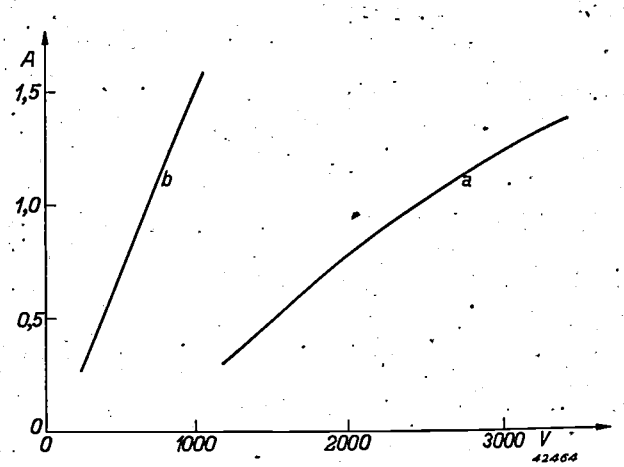


Fig. 7. Characteristics of vacuum rectifiers at high temperatures of the filament (more than 2350 °C): *a*) according to the old construction of fig. 1, *b*) according to the improved construction of fig. 6.

cup-shaped anode is compared with that of a valve with a plane anode. In order to transmit a current of 1 A the improved valve needs a voltage of only 650 V, compared with 2500 V in the case of the old valve. At the same time the cup shape of the anode, which in this case forms part of the wall of the valve, practically entirely prevents secondary electrons from striking the glass wall and there is no longer any fear of a grid action of varying wall charges.

A practical model of such a valve, the "Metalix" rectifier, to be used in air, is shown in *fig. 8*. These valves attracted considerable attention and they may still be found in use in existing diagnostic apparatus. Their further development, however, was curtailed by the introduction of the gas-filled valve. We have seen, indeed, that gas-filled valves

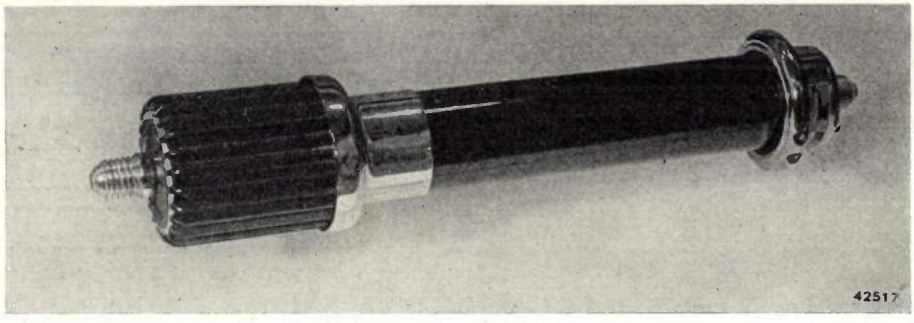


Fig. 8. "Metalix" vacuum rectifier valve. Since this valve is not intended for immersion in oil in the generator container, but for use in air, the envelope is made of non-transparent glass, in order not to give a disturbing light in the X-ray room. The anode is provided with cooling fins. The total length is 55 cm.

are able to satisfy the highest requirements in diagnostics. The objection to gas-filled valves, namely the limitation of their usefulness to the narrow temperature range of 17-40 °C, makes gas-filled valves unsuitable for use in tropical climates or in apparatus for the macroscopic testing of materials, which have to be used under very divergent conditions of temperature. Thus in any case vacuum rectifiers, which are independent of temperature, still command a certain field of application, and for the sake of having a single type of rectifier of universal applicability further development of vacuum rectifiers, *viz.* a still better approximation of the requirements made by diagnostics, certainly seemed worthwhile.

In recent years several improvements have proved possible by which vacuum rectifiers, which had remained at the stage of development shown in fig. 6, could be given a new lease of life. Probably the most important improvement is the employment of thoriated tungsten for the filament. The addition of thorium to the tungsten results in a considerable increase in the electron emission, so that for the same tube current a lower filament temperature is sufficient and a smaller filament power is needed [6]). In practice a temperature of about 1750 °C is used. At this temperature the evaporation of the filament is so slight — although here again the vapour pressure of the pure thorium at the temperature in question is by no means low — that even after several thousand hours of operation no appreciable quantity of evaporated material can be detected as a deposit on the valve wall, not to speak of the occurrence of any burning through of the wire. The resulting, much lower, sensitivity to fluctuations of the heating voltage makes the above-described complications (increase of voltage for each exposure, stabilizer) unnecessary. The filament power is reduced from 125—150 W for the pure tungsten cathode to somewhat more than 30 W for the cathode of thoriated tungsten.

The thoriated tungsten filaments are especially sensitive with respect to the vacuum: traces of gas in the valve may reduce the electron emission considerably. Therefore the use of these cathodes is only possible by taking more care for a very good vacuum than in the case of the old vacuum valves. During manufacture the valves are very carefully evacuated. During operation the high vacuum is maintained with the help of a barium getter.

Another modification introduced in the construction of the "Metalix" rectifier valves of fig. 6, which made them more suitable for modern diagnostic apparatus, may be seen in *fig. 9*. The cup shape of the anode and thus the favourable characteristic

[6]) Because of the reduction in the filament power required, thoriated tungsten cathodes are also used in some modern transmitting valves. See E. G. Dorgelo, Several technical problems in the development of a new series of transmitting valves, Philips Techn. Rev. 6, 253, 1941.

of fig. 7*b* is retained, but now the anode no longer forms part of the valve wall. By means of a thin pin of material with a low heat conductivity (molybdenum) it is connected with a metal cap welded to a hard-glass envelope and serving for

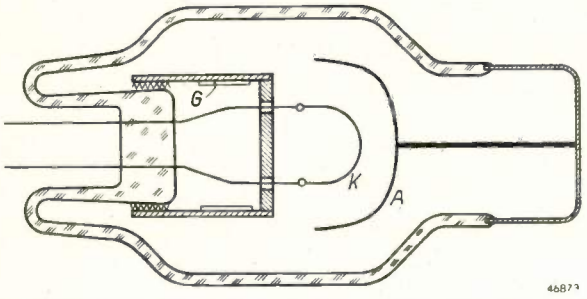


Fig. 9. Diagram of a modern vacuum rectifier. The filament *K* is made of thoriated tungsten. The cup-shaped anode *A* no longer forms a part of the (hard-glass) valve wall, which makes it easier to use the valve under oil. *G* is a getter (barium mirror).

the current supply. Thanks to this construction the anode can be heated to a higher temperature during degassing, while at the same time the field strengths along the glass are lower. By this means, and because of the fact that the metal cap in question remains relatively cold, it becomes easier — as in the case of the above-described gas-filled valves — to make the valves suitable for operation under oil. This leads to an important reduction in dimensions. The valve for 125 kV now has a length of somewhat more than 20 cm with a diameter of 9 cm (see *fig. 10*), *i.e.* about the same dimensions as the three-stage gas-filled valve.

Thanks to the cup shape of the anode in the new construction, practically no secondary electrons can strike the glass wall, and if this does happen the screening of the cathode is an adequate insurance against the occurrence of grid action. For the dissipation of the heat developed on the anode, which in the construction according to fig. 8 took place easily by conduction and convection into the surrounding air, in the new construction use is made of radiation. Because of the large radiating surface of the cup anode, an anode dissipation of 40 to 50 W, as occurs in continuous operation (about 30 mA), offers no difficulties at all, and a peak current of 1 A, at which the anode dissipation amounts to about 600 W, can easily be coped with for a few seconds, since the anode possesses a sufficiently generous heat capacity.

Summarizing, we may say that vacuum rectifier valves in their latest forms give very interesting results and are already beginning to conquer considerable territory. It is possible that ultimately they will again entirely replace gas-filled rectifier valves.

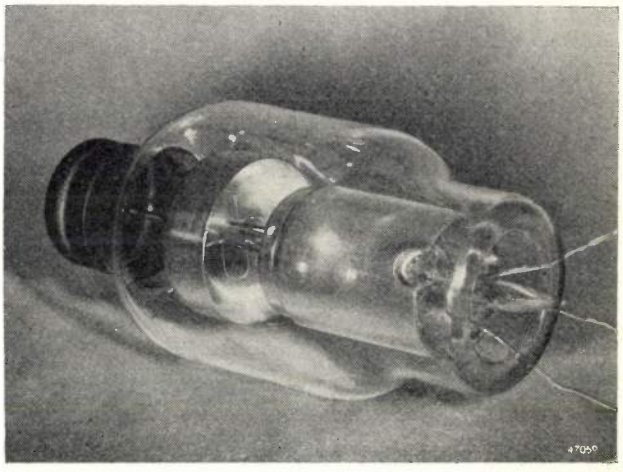


Fig. 10. Modern vacuum rectifier for 125 kV with cup-shaped anode and thoriated tungsten cathode. The valve, more than 20 cm long, is included to be used under oil.

CARRIER TELEGRAPHY

by J. te WINKEL.

621.394.44

In order to improve the economy of telegraphic communication the aim used to be to increase the signalling speed by the construction of special apparatus or by employing devices for the transmission of several telegrams simultaneously over a single pair of conductors. The most modern form of the latter method is voice-frequency or carrier telegraphy, whereby alternating voltages of different frequencies (carriers) are each modulated in the rhythm of the signals of a telegram. After a discussion of the principles of this method an installation for carrier telegraphy is described in which 18 or 9 telegraphy channels, as the case may be, are contained in the frequency band of a single telephone connection, *i.e.* for example in a single channel of a carrier telephony system. Special attention is paid to the circuits employed for modulation of the telegraph signals on their carrier and those for their demodulation.

For many decades telegraphy, the oldest form of electrical communication, was the only available method for the transmission of messages over long distances. When later on, thanks to loading coils and repeaters, telephony could be used for covering large distances it took over a large part of the field of application of telegraphy. The advantage of being able to obtain an answer immediately, as well as that of personal contact, made for many purposes telephony more attractive than telegraphy. As a result the technique of electrical communication has recently been concentrated mainly on the perfecting of the telephone system, which has now attained a high standard of reliability and economy. Nevertheless, telegraphy has always retained its right of existence, especially in cases where it was desirable that the messages transmitted should be recorded in writing, and in general also because telegraphy is more economical and therefore cheaper to the public than telephony. It is remarkable that even the technical development of apparatus for telephony has now led to a new and especially economical telegraph system and thereby served to extend the field of telegraphy. This system, known as voice-frequency or carrier telegraphy, is based on the same principles and for the most part employs the same kind of apparatus as carrier telephony, which has been discussed in a series of articles in this periodical [1]).

Before proceeding to the description of a carrier telegraph system, such as has been developed and put into practical use by Philips, we shall give as an introduction a brief survey of the course of events in ordinary telegraphy, so-called direct-current telegraphy. This will give us the opportunity of mentioning and explaining several elementary concepts and quantities of telegraphy.

[1]) Philips Techn. Rev. 6, 325, 1941; 7, 83, 104 and 184, 1942, 8, 137 and 168, 1946.

Direct-current telegraphy

Fig. 1 shows the circuit diagram of a direct current telegraph connection. At the transmitting station A by means of the switch S, for instance a signalling key, positive or negative voltages can be applied to the line. At the receiving station B the line is completed by a polar relay R. This relay reverses its armature as soon as the current through the relay winding changes its direction, the position of the contact arm T thus changing in the same way

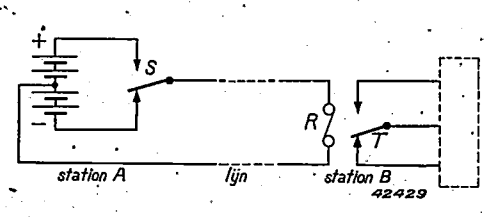


Fig. 1. Diagram of a connection between two stations A and B with direct-current telegraphy. S signalling key, R polar relay with contact arm T.

as the position of the switch S. *Via* the contact T some type of recording apparatus is operated from a source of voltage not shown in the diagram.

Signalling speed

For telegraphing over such a circuit a code is used: each letter is represented by a succession of a certain number of positive and negative voltage impulses of a certain length. The oldest and still commonly used code is the Morse code; for modern recording apparatus, for instance teleprinter apparatus, which are so constructed that each succession of impulses leads immediately to the printing of a certain letter, other codes are used. All the codes, however, have this in common, that the length of each impulse is chosen equal to a whole multiple of a fixed unit, the elementary length τ; see for example *fig. 2a*.

The speed at which one can telegraph is greater

the shorter the time devoted to the elementary signal length. The speed is expressed in baud units (from Baudot, one of the pioneers of telegraphy), *i.e.* the number of elementary lengths transmitted per second.

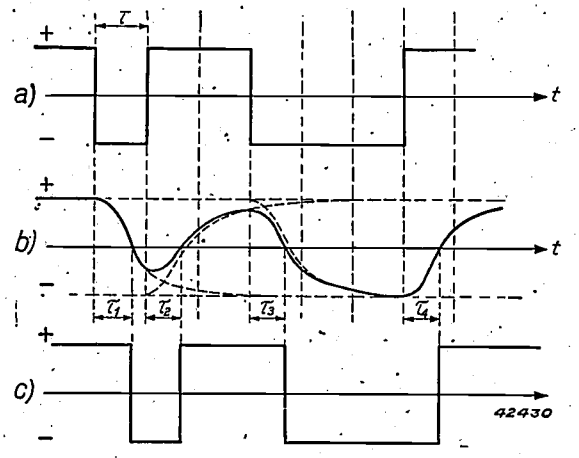


Fig. 2: Shape of the voltage curve at the transmitting end (*a*) and the current curve at the receiving end (*b*). At the zero points of the latter curve the voltage in the receiving apparatus is reversed in polarity by the relay contact T (*c*). Due to the fact that the "propagation times" $\tau_1 \ldots \tau_0$ of the voltage changes differ more or less, *c* is not entirely identical with *a*: distortion has occurred.

The signalling speed is limited by properties of the line, which will be discussed later, and by the inertia of the transmitting and recording apparatus. For teleprinters, which are at present the most common, the signalling speed has been standardized by the C.C.I.T. (Comité Consultatif International Télégraphique) at 50 baud.

Signal distortion

The two conductors of the line possess a certain resistance and capacity which are distributed uniformly over the whole length of the connection. There exists also a certain self-induction and dissipation, which, however, can usually be disregarded. Every time the switch S in fig. 1 is reversed the line capacity is discharged and charged again with opposite polarity. The resistance of the line causes this charging and discharging to occupy some time; thus it will take some time after reversing the key at the transmitting end before the current through the relay at the receiving end reaches its final value. If we know the behaviour of this current for a single change of voltage, we can construct the curve of the current for each letter sign. In *fig. 2b* this has been done for the signal of fig. 2*a*. It may be seen that the current curve involves considerable distortion, namely a rounding off of all transitions. In actual practice it is not a question of the shape of the whole curve, but only of the position of the zero points. As soon as the current in the receiving relay passes through zero the contact T is reversed, and for good transmission of the signals it is sufficient when the contact T follows the key S in the correct rhythm. In *fig. 2c* the position of the contact T is shown as a function of the time for a relay current according to fig. 2*b*. If we compare this position with that of the switch S, it is found that each change of voltage possesses, as it were, a certain time of propagation (τ_1 to τ_4), the value of which is not the same for all changes. The result is that the length of the impulses which are fed to the recording apparatus is no longer exactly equal to the length of the impulses transmitted; the signals are distorted. Signal distortion is defined as the increase or decrease in the length of an impulse expressed as a percentage of the elementary signal length τ, thus

$$100\,\frac{\tau_2-\tau_1}{\tau},\ 100\,\frac{\tau_3-\tau_2}{\tau} \text{ and } 100\,\frac{\tau_4-\tau_3}{\tau}\ \%.$$

It is clear that signal distortion can also occur due to other causes, for example inequality of the positive and negative voltages, incorrect adjustment of the relay or accidental disturbances. The maximum distortion occurring is a measure of the quality of the connection. It may not exceed a certain value, for instance 10 to 20 percent, depending on the type of telegraph apparatus and the margin of safety chosen, so as not to incur the danger that impulses will be omitted and incorrect letters printed.

Economy of the connection and maximum distance to be covered

In order to use a given telegraph line as economically as possible it is desirable to raise the signalling speed as high as possible. For this purpose various systems of high-speed telegraphy have been designed and constructed in the past. From the above, however, it is seen that with the signalling speed ($1/\tau$) the signal distortion will increase. Furthermore the signal distortion (differences in the times of propagation $\tau_1, \tau_2, \ldots$) will be greater, the greater the average time of propagation of the voltage changes. This time is determined by the RC time of the cable, and since R and C are both proportional to the length of the cable it increases proportional to the square of the distance to be covered.

It may therefore be seen that with a given limit for the permissible signal distortion, which limit one can try to place as high as possible by special circuits and constructions, a high signalling speed

leads to a limitation of the distance to be covered and, conversely, a given distance leads to a limitation of the signalling speed.

A different way of using a cable more economically was opened by systems which permitted several telegrams to be sent over the line simultaneously, each at a normal signalling speed. The most modern and at present the most important of these systems is the so-called voice-frequency or carrier telegraphy, which we shall now consider. It must, however, be noted at once, that here, too, is a limitation of the total signalling speed, a limitation which upon closer consideration proves to be based upon the same fundamental principles as in the case of direct-current telegraphy.

Voice-frequency telegraphy

In voice-frequency telegraphy the direct-current impulses are replaced by impulses of an alternating voltage of several hundred or thousand cycles per second. *Fig. 3* shows a simple circuit with which

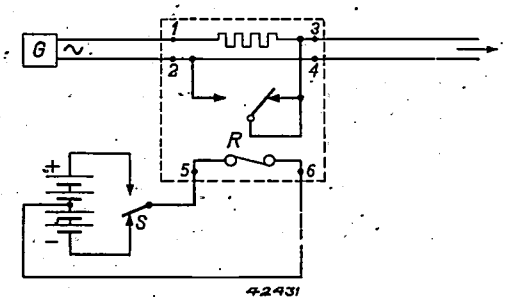


Fig. 3. Simple circuits for passing over from D.C. telegraphy to carrier telegraphy. C_1 generator of an A.C. voltage of several hundred or thousand c/sec, S signalling key, R polar relay which short-circuits or passes the A.C. voltage to be applied to the line in the rhythm of the signals given with S.

signals can be transmitted in this way, while *fig. 4* shows the appearance of a certain signal code in direct-current and in voice-frequency telegraphy. We may consider the alternating voltage as a carrier wave modulated in each case with the telegraph signals to be transmitted according to *fig. 4a* (the "depth of modulation" here is 100 percent). By using several carrier waves with different

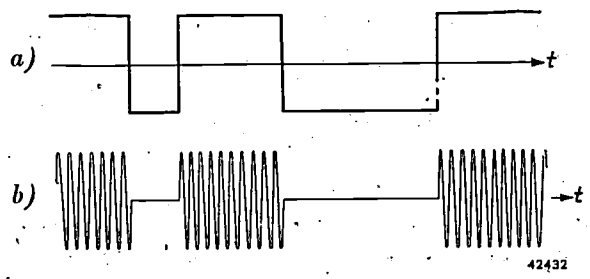


Fig. 4. The same signal code (succession of impulses) in D.C. and in carrier telegraphy.

frequencies, each of which is modulated separately with different signals, a number of telegrams can be sent simultaneously over the same pair of conductors. At the receiving end a series of selective receiving apparatus has to be provided each of which reacts only to one of the carriers.

Frequency band of a telegraph channel

Let us consider a signal in direct-current telegraphy which consists of alternating positive and negative impulses of the unit length τ, see *fig. 5*.

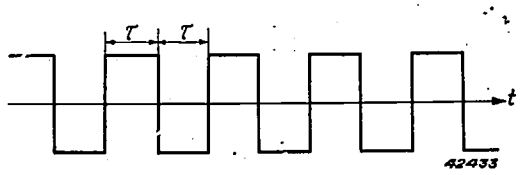


Fig. 5. Succession of positive and negative impulses with the unit length τ.

Such a signal may be conceived as an alternating voltage composed of a series of harmonics. At a signalling speed of 50 baud the fundamental frequency of this alternating voltage is 25 c/sec while further a large number (theroretically an infinite number) of higher harmonics occur. We may likewise assign to any given signal other than that of fig. 5 a complete frequency spectrum, beginning in the most general case with the frequency zero.

The distortion incurred by such a signal in being propagated over a telegraph cable can now also be described in a different way from that given in the foregoing. Due to its capacity and resistance the pair of conductors possesses an attenuation for the currents transmitted, which depends upon the frequency and increases with increasing frequency, since the capacity of the conductors forms a shunt whose impedance decreases with increasing frequency: the cable may be compared to a low-pass filter. In *fig. 6* the curve (attenuation loss *versus* frequency characteristic) of a normal telegraph cable is shown. A signal composed of a number of frequencies is now not only attenuated as a whole, but it is, moreover, distorted due to the fact that the higher harmonics are attenuated more than the

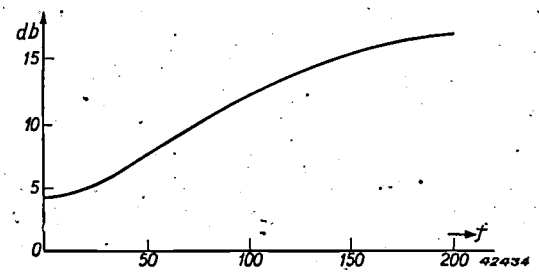


Fig. 6. Attenuation (in db) of a telegraph cable as a function of the frequency f in c/sec.

lower harmonics and the fundamental frequency. The effect of this prejudice to the higher frequencies is a rounding off of jumps and angles in the current curve, as we have seen in figs. 2*a* and *b*.

The relation between distortion, cable length and signalling speed can also easily be explained qualitatively in this way.

Let us now apply these considerations to voice-frequency telegraphy. When a signal with frequency q is modulated on a carrier with frequency p the two side-band frequencies $p \pm q$ occur. Thus in the modulation of a carrier by the telegraph signals two side-bands are formed, one on either side of the carrier frequency, which bands theoretically extend to infinity. These side-bands, however, need not be transmitted in their entirety: all the frequencies farther than a certain distance q_m from the carrier may be suppressed, for instance by means of a band-pass filter. This resolves itself in fact to the suppression of the higher harmonics in the signal, analogous to the unfavourable effect on those harmonics exerted by the cable in direct-current telegraphy. The result is again a distortion (rounding off) of the impulses transmitted, so that the modulated carrier assumes for instance the appearance of *fig. 7*. The degree of distortion per-

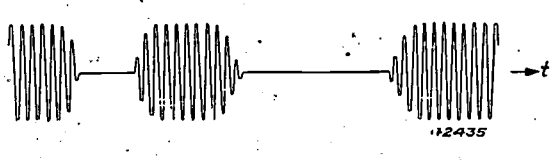


Fig. 7. The same signal code as in fig. 4*b* distorted as a consequence of the cutting off of the harmonics above 40 c/sec by the band-pass filter.

missible for reliable functioning of the receiving apparatus determines the minimum width ($2\ q_m$) which the transmitted frequency band must possess. It has been found that for a signalling speed of 50 baud in general a band width of $2\ q_m = 80$ c/sec is sufficient.

Several telegraph channels on one pair of conductors

If we wish to send several telegrams on different carriers simultaneously over a line without mutual interference we must assign to the carriers frequencies which lie sufficiently far apart, namely at least 80 c/sec, and use for each carrier a band pass filter with the correct transmitting band 80 c/sec in width. At the receiving end the respective carriers with their side-bands are split up again by means of a series of band-pass filters in parallel, identical with the transmitting filters. Following each receiving filter circuits are provided for the demodulation of the carrier, as well as a recording apparatus.

How many of these telegraph channels can now be included in a given frequency band of width Δf? If the loss of the band-pass filters at the boundaries of their transmitting region could jump abruptly to a sufficiently high value (at least 30 db), the distance between two adjacent carriers would not need to amount to more than 80 c/sec. Actually a filter always has a certain transition region between transmitting and attenuating regions, which transition can only be made narrower by a more expensive construction of the filter. In practice for example band-pass filters with the attenuation curve reproduced in *fig. 8* can be used. The neces-

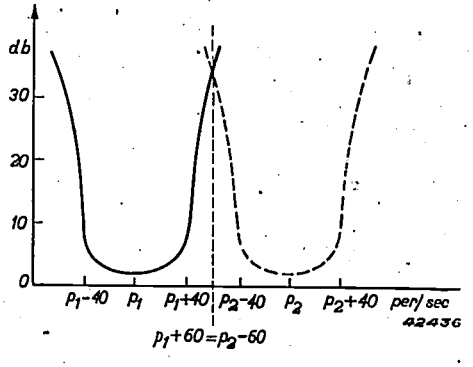


Fig. 8. Attenuation curve of a band-pass filter for the telegraph channel with carrier p_1. The broken line gives the corresponding curve for the adjacent channel with carrier p_1.

sary attenuation is here attained at a distance of 60 c/sec from the carrier, so that a carrier interval of 120 c/sec is required.

A comparison with carrier telephony as previously described shows that in that case — where the width of a channel is much greater (4000 c/sec) — the carrier itself and one of the sidebands are suppressed at the transmitting end. Due to the finite width of the transition region of the filters used, it is practically inevitable that the low frequency part of the transmitted side-band will also be cut off. In telephony this is no objection since the speech frequencies below 300 c/sec are not necessary for an intelligible conversation. In telegraphy, however, where all the modulation frequencies from zero onward have to be transmitted, it is not possible in practice to filter out only one side-band. For this reason the carrier and both side-bands are transmitted.

In a frequency band of width Δf, therefore, $\Delta f/120$ telegraph channels can be laid, so that one may speak of a "total" signalling speed of 50 $\Delta f/120$ baud. An obvious question is whether further profit can be gained by making the signalling speed in each channel greater than 50 baud. In the case of any given telegraph signal, however, all the frequencies of the composite alternating voltages are

proportional to the signalling speed, thus the required frequency band for a channel also exceeds 80 c/sec, and this proportionnally to the signalling speed. If band filters with an attenuation curve as in fig. 8 are always used, where only the frequency scale changes, so that the ratio between the width of transition region and transmission region remains the same, the total signalling speed is independent of the signalling speed in each channel. In other words, the total amount of "information" which can be transmitted is constant. It is of course true that with a greater band width of the filters the transition region can be made relatively narrower without much trouble, so that the signalling speed could be increased somewhat more. This would, however, by no means compensate for the complications and increase in cost which would accompany a higher speed of the transmitting and recording apparatus.

The attainment of a given total signalling speed therefore implies the use of a frequency band of a certain width, and how far one may go with this depends upon the circumstances. For example, if one telegraphs *via* a radio connection, the available frequency region is given by the band width of the radio transmitter. In the case of telegraphy *via* cables also an economic limit is set to the frequency region, since with frequency regions of any great size and with long distances, equalization of the cable attenuation [2]) and amplification (or relaying) of the signals becomes necessary, which considerably increase the cost with increasing width of the frequency region. The situation here is analogous to that in carrier telephony, only in that case these considerations weigh even more heavily owing to the much higher modulation frequencies used in telephony.

Carrier telegraphy in a telephone channel

It is an obvious and much used method in cable telegraphy to make the total frequency band equal to that of an ordinary telephone connection (300-2600 c/sec). In this region, with a band width of 120 c/sec (and beginning at 360 c/sec, see below), no fewer than 18 telegraphy channels can be included.

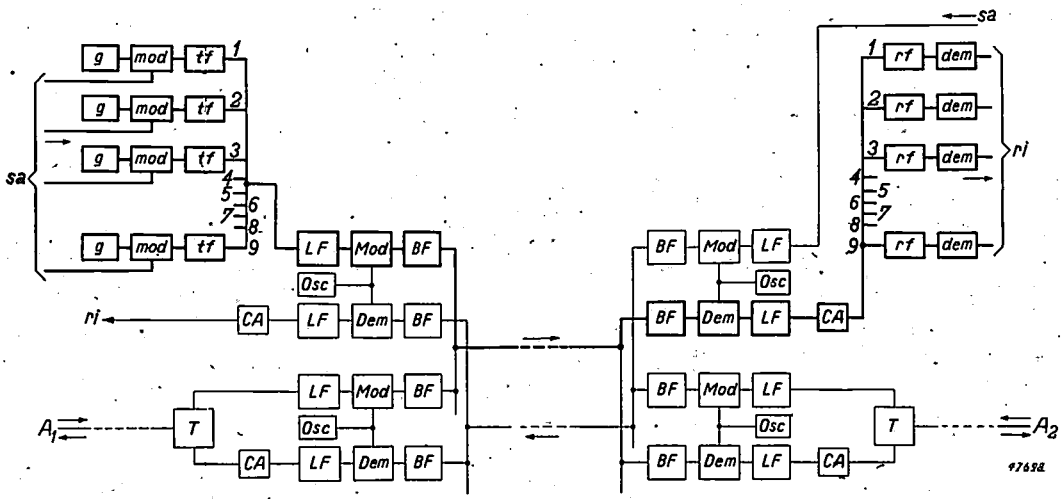


Fig. 9. Block diagram of a carrier telegraph connection using one channel of an installation for carrier telephony. Two channels of the latter are drawn, one for the carrier telegraphy (4 channels of which in one direction are indicated with heavy lines), the second for a normal telephone connection between two subscribers A_1 and A_2. The small letters refer to the telegraphy apparatus: *s.a* signalling apparatus, *r.i.* recording instruments, *g* generators, *mod* modulators, *tf* transmitting filters, *rf* receiving filters, *dem.* demodulators. The capital letters denote the components of the carrier telephone installation: *T* four-wire terminals, *LF* low-pass filters, *Mod* modulators, *BF* filters, *Dem* demodulators, *CA* channel amplifiers, *Osc* oscillators for the addition of the telephone carrier.

These 18 telegraphy channels together can be transmitted in exactly the same way as a normal telephone conversation.

An especially economical telegraph connection is obtained when it is possible to reserve for carrier telegraphy one telephone channel of a carrier telephone installation with for instance 12 or more channels. The group of 18 telegraph carriers, each modulated with its own telegraph signals, is then modulated as a whole on one of the carriers of the telephone apparatus, as is indicated in the block diagram of *fig. 9*.

[2]) Philips Techn. Rev. 6, and 7, 293, 1941 184, 1942.

Since the large number of 18 telegraph channels will often be unnecessary, or since perhaps two instead of one of the many telephone channels in a given connection can be made available for telegraphy, Philips several years ago designed an apparatus whereby only 9 instead of 18 telegraph channels lie in the frequency band in question. This gives the advantage that the carrier interval may now amount to 240 c/sec, so that the band pass filters may have much wider transition regions and may therefore be simpler and cheaper.

As carrier frequencies in the frequency region mentioned from 300 to 2600 c/sec the values recommended by the C.C.I.T. may be taken: 420, 540, 660, etc. to 2460 c/sec, *i.e.* odd multiples of 60 c/sec. The odd multiples have the advantage over the even multiples that second harmonics and sum and difference frequencies of the carriers, which frequencies occur due to non-linear distortion in amplifiers etc., fall just between the carriers where the attenuation of the band-pass filters is high.

Apparatus for carrier telegraphy

The apparatus for carrier telegraphy, as we have seen, contains for each channel a carrier generator, a modulator and a filter for transmission, as well as a filter and a demodulator for reception. A series of other elements is also needed, such as relays, regulators, meters for adjusting the currents, etc. We shall consider here only two components of the apparatus in somewhat more detail, namely the modulator and the demodulator.

Modulator

For the modulation of the telegraph signals on the carrier, instead of the relay used in fig. 3 a device can be used which contains no moving parts and thus has the advantage that neither maintenance nor adjustment is necessary. This device, the telegraph modulator, is practically identical with the modulators used in carrier telephony which have already been discussed in this periodical [3]). The diagram illustrating the principle of the telegraph modulator ("blocker") was also described there, and the circuit diagram is given once more in *fig. 10a*. To the terminals *1-2* the voice-frequency alternating voltage is applied which serves as carrier, to the terminals *5-6* the direct current impulses of the telegraph signals. The resistance of the selenium rectifiers s_1, s_2 for the alternating current depends upon the polarity of the direct voltage acting on them at the same time. If point *6* is positive, the resistance in question is high; if point *5* is positive, it is low. The alternating voltage at the output terminals *3-4* will thus alternate in intensity according to the rhythm in which the voltage on *5-6* is reversed in polarity.

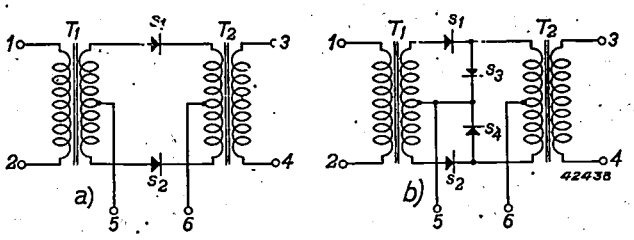


Fig. 10. Modulator connections. T_1, T_2 transformers, $s_1 \ldots s_4$ selenium rectifiers. At *1-2* the carrier is applied, at *5-6* the D.C. telegraphy signals to be transmitted, while at *3-4* the modulated signal is taken off. The modulator replaces in fig. 3 the circuit situated inside the block within the broken line, with the correspondingly numbered connections *1.* to *6* inclusive.

Fig. 10b shows a somewhat modified circuit which, as may easily be seen, has the property that the resistance of the direct-current path between the points *5-6* upon reversing the polarity remains practically constant; this is desirable because the signalling apparatus to be connected at *5-6* is in general constructed for a specific external resistance.

In carrier telephony the audio-frequency microphone currents are applied to the terminals *1-2* of fig. 10*a*, the high-frequency carrier to *5-6*. In our case, on the other hand, the (audio-frequency) "carrier" is applied to *1-2*, while the much more slowly changing D.C. signals are applied to *5-6*. The reason for this difference lies in the fact that in telephony the carrier itself need not be transmitted; this is accomplished by balanced connection with respect to points *5-6*. In telegraphy the carrier has to be transmitted; moreover, the low-frequency telegraph signals, if they were applied to *1-2*, would not be passed on by the transformer T_1.

Demodulator

At the receiving end of the telegraph system the telegraph carriers, after being split up by the receiving band-pass filters, must each be demodulated separately in order to reproduce the telegraph signals. Whereas for the demodulation in carrier telephone channels, containing only one side-band without carrier, the carrier itself must be added again [4]), for the demodulation of the telegraph channels this is not necessary. These channels contain the carrier and both side-bands, so that the demodulation resolves itself into a simple detection, as is employed for instance in radio receiving sets.

[3]) Philips Techn. Rev. 7, 83, 1942.

[4]) See for example Philips Techn. Rev. 8, 137, 1946.

A circuit suitable for this purpose, working with anode detection, is shown in *fig. 11a*. The amplifier valve B has such a large negative grid bias that in the absence of alternating voltage on the grid the anode current is exactly zero (class B adjustment). Now suppose that the telegraph signal shown in fig. 5 is being transmitted. It consists of positive and negative impulses of the unit length τ. An alternating voltage is applied to the grid as shown in *fig. 11b* (left, below), whereby the envelope of the signal, distorted in a certain way, is assumed for the sake of simplicity to consist of straight lines.

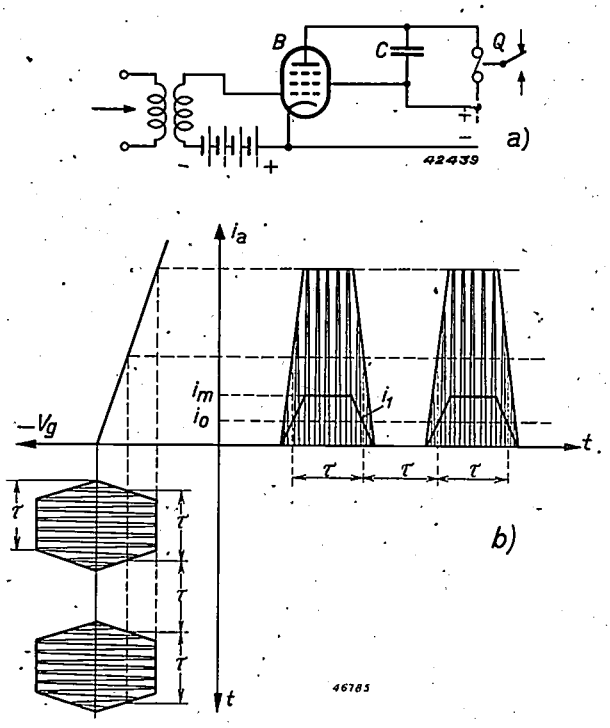


Fig. 11. *a*) Simple connections for the demodulation (detection) of the carrier telegraph signals. The valve B works in class B adjustment.
b) Form of the grid A.C. voltage V_g and the anode current i_a in the case of a signal consisting of impulses of the unit length τ. The low-frequency component i_1 of the anode current flows through the relay Q. This is so adjusted that it reverses its armature at a current $i_0 = i_m/2$.

If we also assume a straight line for the valve characteristic, the anode current i_a takes the form shown in the figure. It can be split up into a low-frequency component i_1, which flows through the receiving relay Q, and a component with carrier frequency, which flows through the condenser C, connected in parallel with the relay. The low-frequency component i_1 is of similar shape as that of the envelope of the anode current (namely divided by a factor π). It may be seen from the figure that the original telegraph signal is reproduced undistorted provided the relay Q is so adjusted that it reverses its armature at a relay current i_0 amounting to one half of the maximum value i_m. (In case the envelope should not be straight, a different but in any case very definite fraction would take its place.) If Q is a polar relay this adjustment can be realized by sending a constant direct current through an auxiliary winding of the relay in such a way that the field of the current i_0 is just cancelled. A simpler solution and one therefore actually employed in our apparatus is to take for Q a normal (non-polar) relay and to adjust the tension of the spring holding the armature in position to the correct value.

Compensation of changes in level

The detection circuit described only functions free of distortion as long as the level of the carrier received remains constant. Actually that level fluctuates more or less, due for example to the fact that the cable attenuation varies with the temperature. If the normal level is such that the low-frequency anode current i_1 flowing through the receiving relay has the shape $ABCD$ in *fig. 12*,

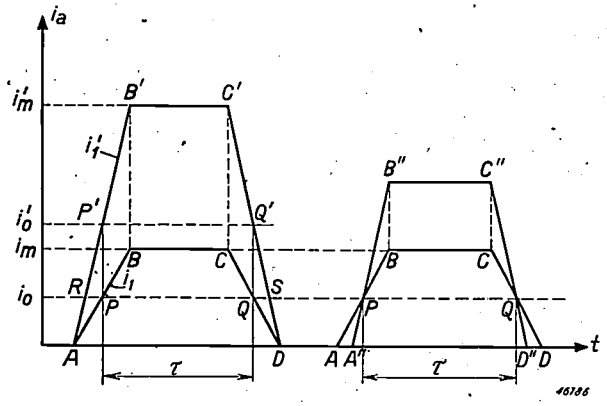


Fig. 12. If the low-frequency component i_1 of the anode current increases to i_1', the impulse received is too long, namely RS instead of PQ. This distortion can be compensated for by decreasing i_1' by a direct current $i_0'-i_0 = (i_m'-i_m)/2$.

at a higher level for example the shape AB^1C^1D will occur. The signal received is now too long, namely RS instead of PQ. A signal distortion thus occurs to the amount of $100\,\dfrac{RS-PQ}{PQ}$ %.

This distortion would be cancelled if the line AB^1C^1D could be lowered by the amount $PP^1 = {}^1/_2\,BB^1$. The shape of the current curve would then be $A''B''C''D''$, which intersects the original line $ABCD$ at the height i_0, thus at the correct points P and Q.

The signal-frequency anode current i_1' must thus be decreased by a direct current i_0'-i_0 equal to one half the increase of its maximum value (i_m'-i_m). A circuit diagram which accomplishes this is shown

schematically in *fig. 13a* [5]). In the grid circuit of the amplifier valve an extra rectifying circuit is introduced, consisting of the selenium rectifier *s* and the resistance R_1, shunted by the condenser C_1. The valve is connected between a tap *1* at the middle of the input transformer and a certain tap *3* on the grid voltage battery. In the non-signalling state (no grid A.C. voltage) a D.C. voltage V_2 therefore acts on the valve in the blocking direction. When a certain A.C. voltage (carrier) whose amplitude is smaller than 2 V_2 acts on the grid, no current will ever flow through the selenium rectifier, since point *1* never becomes positive with respect to point *3*. If, however, due to an increase in the receiving level, the amplitude of the grid A.C. voltage becomes greater than $2V_2$, namely 2 $(V_2 + \Delta V)$, a current flows through the rectifier at the positive peaks. The condenser C_1 is thereby charged, point *2* becomes negative with respect to point *4* and this proceeds until *1* and *3* again have the same potential at the peaks of the grid A.C. voltage. This is the case at a condenser voltage of ΔV. The grid bias therefore automatically becomes more negative by the amount ΔV, *i.e.* half the increase of the carrier amplitude; see *fig. 13b*. It may immediately be seen from the figure that also the envelope of the anode current and thus also the signal-frequency component of the anode current is lowered by half the amount with which its maximum value would have increased in the absence of *s*, C_1 and R_1. This is exactly the shift which was desired.

If therefore the voltage V_2 tapped from the battery is made equal to half the peak value of the carrier at the lowest level occurring, no signal distortion occurs at higher levels. In this way level variations to 6 db above and below the normal level can be permitted. The accuracy of the regulation is only limited by the time constant R_1C_1 which must be made large enough for the extra grid voltage furnished by C_1 to be retained in the intervals between the impulses, *i.e.* where the carrier amplitude is zero for a time. In the sometimes long intervals between successive telegrams, thanks to a suitable connection of the transmitter the carrier is transmitted continuously (in conformity with the recommendation by the C.C.I.T.), so that the grid voltage is set at the correct value immediately a telegram starts.

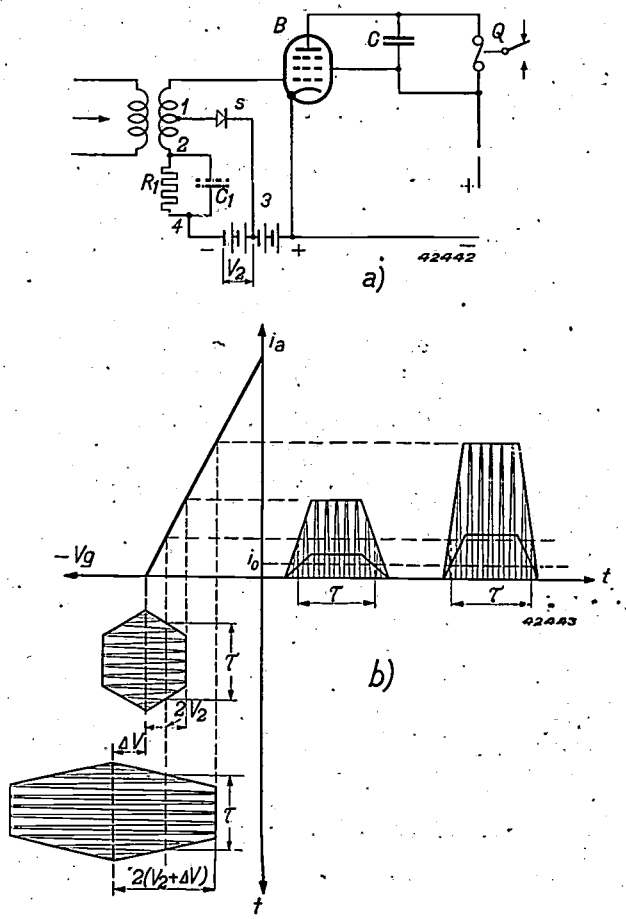


Fig. 13. *a*) Detection circuit with automatic compensation of the signal distortion caused by changes in level according to fig. 12; *s* is a selenium rectifier. Tap *3* of the grid bias is so chosen that $2V_2$ is equal to the amplitude of the grid A.C. voltage at the lowest receiving level occurring.

b) Explanation of the functioning of the circuit shown in the diagram. If the amplitude of the grid A.C. voltage rises from $2V_2$ to 2 $(V_2 + \Delta V)$, by the charging of the condenser C_1 the grid bias becomes more negative by the amount ΔV.

[5]) A similar compensation for changes in level is necessary in signalling in a telephone installation, see Philips Techn. Rev. 8, 168, 1946. The solution there described is somewhat simpler than the one given here. The difference is due to the fact that the impulses received in telephone signalling (dialling impulses) have much steeper fronts than the impulses in telegraphy. With a steeper front the distortions occurring and the required corrections are obviously smaller (*cf.* fig. 12).

VELOCITY-MODULATION VALVES

by F. M. PENNING.

621.385.83

When an ordinary radio valve is used as oscillator, below a certain wave length the efficiency rapidly approaches zero. This behaviour is caused partly by the transit time of the electrons. If one attempts to shorten the transit time by reducing the valve dimensions, the energy which the valve can deliver again decreases rapidly. In order to avoid this difficulty in recent years a different principle has been applied for the excitation of decimeter and centimeter waves, whereby the transit time of the electrons is used to advantage. The beam of fast electrons concentrated by a magnetic or an electric field is made to pass through a longitudinal alternating field which produces a velocity modulation of the beam. As a result electrons which left the source later can overtake those which left earlier and fluctuations of density begin to appear along the beam. The beam is then passed through the openings of a cavity resonator and later received on a large anode. As a consequence of the fluctuations in density electromagnetic oscillations are caused in the cavity resonator by induction. By back-coupling these oscillations to the part where the modulating alternating field is excited — also a cavity resonator or a Lecher system — an oscillator can be realized. In this article it is explained how the electron concentrations depend upon the time and the place in the "velocity-modulator valve", and how with several simplifying assumptions the efficiency can be approximately calculated from two characteristic quantities: the ratio ϱ of external resistance to valve resistance and the product of the number ξ of the concentrations or parcels of electrons in the velocity-modulating space and the back-coupling factor K. For the sake of clarity an example is discussed.

When more than a half a century ago Heinrich Hertz carried out his classical experiments on electromagnetic waves he used a wave length of about 30 cm. For various reasons radio technology developed first mainly in the direction of very long waves; only in recent years has there been greater interest in radio waves of a few decimeters down to a few centimeters.

When ordinary radio valves are used for the excitation of very short waves one encounters various difficulties, as has often been explained in this periodical, difficulties which are caused for example by the fact that the transit times of the electrons are no longer negligibly small compared with the oscillation time, as at low frequencies [1]). This causes a damping in the grid circuit, with the result that oscillations of too high a frequency can no longer be amplified and the valve can no longer oscillate above a certain frequency. Attempts may be made to meet this difficulty by reducing the dimensions, but this is naturally limited, and moreover small dimensions involve limitation to small energies.

Attempts have been made to turn these transit time effects, detrimental in ordinary oscillators, to advantage by changing the construction of the valve or its connections.

This can be done in several ways. The oldest known example is found in the valves according to Barkhausen and Kurz; another example is the magnetron, which was discussed in a previous article in this periodical [2]). In both cases the electrons swing back and forth in the valve. The distances and velocities are so chosen that on an average as much energy as possible is taken per period from the electrons by the alternating field, which energy then appears as energy of oscillation.

Use is made of transit times in a different way in the so-called velocity-modulation valves [3]), which are the subject of this article. In these valves the electrons do not swing back and forth but always move in the same direction. The modulation of the electron current is here produced at the entrance to the tube by periodic variation of the velocity, previously equal for all electrons, as a function of the time. This results in periodic current variations farther along in the tube, because when the electrons which started later have a higher velocity they overtake those which started earlier.

Finally there are oscillator valves in which use is made simultaneously of the velocity modulating

[1]) See for these transit times, for example, C. J. Bakker, Some characteristics of receiving valves in short-wave reception, Philips Techn. Rev. **1**, 171, 1936; M. J. O. Strutt and A. van der Ziel, The behaviour of amplifier valves at very high frequencies, Philips Techn. Rev. **3**, 103, 1938.

[2]) Philips Techn. Rev. **4**, 189, 1939.

[3]) The literature on velocity modulator valves has become very extensive in recent years. We mention here only the first articles by the pioneers in this field. W. C. Hahn and G. F. Metcalf, Proc. Inst. Radio Eng. **27**, 106, 1939; R. H. Varian and S. Varian, J. Applied Phys. **10**, 321, 1939; D. L. Webster, J. Applied Phys. **10**, 501 and 864, 1939.

effect just described and of the swinging back and forth of the electrons, which is characteristic of Barkhausen-Kurz valves. These oscillator valves are also very important in practice. An interesting example will shortly be discussed in this periodical.

Before we now pass on to the elucidation of the principles on which the functioning of velocity-modulating valves is based, we shall make a few general remarks about the way in which the high-frequency energy of the electrons is transmitted to the oscillation circuit.

In ordinary transmitting valves the electrons finally strike an anode forming part of the oscillation circuit, which with the small dimensions of the valves results in the above-mentioned limitation of energy. On the other hand in the velocity-modulator valves and also in other oscillator valves based upon transit time effects, the electrons are made to pass along the "oscillation circuit" (a cavity resonator is often used for this) and give off their energy inductively [4]). The anode is placed farther on in the valve and may have any desired size.

Principle of the velocity-modulator valve

Fig. 1, which also gives the notation used, shows diagrammatically the arrangement of a velocity modulator valve. The oscillator system contains two grids A_1B_1 (modulator) placed very close together and considered to be absolutely permeable, and another similar pair A_2B_2 (inductor) separated from A_1B_1 by a space free of field B_1A_2 (velocity-modulating space). The grid A_1 is at a potential V_0 with respect to the cathode. If an *A.C.* voltage $V_1 \sin \omega t_1 = aV_0 \sin \omega t_1$ is now applied between B_1 and A_1 ($a = V_1/V_0$ = depth of modulation), the voltage on B_1 is equal to $V_0 (1 + a \sin \omega t_1)$. The electrons arrive on the grid A_1 with a velocity v_0, which corresponds to the voltage V_0, and leave grid B_1 with a velocity v, which depends upon t_1 according to the relation

$$v = v_0 \sqrt{1 + a \sin \omega t_1} = v_0 \sqrt{1 + a \sin 2\pi t_1/T} \quad (1)$$

(T = time of oscillation). In this way the beam of electrons receives a velocity modulation. For $-T/4 < t_1 < T/4$, v increases with t_1, and thus electrons leaving B_1 later will be able to overtake the electrons leaving B_1 earlier. This can be seen clearly in *fig. 2*. The vertical x axis represents the direction of motion of the electrons. The modulator A_1B_1 lies at the point $x = 0$. We assume that at regular intervals of $1/12\ T$ electrons with equal velocity v_0 move from below towards the system A_1B_1. In the diagram of fig. 2 the distance x covered by each electron can be read off as a function of t.

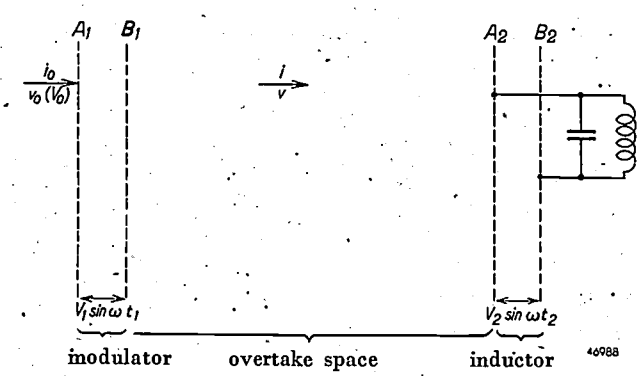


Fig. 1. Diagram showing the principle of the velocity modulator valve. A beam of electrons with a current i_0 amperes and an energy V_0 electron volts (velocity v_0) first passes two closely placed grids A_1B_1 (modulator), between which an A.C. voltage $V_1 \sin \omega t$ is applied.
The velocity modulation of the electrons thereby excited is converted in the "overtake space" B_1A_2, which is free of field, into a modulation of the convection current i. This modulated current gives off energy to an oscillation circuit (inductor) connected between the grids A_2 and B_2, which is represented diagrammatically by a self-induction and a capacity, although a separate L and C are not used, but for instance a cavity resonator.

The slope of the lines is here a measure of the velocity of the electrons. In the lower part of the figure it amounts to 45°, since the quantity x/v_0T is plotted along the vertical axis and along the horizontal axis the quantity t/T. The lines representing the behaviour of the different electrons are therefore parallel at the bottom and also equidistant. Due to the velocity modulation this is no longer true in the upper part of the figure. Upon passing the point $x = 0$ the lines in question will therefore show a bend the degree and direction of which are determined by the depth of modulation a and the moment t_1 at which the electron passes the modulator, that is the point $x = 0$. (For a given line t_1 is the abscissa of the point of intersection with the t-axis; t_1 is therefore the moment at which the electron that was at x at the moment t left the modulator ($x = 0$).) Due to the modulating voltage these bends will form a pattern in the part of the figure above the line $x = 0$ in which intersections of the lines in question occur, which means that the electrons overtake each other. The character of this pattern depends upon the value of the depth of modulation a. In fig. 2 the value $a = 0{\cdot}2$ has been chosen.

We are now able to follow the course of a given group of electrons in their passage through the valve. We draw for example a vertical line at $t = {}^1/_2\ T$ and cut it by the 45° line through the origin and 12 neighbouring lines. The ordinates of the

[4]) See for example p. 159 of the article by C. G. A. von Lindern and G. de Vries "Flat cavities as electrical resonators" Philips Techn. Rev. 8, 149, 1946.

intersection points, given in the figure by dots, represent the positions of 13 equidistant electrons which have not yet reached the modulator at the moment $t = -{}^1/_2T$. In order to discover what happens to this group of electrons at later moments, $t = 0, {}^1/_2T, T$ etc. we must cut the bent lines by the vertical lines corresponding to these moments. These points are also indicated in the figure and it is clear how a concentration gradually occurs in the middle of the group.

This concentration takes place around the electron which passes the modulator at the time $t = 0$, whose velocity v_0 thus remains unchanged. The concentration is thus likewise propagated through the valve at the velocity v_0. It increases in strength and at about $t = 1{}^1/_2\ T$ reaches a maximum, while for $t > {}^1/_2T$ the order of several electrons is found to have changed.

In a similar way it is possible to study an intersection of the network of lines of fig. 2 with a horizontal line, for instance the line $x/v_0T = 1$. This gives an idea as to what takes place at a given point in the valve in the course of time. The points of intersection of the chosen horizontal line with the oblique lines represent the moments at which the various electrons pass the point in question. It is clear that the closer the intersections lie on the horizontal line the larger the current at the corresponding moment. From the figure it can be read off quite easily that for a given value of x the current is a purely periodical function of time, but that for different values of x the character of this periodical function differs.

The model of *fig. 3* shows us the form of this periodic function for a number of values of x/v_0T.

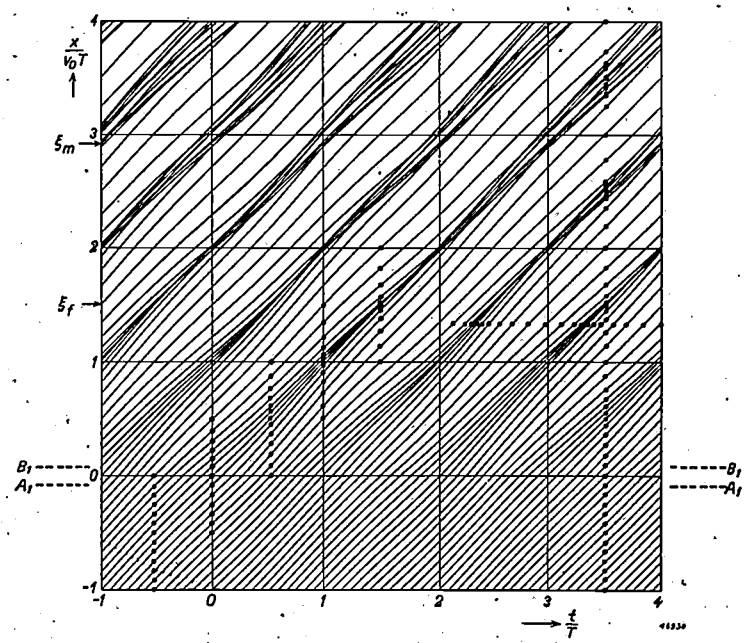


Fig. 2. Distance-time diagram for electrons which pass the grids A_1B_1 of fig. 1 at regular time intervals $T/12$. The quantities x/v_0T and t/T are plotted. Each of the continuous oblique lines represents the behaviour of x as a function of t for a given electron. The bend which the lines exhibit at $x = 0$ is determined by the depth of modulation a, for which in this case the value 0.2 is chosen. It may be seen how concentrations and rarefactions occur in the beam. At ξ_f (focus) the current first becomes infinite; at ξ_m (maximum plane) the first Fourier component of the A.C. is a maximum.

(In accordance with the abbreviations to be introduced presently, the quantity x/v_0T is indicated in the figure by ξ and t/T by τ.) Fig. 3 shows more clearly than fig. 2 how a concentration, which, for example, begins to form at the moment $t = 0$ and at $x = 0$ is propagated through the valve in the direction of increasing x, increases in density. The distance-time line of the current concentration makes an angle of 45° with the horizontal axis, which means — in reference to the scales chosen — that the concentration moves through the valve with a velocity v_0, as already noted above.

Fig. 4 is a reproduction of an instantaneous picture of the current through the valve (cross-section of the surface of fig. 3 for $\tau = 1.52$, the moment at which the current first becomes infinite). It may be seen how the centres of the current concentrations are separated from each other by a distance of about $\lambda_1 = v_0T$, which may be called

the "concentration wave length". Since the phenomena are repeated periodically with the oscillation time T, the instantaneous pictures, differing only by a time T, are identical. The concentrations thus travel through the valve as a kind of wave with an

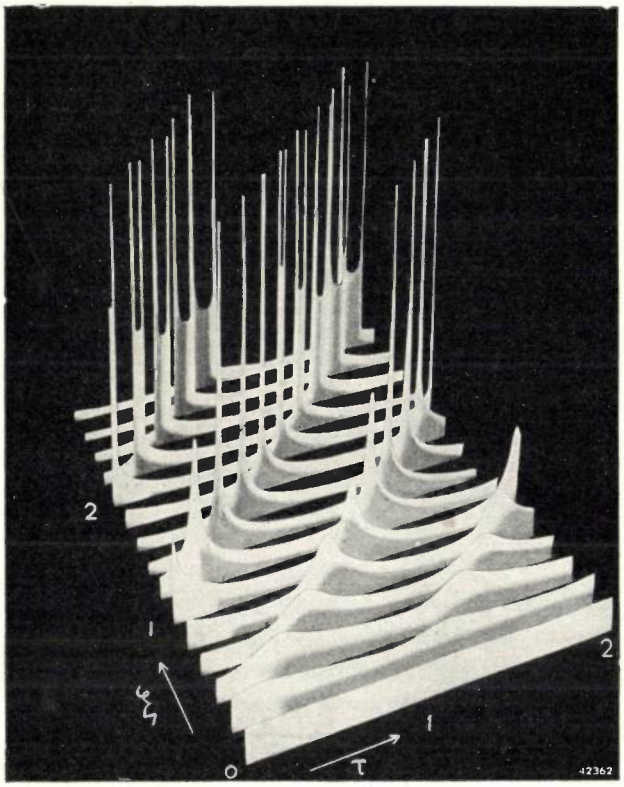


Fig. 3. The variation of the current i in a velocity modulator as a function of the time t and the distance x to the modulator. Plotted horizontally are the values of $x/v_0\ T\ (=\xi)$ and t/T $(=\tau)$, vertically the value of i/i_0 where i_0 is the unmodulated current. The upper edges of the vertical strips thus indicate how the current at a given place varies as a function of the time. The current maxima are propagated at a velocity v_0 in the longitudinal direction of the valve (ξ direction). They meanwhile first increase in height and later split up into two peaks which move farther and farther apart.

oscillation time T, a velocity of propagation v_0 and a wave length λ_1. The height of the top of the wave thereby increases and becomes infinite at a certain spot. The peak then splits up into two parts which become farther and farther apart. It is to be noted that the electromagnetic oscillations of the same oscillation time T have a velocity c and a wave length $\lambda = cT$. For the relation between λ_1 and λ the following relation is apparently valid:

$$\lambda_i/\lambda = v_0/c = 0{,}002\sqrt{V_0}\ (V_0 \text{ in volts}). \quad (2)$$

In the foregoing and with reference to figs. 3 and 4 we have discussed the behaviour of the electron current i as a function of the distance x from the modulator and of the time t. We shall now show how this function and thus also the figures can be calculated. We shall then particularize the calculations for the case, so important in practice, where the velocity-modulator valve oscillates at the fundamental frequency ω and the depth of modulation a is small ($a \ll 1$). The results of these calculations will enable us to discuss the theory of the velocity modulator as amplifier and as oscillator and to explain it with an empirical example.

Calculation of the current variations

For the calculation of i as a function of x and t it is helpful to use as auxiliary quantity the time t_1 already introduced. Obviously

$$x = v\,(t - t_1),$$

where v, the velocity of the electron, is given by equation (1). When this value of v is substituted in the above equation we find for the relation between t_1, t and x:

$$t = t_1 + x/v = t_1 + x/v_0\sqrt{1 + a\sin\omega t_1}. \quad (3)$$

The latter equation and those following can be more simply written when one uses as variables, instead of the distance x and the time t, the non-dimensional quantities used already in figs. 2, 3 and 4.

$$\xi = x/v_0T \text{ and } \tau = t/T. \quad (4)$$

Then according to (2)

$$\xi = x/v_0T = x/\lambda_i = 500\,x/\lambda\sqrt{V_0}.\ (V_0 \text{ in volts}) \quad (5)$$

The quantity ξ has a simple physical significance. Since the distance between the centres of two current concentrations according to fig. 4 amounts to λ_1, for the case where ξ is a whole number $\xi = x/\lambda_i$ represents the number of concentrations or parcels situated within the distance x. With an obvious extension even when ξ is not a whole number it can still be called the "number of parcels" in the distance x; we then speak, for example, of $^3/_4$ parcel in the distance d when the distance between two parcels is $4/3\ d$.

With the help of the new variables ξ and τ equation (3) may be written as follows:

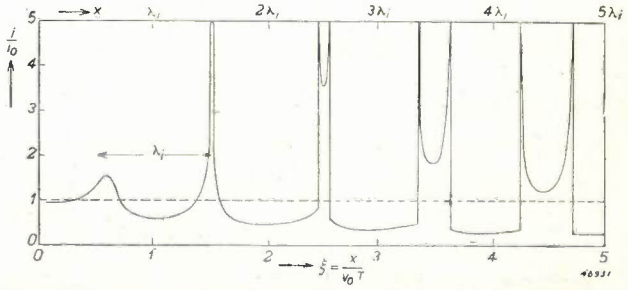


Fig. 4. Instantaneous picture (for $\tau = 1.52$) of the current variation in the valve as a function of the distance x to the modulator. The centres of the current concentrations have a mutual distance approximately equal to $\eta_1 = v_0T$ and are propagated in the valve as a wave of wave length λ_i, velocity v_0 and oscillation time T.

$$\tau = \tau_1 + \xi/\sqrt{1 + \alpha \sin 2\pi\tau_1} \quad . \quad . \quad . \quad . \quad (6)$$

This relation between the reduced times of arrival (τ) and departure (τ_1) is shown in *fig. 5* for a number of values of the reduced distance ξ (it is assumed that $\alpha = 0.2$). It may be seen in the figure that for sufficiently large values of ξ, namely $\xi > \xi_f$, three values of τ_1 correspond to one value of τ, *i.e.* at a single given moment, τ, electrons pass the spot in question ξ, which left the modulator at three different moments τ_1. On the other hand for $\xi > \xi_f$ only one value of τ_1 corresponds to one value of τ. We shall revert to the significance of ξ_f later on.

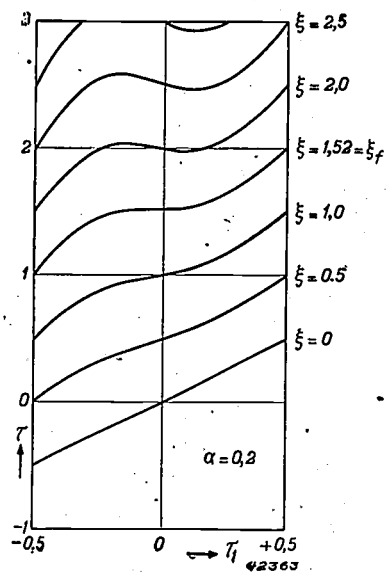


Fig. 5. Reduced time of arrival τ of electrons, which have covered a reduced distance ξ, as a function of the reduced time of departure τ_1. The scale for τ_1 is taken twice as large as that for τ. Depth of modulation $\alpha = 0.2$. At $\xi = 1.52$ the value of $d\tau_1/d\tau = i/i_0$ first becomes infinite (focus).

After these preparatory remarks the ratio of the current i at x to the unmodulated current i_0 entering the modulator can easily be derived. Let us consider a group of electrons which pass x between t and $t + dt$ and there give rise to a current i, *i.e.* have a charge $i\,|dt|$. If $x \leqq v_0 T \xi_f$ the group in question is made up exclusively of electrons which have left the modulator in the same time interval $(t_1, t_1 + dt_1)$ and which there had the charge $i_0|dt_1|$. Since no electrons have been lost en route

$$i\,|dt| = i_0|dt_1| \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

If on the other hand $x > v_0 T \xi_f$ the group of electrons in question is actually the sum of different groups, in the case of fig. 5 three groups which left the modulator in three different time intervals $(t_1^{(1)}, t_1^{(1)} + dt_1^{(1)})$, $(t_1^{(2)}, t_1^{(2)} + dt_1^{(2)})$ and $(t_1^{(3)}, t_1^{(3)} + dt_1^{(3)})$. Equation (7) must accordingly be replaced by

$$i\,|dt| = i_0 \sum_{k=1}^{3} |dt_1^{(k)}| \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

We now introduce the reduced variables (4) again. Moreover, in order not to have to distinguish each time between the cases (7) and (8) we shall collect the two cases into one equation:

$$i\,|d\tau| = i_0 \Sigma |d\tau_1| \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

on the understanding that here and in the equations following from it the sign of summation can be omitted when $\xi < \xi_f$.

$$i = i_0 \Sigma \frac{1}{|d\tau/d\tau_1|} \quad . \quad . \quad . \quad . \quad . \quad (10)$$

If one substitutes here the value of $d\tau/d\tau_1$, which can easily be calculated from (6), the result is:

$$i/i_0 = \Sigma \frac{1}{|1 - \pi\alpha\xi\,(1 + \alpha \sin 2\pi\,\tau_1)^{-3/2} \cos 2\pi\,\tau_1|} \quad . \quad (11)$$

Equation (11) gives i/i_0 as a function of τ_1, but we wish to know i/i_0 as a function of τ and must therefore eliminate τ_1 from (11) and (6), which, however, is not possible explicitly. Figures 3 and 4 are calculated by determining corresponding values of i/i_0 and τ for a number of values of τ_1 or ξ from (6) and (11).

Figs. 2 to 5 can also be derived directly from the surface shown in *fig. 6*, which represents τ as a function of τ_1 and ξ according to equation (6). For the sake of clearness α is there assumed to be equal to 0.4. The cross-sections perpendicular to the τ_1 axis are straight lines which satisfy (6). If these straight lines are projected on the τ-ξ surface fig. 2 is obtained, while fig. 3 appears when the derivate $d\tau_1/d\tau = i/i_0$ from fig. 6 is determined and plotted as a function of τ and ξ. Fig. 4 is a cross-section of fig. 3 for τ constant, while the curves in fig. 5 are cross-sections of fig. 6 for ξ constant.

We now return to the significance of the above-mentioned critical value ξ_f of $\xi = x/v_0 T$. According to equation (10) the current i at x becomes infinite [5]) when $d\tau/d\tau_1 = 0$, *i.e.* when the curve corresponding to $\xi = x/v_0 T$ in fig. 5 is parallel to the τ_1 axis. It is clear from the figure that for $\xi < \xi_f$ this is never the case, for $\xi = \xi_f$ once per period and for $\xi > \xi_f$ twice per period. The critical value ξ is thus the smallest value of ξ for which the current becomes infinite. For larger values of ξ it even becomes infinite twice per period, corresponding to the double peaks in figs. 3 and 4. For small values of ξ on the other hand it always remains finite and from fig. 3 it may be seen that for small values of ξ, *i.e.*

[5]) The current maxima in question remain finite in a theory which takes the space charge into account.

close to the modulator, there are in fact no infinite peaks.

The spot in the valve for which $\xi = \xi_f$ is called the "focus", because at that point (in connection with the fact that the point for which $d\tau/d\tau_1 = 0$ is in this case at the same time a point of inflexion) the highest concentration of electrons occurs.

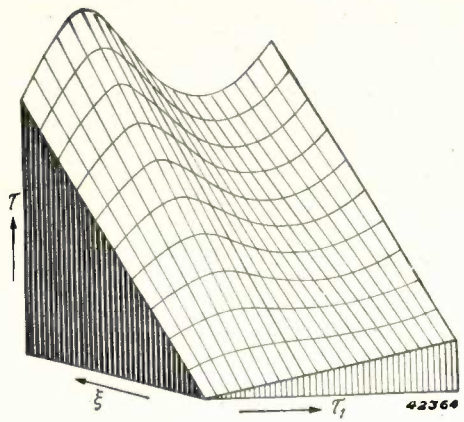


Fig. 6. The τ_1 τ_1, ξ plane (*cf.* fig. 5). The scale for τ_1 is five times as large as that for τ; $a = 0.4$.

This concentration of the electrons shows a strong analogy to the focussing of light rays by a lens and has therefore been called "phase focussing" by Brüche and Recknagel [6]). In the optical case rays of originally the same direction are focussed by a lens at a point in space; in the case of the velocity modulator electrons of originally equal velocities are focussed by an alternating field at a point in time (phase point). In the case of the lens the changes in direction of the rays depend upon the place, where they strike the lens, while in the case of the alternating field the changes in velocity of the electrons are determined by the time, or in other words by the phase at which they enter the field. The analogy is also clear from fig. 2, where the originally parallel distance-time lines (at the bottom of the figure) are broken by the A.C. field and brought to a "focus". If in fig. 2 the time axis is considered as a distance axis the case is the same as that of optical focussing.

In order to calculate ξ_f explicity the smallest value of ξ must be sought for which $d\tau/d\tau_1 = 0$, *i.e.* for which one of the denominators in (11) becomes zero. The values of ξ for which this is the case are given by

$$\frac{(1 + a \sin 2\pi\tau_1)^{3/2}}{\cos 2\pi\tau_1} \quad . \quad . \quad . \quad . \quad (12)$$

The minimum of this is found by differentiation with respect to τ_1. For $\tau \ll 1$ the following can be derived from (12):

$$\xi_f = 1/\pi a \ \ (\text{at } a \ll 1); \quad . \quad . \quad . \quad . \quad . \quad (13)$$

the point of inflection then lies at $\tau_1 = 0$. For larger values of a, ξ_f becomes smaller than $1/\pi a$ In fig. 5, where $a = 0.2$, equation (13) is fairly well satisfied by $\xi_f = 1.52$ instead of $1/0.2\,\pi = 1.59$.

Calculation of the fundamental component of the current variations

Although, as stated above, it is impossible to find an explicit expression for the current variations as a function of time and place, it is possible to calculate the current value for the fundamental frequency of the current, thus the Fourier component with the frequency ω, at least if we confine ourselves to the case just considered where $a \ll 1$.

This component is usually the most important in practice, because the circuit in which it is desired to cause oscillations at the output end is tuned to the fundamental frequency and possesses a large impedance only for that frequency.

For $a \ll 1$ (6) and (11) can be developed in powers of a and the linear terms in a are sufficient. One then obtains [7])

$$2\pi\,(\tau - \xi) = 2\pi\tau_1 - \pi a\,\xi \sin 2\pi\,\tau_1 \quad . \quad . \quad (14)$$

and

$$i/i_0 = \Sigma\, 1/(1 - \pi a \xi \cos 2\pi\tau_1) \quad . \quad . \quad . \quad . \quad . \quad (15)$$

In connection with equation (13) the summation symbol in (15) may be omitted for $\pi a\xi \leqq 1$. Calculation shows that the fundamental component of the A.C. part of i can be represented, for $\pi a\xi \leqq 1$ as well as for $\pi a\xi > 1$, by [8])

$$i_2\,(\xi, \tau) = 2 i_0\, J_1\,(\pi a \xi) \cos 2\pi\,(\tau - \xi).\ (a \ll 1) \quad (16)$$

J_1 is here the Bessel function of the first order. The amplitude I_2 of the fundamental component at the position ξ thus has the value

$$I_2\,(\xi) = 2\, i_0\, J_1\,(\pi a \xi) \quad . \quad . \quad . \quad . \quad . \quad (17)$$

From (17) it follows that for $a < 1$ the amplitude I_2 of the A.C. depends only on $a\xi$ and no longer on a alone, *i.e.* with varying a the same value of I_2 is retained when the distance covered by the electrons varies inversely proportional to a. For the focus ($\pi a\xi = 1$) I_2 is equal to $2\, i_0\, J_1\,(1) = 0.88\, i_0$; the maximum amplitude, however, is reached for

[6]) E. Brüche and H. Recknagel Z. Phys. **108**, 459, 1938.

[7]) When the summation sign is omitted in (15), for $\pi a\xi < 1$, these are the equations of an extended cycloid with $2\pi\,(\tau - \xi)$ and the reciprocal of i/i_0, *i.e.* i_0/i, as coordinates and τ_1 as variable parameter. For $\pi a\xi > 1$ the equation is obtained of a shortened cycloid; for certain values of τ the quantity $1/(1 - \pi a\xi \cos 2\pi\tau_1)$ then has more than one value and in order to calculate the current equation (15) must be used with the sign of summation.

[8]) Equation (14) is identical with Keppler's equation for the movement of the planets, from which Bessel solved τ_1 as a function of τ with the aid of the functions which were later named after him. The solution of $d\tau_1/d\tau = i/i_0$ as a function of τ leads to (16).

$\pi a\xi = 1.84$ and amounts to $2\, i_0\, J_1\,(1.84) = 1.16\, i_0$. The plane in which this happens we call the "maximum plane", to distinguish it from the focus; the following holds for that plane:

$$\pi(\xi a)_m = 1.84, \quad \ldots\ldots \quad (18a)$$

Hence if a is taken as constant:

$$\xi_m = 1.84/\pi a. \quad \ldots\ldots \quad (18b)$$

The velocity modulator as amplifier

Before considering the velocity modulator as an exciter of oscillations we shall first study its action as an amplifier. For that purpose an A.C. voltage $aV_0 \sin \omega t$ is applied between the grids B_1 and A_1 (modulator). The energy of the beam of electrons modulated by this A.C. voltage is given off to an oscillation circuit with impedance Z tuned to ω and connected between A_2 and B_2 (inductor). We shall now assume that the distance d_2 between A_2 and B_2 is very small as compared with the concentration wave length λ_i. Then the A.C. which flows from B_2 through the impedance Z to A_2 is exactly equal to the value i_2 in (16) of the fundamental component of the current through the valve with the amplitude I_2 [9]). If V_2 is the amplitude of the A.C. voltage between B_2 and A_2 and the impedance for the higher harmonics may be ignored, compared with Z, the following holds: $V_2 = I_2 Z$, and the total oscillation energy which flows through the output circuit per second is $^1/_2\, I_2 V_2$. If, further, we understand by the efficiency η the ratio of this total oscillation energy to the D. C. energy $i_0 V_0$ given off to the electrons, then according to (17)

$$\eta = \frac{I_2\, V_2}{2 i_0 V_0} = \frac{V_2}{V_0}\, J_1\,(\pi a\xi) \quad \ldots\ldots \quad (19)$$

In order to give this efficiency its maximum value two conditions must be fulfilled, namely $I_2/2\, i_0$ or $J_1\,(\pi a\xi)$ must be a maximum and V_2/V_0 must be a maximum. The first condition implies that the maximum plane must be situated at the position of the inductor; the corresponding value ξ_m of ξ follows from (18b), while

$$I_2/2 i_0 = J_1\,(\pi a \xi_m) = J_1\,(1.84) = 0.58\,. \quad (20)$$

As to the second condition it may be noted that for V_2/V_0 electrons will be thrown back before they reach the inductor. We thus assume that V_2 can at the most be equal to V_0. It then follows from (19) that

$$\eta_{\max} = 0.58 \quad \ldots\ldots \quad (21)$$

The amplification factor, which is obviously given by V_2/aV_0, is therefore in the case of maximum efficiency equal to the reciprocal $1/a$ of the depth of modulation.

The conditions which must be satisfied by i_0, V_0 and Z at given values of the depth of modulation a and length of the overtake space in order to attain this maximum efficiency follow from (5) and the relation $V_2 = V_0$.

According to (5) the following holds:

$$\sqrt{V_0} = 500\, x/\lambda\xi \quad (V_0 \text{ in volt}), \quad \ldots \quad (22)$$

thus, on the basis of (18)

$$\sqrt{V_0} = 855\, ad/\lambda \ (V_0 \text{ in volt}) \quad \ldots \quad (23)$$

Further, for the realization of $\eta_{\max}$, Z must have a value Z_m which satisfies $V_2 = 1.16\, i_0 Z_m = V_0$; thus $Z_m = V_0/1.16\, i_0$.

If we denote V_0/i_0, the "valve resistance", by Z_b and the ratio of external resistance Z to valve resistance Z_b by ϱ, then

$$\varrho = Z/Z_b = Z\, i_0/V_0 \quad \ldots\ldots \quad (24)$$

$$\varrho_m = Z_m/Z_b = 1/1.16 = 0.86 \quad \ldots \quad (25)$$

At a depth of modulation a, therefore, theoretically the maximum efficiency of 58% is obtained at $\sqrt{V_0} = 855\, ad/\lambda$ and $\varphi = \varrho_m = 0.86$.

The velocity modulator as oscillator

In order to use the velocity modulator as oscillator there must be a back-coupling of the inductor to the modulator. Just as in the case of triode oscillators, this back-coupling must have a certain minimum value to make the occurrence of oscillations possible. The "back-coupling condition" can easily be indicated, after the foregoing.

When an A.C. voltage $V_1 \sin 2\pi\tau$ is applied between B_1 and A_1 an A.C. occurs in the impedance of the inductor, whose fundamental frequency i_2 is determined by (16), so that for the voltage over the impedanze Z between B_2 and A_2 the following is valid:

$$-i_2 Z = -2\, i_0 Z\, J_1\,(\pi a\xi) \cos 2\pi\,(\tau - \xi) = \\ = -V_2 \cos 2\pi\,(\tau - \xi)\,. \quad (26)$$

The amplitude of this A.C. voltage thus amounts to

$$V_2 = 2 i_0\, Z\, J_1\,(\pi a\xi) \quad \ldots\ldots \quad (27)$$

If it is desired to maintain oscillations a part K (back-coupling factor) of V_2 must be sent back to

[9]) See for example C. J. Bakker and G. de Vries, Physica, 2, 683, 1935, where the case for finite $d = d_2/\lambda_1$ is discussed. For this the amplitude becomes

$$2\, i_0\, \frac{\sin \pi\tau}{\pi\tau}\, J_1\,(\pi a\xi).$$

A_1B_1 and must there be equal to V_1 in amplitude and phase. We shall return to the phase condition later. The amplitude condition obviously becomes

$$KV_2 = V_1 = aV_0, \quad . \quad . \quad . \quad . \quad (28)$$

or in connection with (27) and (24)

$$K \cdot 2i_0 Z J_1(\pi a\xi) = aV_0,$$

so that

$$K = a/2\varrho J_1(\pi a\xi) \quad . \quad . \quad . \quad . \quad . \quad (29)$$

At the beginning of the oscillations when a is practically equal to zero, K must have a value $K_{\min}$ which according to (29) satisfies

$$K_{\min} = 1/\pi\varrho\xi, \quad . \quad . \quad . \quad . \quad . \quad (30)$$

because for $a \ll 1$, $2J_1(\pi a\xi) = \pi a\xi$.

When one takes $K > K_{\min}$ the amplitude of the oscillations increases; the value of the depth of modulation a in the modulator then follows from (29).

The efficiency is again determined by (19), for which, with the relation $V_2/V_0 = a/K$ following from (28), one may also write

$$\eta = \frac{I_2}{2i_0}\frac{V_2}{V_0} = \frac{a}{K} J_1(\pi a\xi) \quad . \quad . \quad . \quad (31)$$

Just as in the case of the amplifier, η is a maximum when $I_2/2i_0$ and V_2/V_0 or a/K are maxima. The latter is again true for $V_2 = V_0$ or $a = K$. The conditions for maximum efficiency (18a) and (25) hereby pass over into:

$$\left.\begin{aligned} \pi(\xi K)_m &= 1.84 \quad (32) \\ \varrho_m &= 0.86 \quad (33) \end{aligned}\right\} \eta_{\max} = 0.58.$$

Equations (29) and (31), however, lead not only to the conditions for $\eta_{\max}$ but also give the efficiency for any other case, since it depends upon the two characteristic quantities ϱ and ξK. This may be seen more clearly when (29) and (31) are written in the following form:

$$\varrho\xi K = a\xi/2J_1(\pi a\xi) \quad . \quad . \quad . \quad . \quad . \quad (34)$$

$$\eta\xi K = a\xi J_1(\pi a\xi) \quad . \quad . \quad . \quad . \quad . \quad (35)$$

From equation (34) the value of $a\xi$ at given values of ϱ and ξK follows, while substitution of this in (35) gives the efficiency. It is also possible to calculate both $\varrho\xi K$ and $\eta\xi K$ as functions of $a\xi$ according to (34) and (35) and then represent $\eta\xi K$ as a function of $\varrho\xi K$ in a single curve.

In *fig.* 7 η is plotted as a function of ξK with ϱ as a parameter. At a constant value of ϱ the maximum of η is reached when the inductor lies in the maximum plane. For $\varrho = 0.86$ this maximum reaches its optimum value of 58%. For $\varrho > 0.86$ only those parts of the curve are drawn for which $V_2 < V_0$, *i.e.* $a < K$, since with $V_2 > V_0$ electrons would be reflected. It may be seen from the figure that for these values of ϱ the efficiency always remains 58%.

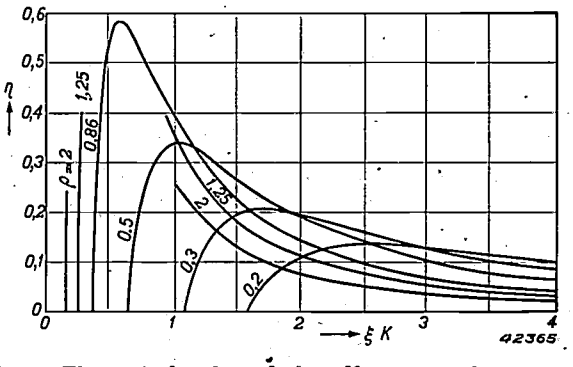


Fig. 7. Theoretical value of the efficiency η of a velocity-modulator valve as a function of ξK at constant ϱ. K = back-coupling factor, ξ = number of parcels between the modulator and the inductor, ϱ = ratio of external impedance Z to valve resistance V_0/i_0. The theoretical optimum efficiency of 58% is attained at the maximum of the curve for $\varrho = 0.86$.

Fig. 7 may also be interpreted as follows: for $\varrho < 0.86$ the current through the valve can be entirely modulated by taking $\pi a\xi = 1.84$; the impedance, however, is then still too small to modulate also the voltage to its maximum value ($V_2 = V_0$). This latter takes place only at $\varrho = 0.86$. For $\varrho > 0.86$ on the other hand the impedance is so large that the voltage can be entirely modulated; this takes place at constant ϱ for two different values of ξK and $a\xi$ respectively (in the figure the points where the curves end). For these values of ϱ, however, the current at the inductor is again not entirely modulated, so that η remains smaller than 58%.

According to (35) the maxima of η for $\varrho < 0.86$ (current maximum) in fig. 7 lie on the equilateral hyperbola $\pi\eta\xi K = 1.84\, J_1(1.84)$. For $\varrho > 0.86$ the curves end at the points for which $V_2 = V_0$ (voltage maximum) or $a = K$; these points, according to (35), lie on the curve $\xi = J_1(\pi\xi K)$ and according to (34) satisfy the relation $2\varrho J_1(\pi\xi K) = 2\varrho\eta = 1$.

In the foregoing no use has been made of the phase condition of the back-coupling. According to (26) the voltage between B_2 and A_2 is

$$-V_2 \cos 2\pi(\tau - \xi) = V_2 \sin 2\pi(\tau - \xi - {}^1/_4), \quad (36)$$

while that between B_2 and A_2 was taken equal to

$$V_1 \sin 2\pi\tau.$$

For the phase difference Φ between the two voltages, therefore, the following holds:

$$\Phi = (\xi + {}^1/_4)\, 2\pi.$$

Due to the back-coupling the phase is rotated through an angle φ. The phase condition of the back-coupling is now apparently:

$$\Phi + \varphi = (\xi + {}^1/_4)\, 2\pi + \varphi = n \cdot 2\pi \text{ } (n \text{ a whole number}),$$
$$\xi = n - {}^1/_4 - \varphi/2\pi \quad . \quad . \quad . \quad . \quad (37)$$

Further, according to (22) V_0 is connected with ξ, so that finally the value of V_0 is determined by

$$\sqrt{V_0} = 500\, d/\lambda\, (n - {}^1/_4 - \varphi/2\pi)\,. \quad (V_0 \text{ in volts}) \quad (38)$$

When the values of φ, d, λ, K and Z are determined by the oscillator system (modulator circuit, inductor circuit and back-coupling connections), the values of V_0 and i_0 may still be chosen. For the excitation of oscillations ξ must satisfy (37). Hereby several discrete values of ξ are fixed, and thus, in connection with (38), also of V_0. The value of i_0 can then still be chosen arbitrarily. This determines $\varrho = i_0 Z/V_0$. The efficiency, which is a function of ϱ and ξK, then follows from fig. 7.

We have defined the efficiency η as the ratio of the total oscillation energy to the energy $i_0 V_0$ given off to the electrons. In practice the "aerial efficiency" η_a is of more importance. By this we understand the quotient of the aerial power and $i_0 V_0$. Obviously $\eta = \eta_a + \eta_s$ where η_s refers to the oscillator system.

If now the impedances Z_a and Z_s observed from A_2B_2 (fig. 1) of the aerial and the oscillator system respectively are known, the following relation holds for the total impedance Z: $1/Z = 1/Z_a + 1/Z_s$; ϱ is then known, and from fig. 7 η and, after some calculation, η_a also can be derived.

In velocity modulators η_a is often appreciably smaller than η, since Z_s cannot be taken large enough. Usually one works here with high voltages and currents whereby V_0/i_0 becomes high and ϱ low. As an example we choose $V_0 = 5$ kV, $i_0 = 50$ mA and $\xi K = 1$. Then $Z_b = 100\,000\ \Omega$. For the occurrence of oscillations according to (30) $\varrho \xi K$ must at least be equal to $1/\pi$, thus ϱ at least $1/\pi$ and Z at least 32 000 Ω. If Z_s is little larger than this amount, Z_a must be taken very large, since otherwise the total Z becomes too small and the oscillations stop. The aerial can then take up only little energy and η_a will be much smaller than the value of η without aerial (η_0) following from fig. 7. Thus for example for $Z_s =$ 40 000 Ω, for the case with no aerial $\eta_0 = 0.22$, while with an aerial the highest value attainable for η_a is only 0.01. With increasing Z_s the ratio η_a/η_0 becomes more favourable; for $Z_s = 100\,000\ \Omega$ thus $\varrho = 1$, we arrive at $\eta_0 = 0.38$ (without aerial) and $\eta_a = 0.17$ maximum. For $Z_s = \infty$ all the energy can be concentrated in the aerial.

Finally it must be remarked that in comparing the experimental results with the results of the above theory no precise agreement can be expected, since the theory is derived on the following simplifying assumptions:

1) the space charge may be disregarded;
2) the depth of modulation is small ($\alpha \ll 1$);
3) the grids are completely permeable for electrons;
4) the electrons always move parallel to the axis;
5) the grid distances are infinitesimally small.

As far as point 4) is concerned it may be noted that when slits with no grids are used (see below) the field is very unhomogeneous and, moreover, the magnetic field holding the electron beam together causes the electrons to describe spiral paths. This may cause a flattening of the maxima along the electron beam.

Empirical example

In order to elucidate the above theoretical considerations we shall discuss an experimental case which has indeed only a small efficiency but in which the various characteristic quantities are easy to determine. The arrangement is shown in *fig. 8*. The beam of electrons here passes through a tube made of silica to reduce the dielectric losses. With the aid of a coil, not shown in the figure, the beam of electrons is kept concentrated along the axis of the tube. The oscillation system, shown in cross-section for the sake of clarity, consists of a co-axial Lecher system of copper outside the silica tube; the inner and outer cylinders are kept at the proper distance from each other by two insulating discs.

In the silica tube itself there are no grids of any kind for varying the velocities of the electrons, the beam being affected only by external electrodes. In the left-hand slit the electron velocity is modulated, in the inner cylinder the velocity modulation is converted into a density modulation and this density modulation induces again an A.C. field in the right-hand slit. As a result the whole system oscillates in such a way that voltage maxima occur at the slits, while the length of the inner cylinder is for example about ${}^1/_2\ \lambda$. The energy can be taken off by a loop, which is brought into the interior through a slit, which may be seen at the middle of the upper side of the outer cylinder.

Instead of the grids A_1B_1 and A_2B_2 of fig. 1 we here have the slits S_1T_1 and S_2T_2 (*fig. 9*). The impedance Z between S_2 and T_2 is approximately equal to the impedance between S_2 and U_2 of the open ${}^1/_2\ \lambda$ system $T_1\ U_1\ U_2\ S_2$ and can be calculated by means of the formula [10])

$$Z = \frac{240}{\pi\delta} \frac{R_1 R_2}{R_1 + R_2} \left(\ln \frac{R_2}{R_1}\right)^2,$$

where R_2, the radius of the outer cylinder, equals 3 cm, R_1, that of the inner cylinder, equals 0.95 cm and δ, the depth of penetration of the skin effect,

[10]) See C. G. A. von Lindern and G. de Vries Lecher systems, Philips Techn. Rev. 6, 240, 1941.

is $4\times10^{-5}\sqrt{\lambda}$ (copper). With $\lambda = 35$ cm in the case in question this gives approximately

$$Z = 3 \times 10^6\ \Omega.$$

The inside of the cylinder $T_1\ S_2$ (length $d = 15$ cm) here serves as overtake space, while the outside functions as back-coupling connection of about $^1/_2\lambda$ length. Since T_1 and S_2 are opposite in phase, the angle φ, through which the phase is rotated upon back-coupling, is zero, so that according to (37) the number of parcels in the length d becomes

$$\xi = n - {}^1/_4.$$

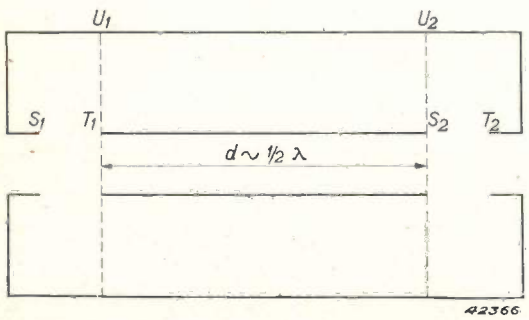


Fig. 9. Diagram of the oscillator system of fig. 8. The system oscillates in about the same way as the open Lecher system $T_1\ U_1\ U_2\ S_2$ with a length of $^1/_2\ \lambda$. The outside of the cylinder $T_1\ S_2$, inside of which is the overtake space, serves as back-coupling connection with $K = 1$.

According to (38) the tube voltage is thereby fixed:

$$\sqrt{V_0} \approx 250/(n-{}^1/_4),\ (V_0 \text{ in volts}) \qquad (39)$$

since here $d \sim {}^1/_2\ \lambda$. For the back-coupling factor K the following holds:

$$K = 1.$$

The valve oscillates at $V_0 = 5000$ volts, for which according to (39) n is approximately equal to 4, so that $3^3/_4$ parcels are situated within the overtake space [11]).

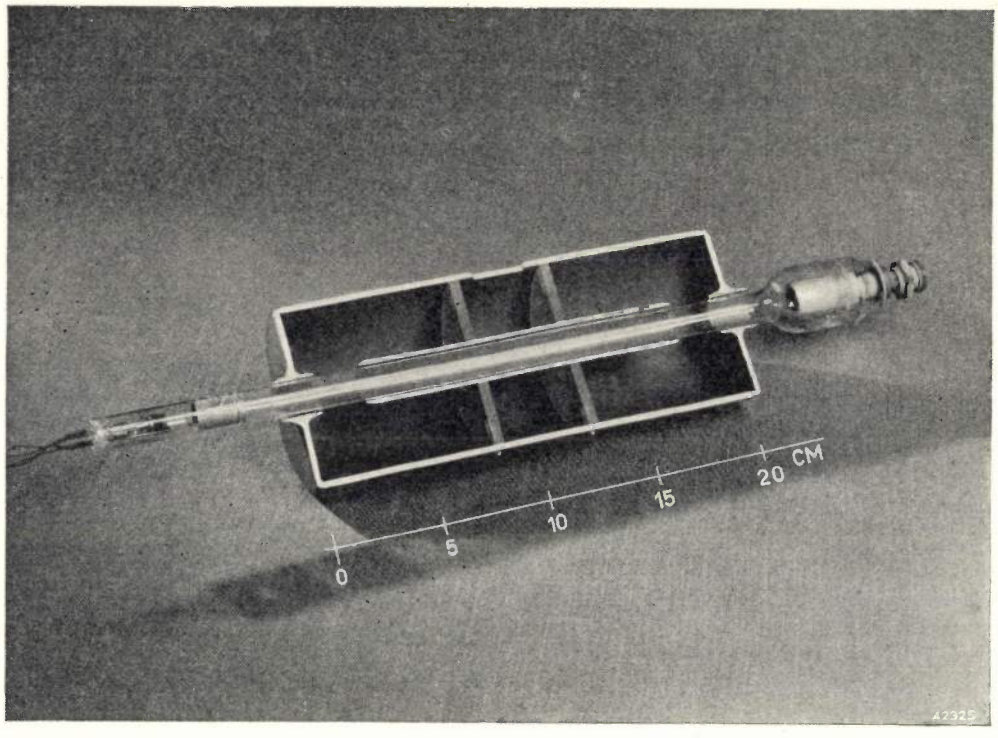


Fig. 8. External oscillator system consisting of two co-axial cylinders (shown in cross-section) with two slits. The oscillations are excited by the electron beam in a silica tube inserted along the axis of the oscillator system. The whole is placed in a co-axial magnetic field.

According to (30) the oscillations begin at $\varrho\xi K = 1/\pi$. The width of the slit here, however, is not infinitesimal, which makes the minimum value of $\varrho\xi K$ larger (*cf.* footnote 9)). When this is taken into account, with $\xi = 3^3/_4$, $V_0 = 5000$ volts and $Z = 3 \times 10^5\ \Omega$, one finds for the current at which the oscillations begin a value of 2 mA. The experimental value was $3^1/_2$ mA.

Instead of having the oscillator system external it may also form part of the valve itself. In that way oscillations of considerably higher frequency can be excited.

[11]) With an external oscillator system deviations from equation (39) may occur, since due to wall charges the energy of the electrons need not correspond to the anode voltage V_0.

It is clear that with the system of fig. 8 only a low efficiency can be expected, since according to fig. 7 when $\xi \approx 4$ and $K = 1$, thus $\xi K = 4$, the total oscillation efficiency η is not more than 10%, only part of which reaches the aerial. Experimentally in the case of fig. 8 a power of 8 watts could be realized in an incandescent lamp, which corresponds to an efficiency of $\eta_a = 3\%$. In judging this result account must be taken of the simplifications introduced into the theory (see above).

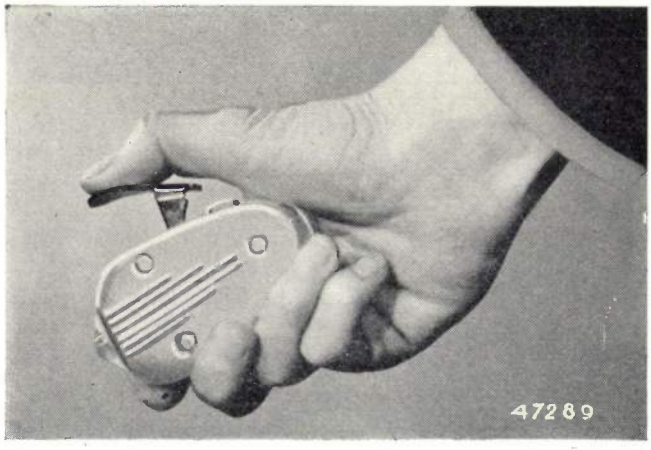
In order to obtain the theoretical maximum efficiency of 58%, according to (32) $\pi\xi K$ should be taken equal to 1.84. On the other hand for $K = 1$ ξ must equal $n-{}^1/_4$ (n a whole number); the optimum value of ξ is thus ${}^3/_4$. This leads, however, in connection with (39), to an unusable value for V_0 ($> 100\,000$ volts). Therefore in order to attain high efficiencies the oscillator systems must be designed otherwise, but we shall not go more deeply into that here.

VOL. 8 No. 8, p. 225-256 AUGUST 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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



THE DYNAMO POCKET TORCH

by E. A. van IJZEREN.

621.313.39-871

Description of the Philips dynamo pocket torch type No. 7424. When squeezed "normally" it generates a light flux of 2·4 lumens, for which the user has to apply about 0·9 watt of mechanical energy. In spite of the intermittent excitation the light flux varies only little while the dynamo is being squeezed. This is due partly to a large moment of inertia and high speed of rotation of the rotor and partly to the flatness of the voltage — speed characteristic. The dynamo, which on broad lines is analogous to a bicycle dynamo, contains a permanent magnet of eight poles made of "Ticonal" 0.8 as rotor. The armature winding is on a stator of a construction very different from that commonly used and adapted to a high degree to mass production. The dynamo is made resistant to short-circuiting, as will be explained.

In this mechanized age it is not often that man reverts to the use of his own physical strength to perform work which he has long been accustomed to have performed by other sources of energy. That is, however, what has occurred on a miniature scale in recent years since the dynamo pocket torch came to occupy an important position side by side with the traditional battery pocket torch. This development was stimulated by the fact that during the war batteries became scarce, while black-out precautions increased the need of a pocket torch. But now that war scarcity and blackout are things of the past the dynamo pocket torch still appears to be in great demand, obviously because the public has become more conscious of the disadvantages of the battery torch even in peace time: the necessity of repeatedly buying new batteries and the chance that at a critical moment the batteries kept on hand will be found to be exhausted when the torch has not been used for some time. These disadvantages are especially pronounced in tropical countries, where batteries deteriorate more quickly and fresh supplies have sometimes to be fetched a long distance away.

There is another example where light is generated by physical strength; the bicycle lamp. In the development of the dynamo pocket torch extensive use has been made of the experience gained in the

construction of modern bicycle dynamos. This will become evident in the following description of the Philips dynamo pocket torch, type no. 7424 [1])

General construction

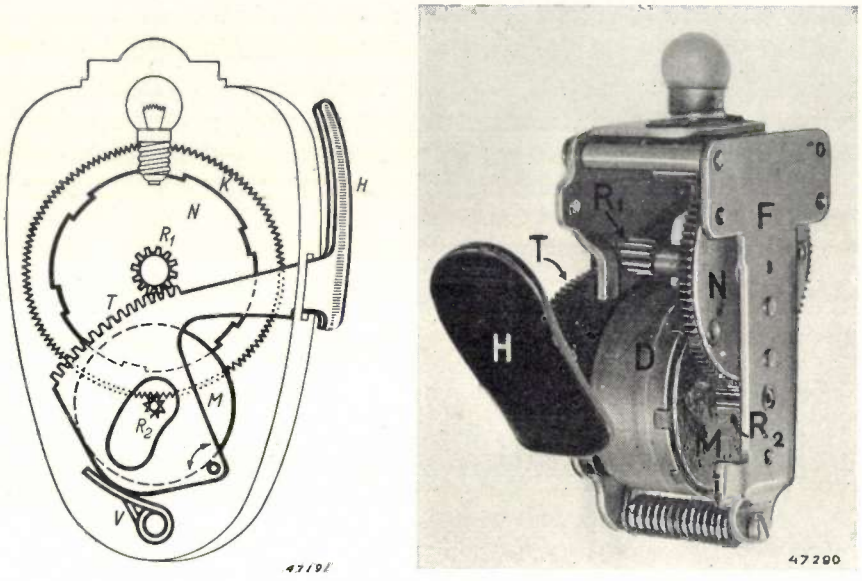
The construction of the dynamo itself is very similar to that of a bicycle dynamo. The magnetic field is produced by a permanent magnet in the form of a rotor, while the armature winding, in which an alternating e.m.f. is generated by the rotation of the magnet, is on a soft iron stator. With this principle of construction, which has become a practical possibility since magnet steels of very high magnetic power have become available, the collector with brushes and the separate excitation of the field found in ordinary dynamos are dispensed with, and these simplifications alone have made it possible to produce dynamos sufficiently small and reliable for the purpose.

The small lamp fed by the dynamo is mounted on the dynamo itself and the light emitted is concentrated by a small lens and reflector.

Since the rotor of the dynamo has to be driven by the same hand that holds the torch and directs its light, rotary drive is out of the question. A squeezing motion, either with the palm or with the thumb, is practically the only motion that can be reconciled with the other two functions of the hand.

In our case thumb pressure has been chosen. The mechanism of the motive power based on this principle is shown diagrammatically in *fig. 1a*. The lever, which is pressed in periodically against the action of a spring, is attached to a gear rack which sets the rotor of the dynamo in motion *via* a pinion and cogwheel transmission with freewheel. Thanks to the freewheel the rotor continues to move without hindrance when the lever springs back. The rotor, driving cogwheel and lever spindle bearings are in a frame entirely enclosed in the casing. Lubrication is provided by a small piece of cloth drenched with oil and clamped to the bearings. The photograph in *fig. 1b* shows how all the components are assembled in the frame (with the casing removed).

With "normal" squeezing the dynamo yields an average effective voltage of 2.5 volts and supplies the lamp with a power of about 0.3 watt. Losses due to friction, eddy currents, etc, make it necessary for the user to apply a mechanical energy of 0.9 watt.



a *b*

Fig. 1. *a*) Mechanism of the rotor in the Philips dynamo pocket torch type no. 7424. The lever *H* with the rack *T* is pressed in periodically with the thumb against the force of the spring *V*. *Via* the pinion R_1, *T* drives the freewheel *N*. When the lever is pushed in *N* carries with it the toothed rim *K*, which, *via* the pinion R_2, drives the rotor *M*, while upon the lever *H* being allowed to spring back the toothed rim and the rotor continue to turn freely. *b*) The dynamo pocket torch with casing removed. *F* frame, *H* lever with rack *T*, R_1 pinion, *N* cog wheel transmission with freewheel, R_2 pinion on the spindle of the rotor *M*, *D* dynamo.

[1]) In addition to the type described, a second, smaller dynamo pocket torch (the "Byou", type no. 7426) has also been developed by Philips. We shall, however, confine ourselves here to the first-mentioned type.

Since the lamp will usually only be in continuous use for relatively short times, a total lifetime of 10 hours is considered sufficient. The lamp therefore burns as it were on "overvoltage" for the sake of better light efficiency. The light flux obtained (always with "normal" squeezing) amounts to 2.4 lumens, half of which is irradiated from the lamp as effective light. In the beam proper at a distance of 1 m an illumination intensity of about 4 lux is obtained.

The overall dimensions of the torch are 8.5 × 5 × 3 cm, with a total weight of only 215 grams, so that it can very well be carried in the coat or trousers pocket or in a lady's handbag.

Mechanical details of the construction

A bicycle dynamo gives off about 3 watts electrical power and requires about 9 watts mechanical energy from the cyclist. The power of the pocket torch dynamo is therefore about 1/10 of that of a bicycle dynamo. This ratio is not unreasonable, considering the volumes of the muscles concerned in each case in the production of the mechanical energy and taking into account that the leg muscles have to drive not only the dynamo but also the bicycle [2]). A more direct method of obtaining some idea of the mechanical energy which can be demanded of the thumb is to calculate the squeezing power. In squeezing, the thumb can only make a stroke of a few centimeters and cannot repeat the squeezing motion more than a few times per second. If s is the stroke in cm and n the number of strokes per minute, the squeezing power P which the thumb must exert upon pressing in the lever in order to produce an average power of N watts is

$$P = 612 \frac{N}{s \cdot n} \text{ kg.}$$

In the case of the dynamo pocket torch described here the stroke at the extremity of the lever where the thumb acts is about 3 cm. From tests with different persons it has been found that a frequency of about 180 strokes per minute may be considered as normal. The force of the squeeze required for $N = 0.9$ W is then about 1 kg. Added to this is the force of about 0.5 kg necessary to overcome the pressure of the spring tending to push the lever back. With a total force of 1.5 kg one is clearly not far from the limit of what can be demanded of the thumb muscles: a much greater force of squeeze would in most cases lead too quickly to fatigue.

The transmission ratio of the cogwheels is so chosen that, at the normal frequency mentioned and with a ratio of about 1 : 1 between the times for pressing down the lever and its return, the rotor attains a speed of about 4000 r.p.m. Such a high speed is desired for the dynamo to function in a sufficiently flat part of its voltage—speed characteristic, a requirement which will be discussed later. Moreover, a high number of revolutions makes it possible to accumulate rotation energy in the rotor with a relatively low moment of inertia, thus small dimensions and light weight, so that in spite of the periodic interruption of the motive power the rotor continues to run with sufficient uniformity. Much higher speeds than the one mentioned can hardly be considered, because a second cogwheel transmission would then be required, which would reduce too much the mechanical efficiency and cause the noise of the mechanism to assume undesirable proportions. In order to minimize the noise as far as possible, the toothed rim on the freewheel is made of the well-known "hard tissue" frequently used for damping the sound of gearwheels, which consists of compressed layers of fabric impregnated with artificial resin ("Philitext").

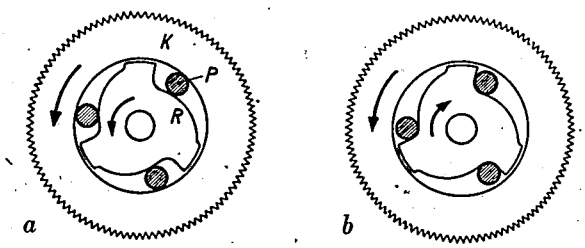


Fig. 2. Construction of the freewheel. Upon the hub being rotated in one direction the three rollers P roll along the face R of the three teeth upwards (a), become jammed between R and the toothed rim K and thus carry the latter along. When the hub is rotated in the opposite direction the rollers run down again (b) and no longer make contact with the rim K, which can thus continue to turn freely in the direction in which it started.

The construction of the freewheel (*fig. 2*), which periodically carries the toothed rim along with it and then allows it to run on freely, is also designed to produce little sound and, moreover, as little friction as possible during the freewheeling.

The desired moment of inertia of the rotor can be obtained by affixing a lead disc on the rotor spindle beside the permanent magnet. The whole must be well balanced in order to ensure smooth running of the rotor, without vibrations.

[2]) With an ordinary bicycle and moderately good tyres a power of, for instance, 30 to 100 watts, according to the condition of the road surface, in very unfavourable conditions even up to 200 watts, is necessary for a speed of 15 km/hr.

Electrical and magnetic details

The voltage—speed characteristic

At low speeds the terminal voltage of an A.C. dynamo under a certain permanent load resistance is proportional to the number of revolutions, but at higher speeds it becomes more and more independent of the number of revolutions.

This phenomenon and the measures which have to be taken to obtain the flattest possible voltage—speed characteristic have already been discussed in this periodical in connection with the description of a bicycle dynamo [3]). Since the same considerations apply to the pocket torch dynamo we shall go into them in somewhat more detail.

To begin with it must be explained in how far a flat characteristic is of importance for the torch dynamo. Though a cyclist usually rides at a fairly constant speed, this will sometimes be high (for instance 25 km/hr) and sometimes lower, while there are also cases where he dismounts and wheels the bicycle along a short distance (5 km/hr). A flat characteristic is therefore necessary for a bicycle dynamo in order on the one hand to obtain sufficient light from the lamp at 5 km/hr and on the other hand not to overload the lamp at 25 km/hr. In the case of the dynamo pocket torch the situation is somewhat different, for the average speed of squeezing will vary only slightly at different times, certainly not more than by a factor 2, compared with a factor 5 in the case of the bicycle dynamo. Thus for the same voltage variation a less flat characteristic would be sufficient. Here, however, we have a new factor: the variation in the speed of the rotor during the squeezing. Although an attempt is made to reduce these variations as much as possible by the above-mentioned enlargement of the moment of inertia, it is clear that a limit is set by the fact that the torch has to be kept small and light. The fluctuations in speed which may still remain are shown in *fig. 3* for a certain, rather unfavourable case. If it is desired to prevent these speed variations from causing a disturbing fluctuation of the light flux, the solution must again be sought in making the voltage—speed characteristic as flat as possible, especially because of the very small heat inertia of the filament, which is only 12 μ thick.

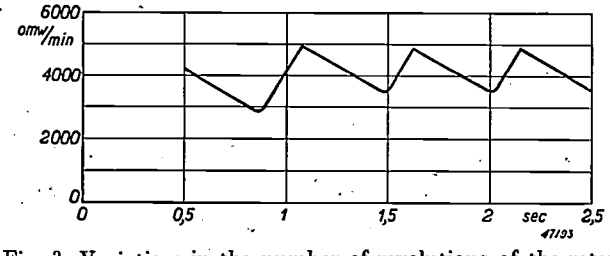


Fig. 3. Variations in the number of revolutions of the rotor caused by the intermittent drive, upon periodic squeezing in this case about 120 strokes per minute and a ratio of about 1 : 2 between the times of pressing in and the releasing of the thumb.

The terminal voltage becomes independent of the number of revolutions when the inductive resistance of the armature winding, which is proportional to the frequency of the A.C. voltage excited and thus to the speed of revolution, is large compared with the ohmic resistance of the armature winding plus lamp. Since the inductive resistance increases according as a higher flux is excited in the armature by a given current, it is desirable to keep the magnetic resistance of the armature circuit low. Furthermore, it is desirable to have a large number of poles on magnet and stator, since at a low speed the voltage generated still has a high frequency.

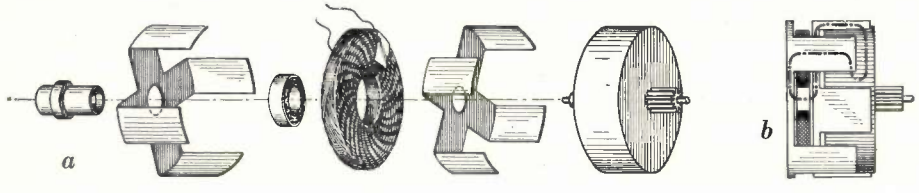
The magnet of the pocket torch dynamo has eight poles, so that at 4000 r.p.m. the frequency of the A.C. voltage generated amounts to about 250 c/sec. For a compact construction it is important that no actual projections need to correspond to these poles: the magnet (like that of a bicycle dynamo) consists of a short smooth cylinder, magnetized in eight poles along the periphery. Such a construction was very uneconomical with the magnet steels formerly available, for, as soon as the magnet is transferred from the magnetizing apparatus to the ultimate circuit, the flux produced by two neighbouring magnet poles then has a certain tendency to close inside the cylindrical shell through the steel, instead of acting outwards. In order to prevent this it is necessary that the magnet steel lying between the poles along the circumference of the cylinder should offer a high resistance to the reversal of the direction of magnetization thereby occurring. The magnet steel "Ticonal" 0.8 used for the pocket torch dynamo possesses this property to a high degree [4]).

Accompanying the property mentioned there is generally a low reversible permeability, *i.e.* the magnetization is little affected by small fluctuations of the magnetizing field. In particular, therefore,

[3]) H. A. G. Hazeu and M. Kiek, An alternating current dynamo with a flat characteristic for bicycle lighting, Philips Techn. Rev. **3**, 87, 1938.

[4]) As a closer consideration will show, the condition can be formulated more precisely as follows: the demagnetization curve, especially near the turning point of the magnetization, must have a small slope. For this purpose it is desirable that the steel, compared to the residuum, should have a large coercive force — a property which the older magnetic steels (tungsten steel) possessed to a much smaller degree; *cf*, the full-drawn and the dotted magnetization curves in fig. 6.

the armature reaction has little effect on the magnet, which promotes the stability of the whole. On the other hand, if we desire to obtain a flat voltage-speed characteristic, a low reversible permeability is an obstacle to the second requirement that the current through the armature winding must excite a strong counter flux. In as far as it has to pass through the magnet this field then experiences a fairly high magnetic resistance. The counter-field is therefore given the opportunity of completing itself for a large part outside the magnet, *i.e.* the magnetic circuit of the armature winding is made with a wide spread.



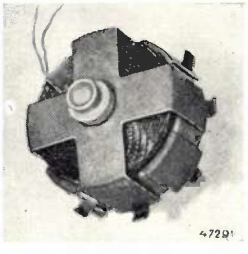


Fig. 4. *a*) The components of the dynamo ("exploded" view). The two cross-shaped stator pieces are rotated 45° with respect to each other and affixed to the spindle on either side of the armature coil, which is passed over a cylindrical core. The bent-down ends of the two crosses form the eight poles of the stator and enclose a cylindrical space in which the magnet (extreme right) rotates.
b) The magnetic circuit assembled. The dot-dash line represents the path of one line of force.
c) View of the finished circuit.

The form of the magnetic circuit, which deviates very much from the usual form, may be seen in *fig. 4*. It is formed by two cross-shaped pieces of soft iron sheet with the ends of the arms bent over at an angle of 90°, and a cylindrical soft iron core over which the armature winding is placed. The two crosses are fastened against the ends of the cylindrical core. The inner cross has slightly shorter arms than the outer one. Due to the fact that the two crosses are rotated 45° with respect to each other, the four bent ends of one cross fall between those of the other, and the eight ends, which function as stator poles, together enclose a cylindrical space in which the magnet rotates. Armature coil and magnet in this peculiar circuit are thus arranged as co-axial cylindrical discs. In fig. 6*b* the course of the lines of force can be followed with the help of the example given; fig. 6*c* is a photograph of the circuit fully assembled.

It is clear that this circuit will have the desired large spread, considering the great length over which it travels between neighbouring stator poles adjacent to each other at such relatively short intervals.

In *fig. 5* the voltage—speed characteristic is reproduced, which is obtained with the construction described. Thanks to its flatness, in spite of the fluctuations in the speed during the periodic squeezing only a slight flicker occurs in the light of the lamp, which is scarcely perceptible.

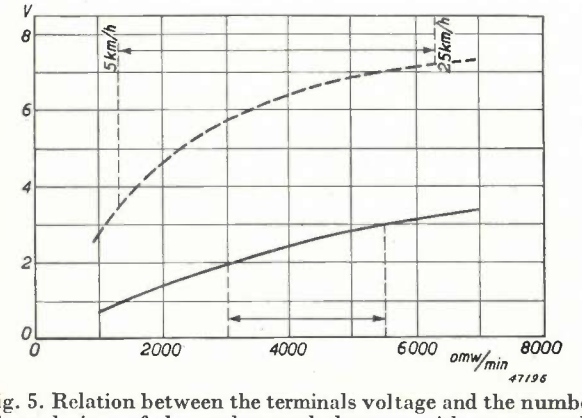


Fig. 5. Relation between the terminals voltage and the number of revolutions of the pocket torch dynamo with permanently connected lamp for 2·5 V, 0·1 A. The dotted line represents the corresponding relation for the Philips bicycle dynamo type no. 7722 equipped with a lamp for 6 V, 0·5 A.

Owing to its peculiar construction the magnetic circuit is specially adapted to the requirements of mass production. An important point, for example, is that the armature winding can be prepared in advance on a winding machine in the form of a simple disc-like coil, to be passed over the core before the two cross-shaped pole plates are put on.

The fact that the magnet is constructed as a smooth cylinder without projecting poles is also very important for mass production. The magnet steel, cast in the desired shape and hardened, is so hard and brittle as to allow of practically no other processing than grinding. The prescribed simple cylindrical shape makes it possible for the grinding to the precise dimensions of the magnet to be carried out by a method particularly suitable for mass production, namely on a so-called centerless grinding machine, which process costs relatively little in time and energy.

Safety measures against short-circuiting.

Due to the wide spread of the magnetic circuit, the necessity of which was explained above, magnetic energy is "spilt". Thanks to the high output of magnetic energy per unit of volume of the magnet steel "Ticonal" 0.8 this extravagance is permissible, without it being necessary to have recourse to too heavy a magnet, in order to excite the necessary flux in the armature. It is even possible to go farther and to sacrifice still more of the magnetic energy to make the dynamo insensitive

to short-circuiting, as will be seen from the following.

Fig. 6 gives the demagnetization curve of the magnet steel, *i.e.* that part of the magnetization curve which gives the induction (in gauss) as a function of a demagnetizing field (in oersted). Given the magnetic resistance of the circuit, the relation between the induction and the field is represented by a straight line through the origin,

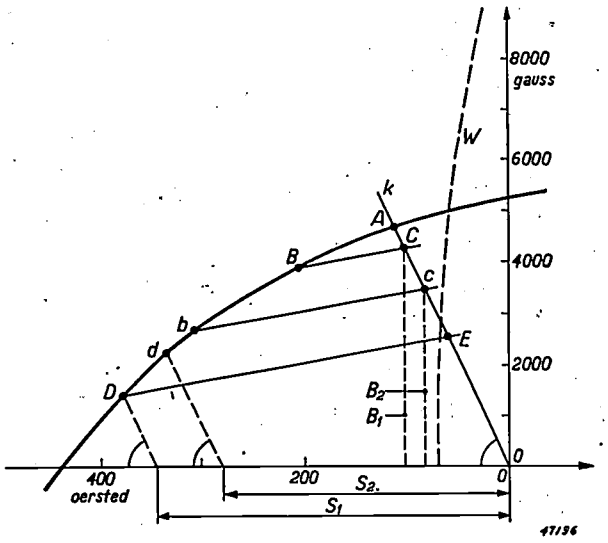


Fig. 6. Demagnetization curve of the magnet steel "Ticonal" 0·8 used in the pocket torch dynamo. In ordinary use the working point of the magnet steel (of a given part in the poles) runs for example back and forth on the line *BC*. Due to a temporary short circuit this point would fall to *D* through the strong counter-field S_1. When the magnet is demagnetized in advance to the working line *bc* and is given correspondingly larger dimensions to obtain the same flux, a weaker counter-field, S_2, occurs upon short-circuiting (where $S_2 : S_1 = B_2 : B_1$), so that the working point only falls to *d*. The dotted curve *W* represents, for the sake of comparison, the demagnetization curve of tungsten steel.

for instance *k* in fig. 6. At the point of intersection *A* of this line with the curve the induction may be read off which the steel will retain when, after magnetization to saturation, it is placed in the circuit in question. When, due to the rotation of the magnet, a current begins to flow in the armature winding, it will excite a counter-field, which causes the induction to fall from *A* to *B*; upon the current falling to zero the induction increases again, not along the magnetization curve but along the practically straight line *BC*, the slope of which is equal to the previously mentioned reversible permeability. In the state of rest the working point remains at *C*, while when the torch is functioning it runs back and forth along the line *BC* (strictly speaking it describes a very narrow loop).

This is true, however, only as long as the counter-field due to the current taken off does not become stronger than corresponds to point *B*. If a short circuit occurs, when a much larger current flows for a very short time, the induction falls for instance to *D*, the short circuit is ended and the induction ascends along the line *DE* instead of returning to *C*. A new working point *E* is then reached, which may have a much less favourable position. This means that after one short-circuiting the dynamo will no longer produce sufficient voltage, unless the magnet is magnetized anew by means of special devices required for that purpose.

This disastrous effect of a short-circuit can be eliminated in a simple manner by using a somewhat larger magnet and demagnetizing it slightly when assembling the dynamo, for instance to point *b* in *fig. 6*, so that the working point lies at *c* [5]). The magnet has to be so much larger as to produce the same flux as before, since under otherwise equal conditions the flux determines the A.C. voltage generated. Also the current upon short-circuiting occurring is the same, but the counter m.m.f. thereby set up is distributed over a greater length of magnet and the demagnetizing field (S_2 in fig. 6) is thus proportionally smaller. The point *d* reached when short-circuiting occurs will therefore not lie much lower than *b* and in any case much higher than *D*. With a suitable choice of the initial state *c*, the working point may even in the event of a short-circuit still remain on the recoil line *bc*, so that the short-circuit then causes no permanent change of any kind.

The method described is employed in the pocket torch dynamo in order to exclude the possibility of it being rendered unserviceable by a single temporary short-circuit, although there is very little chance of short-circuiting thanks to the completely closed construction of the torch. The slightly heavier weight of the magnet constitutes no objection, especially since it could be set against the above-mentioned lead disc added to increase the moment of inertia. In this case it would have been of practically no advantage to reduce the amount of magnet steel by substituting one of the still better kinds of magnet steel for the "Ticonal" 0.8 [6]).

---

[5]) Point c will only lie on the same line *b* when the air gap is enlarged in the same proportion as the magnet. In practice this will not be done; the considerations become somewhat more complicated but the conclusions remain the same.

[6]) See, for example, B. Jonas and H. J. Meerkamp van Embden, New kinds of steel of high magnetic power, Philips Techn. Rev. 6, 8, 1941.

THE VITREOUS STATE

by J. M. STEVELS. 539.213.1

The vitreous state of matter is the solid state of aggregation which occurs when a liquid solidifies without crystallizing. In this article the factors are discussed which exert influence on the tendency of a substance to solidify in the vitreous form. The most important cause of the formation of a glass is an association of the atoms of the liquid which leads to larger and larger complexes upon decreasing temperature of the liquid. Under certain circumstances, discussed in this article, these complexes have practically no tendency to pass over into a crystalline structure. The mass then solidifies in such a way that the arrangement of the atoms, considered over short distances, exhibits much similarity with that in a crystal, while over distances of more than a few atoms there is no regular arrangement. In the case of the glass-forming oxides a detailed study is made of the characteristics of their structure responsible for their tendency to form glasses, following in the main the conceptions of Zachariasen. In conclusion a discussion is given of the inorganic glasses formed by fusing together different oxides. It is shown that a glass can best be composed of a suitable acid oxide as glass-former and basic oxides which lower the temperature of fusion.

Most substances are known in three states of aggregation: gaseous, liquid and solid. In the gaseous state the molecules of the substance have no mutual cohesion, so that a given quantity of it can fill a space of any given form and size. In the liquid state the volume of a quantity of substance is almost independent of pressure and temperature. The form of the space filled, however, is still arbitrary, while in the solid state the form is also fixed.

Given such a description the question immediately arises as to whether it is clear in every case what state of aggregation one is dealing with. If, for example, a liquid is very compressible, its behaviour under high pressure does not differ in character from that of a gas. If, on the other hand, a liquid has a high viscosity, so that it accommodates itself only slowly to the shape of the container, doubt may be felt whether one is not dealing with a solid. It is well known that a gradual change of shape (plasticity) is possible in a solid under sufficiently high pressure. Thus if the states of aggregation of matter were distinguished merely by the characteristic properties just outlined, the state of matter could always be considered more or less as a transitional case between different states of aggregation.

The fact that the states of aggregation of the substances occurring in nature are usually unambiguously determined is thus actually a remarkable phenomenon. The reason for this fact by no means lies in the properties which the substance possesses in each of those states of aggregation, but in a phenomenon which has not yet been mentioned, namely that of discontinuous phase transitions. At very definite temperatures the substance suddenly changes its structure, and with the absorption or giving off of heat it passes over from one state of aggregation to another. The range of existence of the states of aggregation is sharply limited by theses phase transitions, so that all uncertainty is removed.

If we now consider in particular the transition between the liquid and the solid substance, the change in structure consists briefly in the fact that the molecules in the liquid state exhibit a more or less random arrangement, while the solid substance has a definite crystalline structure. Upon further cooling of the solidified material this structure may change one or more times (again with the giving off of heat). It is thus not in principle necessary that a phase transition should be accompanied by a change in the state of aggregation.

The vitreous state

In the case of many substances it is observed that upon cooling the liquid mass solidifies gradually, without there being any question of a phase transition. The viscosity of the liquid increases steadily with decreasing temperature until the deformation of the mass has finally become so difficult that the substance must be called solid. Conversely, upon heating such substances above a certain temperature they do not suddenly become liquid, but only very gradually. Apparently no sudden change of structure occurs during the softening or solidifying.

A substance which behaves in this way is called a glass in the solid state. From X-ray diffraction photographs of glasses it is indeed evident that the atoms exhibit the same random arrangement in

the solid state as in the liquid state. This may be due to various causes. In the first place it might be held that the molecules or parts of molecules from which the glass is built up are unsuitable for the composition of a regular crystalline structure. This, however, is a practically unknown phenomenon, at least in inorganic chemistry. It is more likely that the crystalline state does indeed exist but that the molecules are too much hindered in their movements during the solidification to achieve this regular arrangement. Glass is therefore called a supercooled liquid. By this is meant that upon cooling the liquid has indeed passed through the temperature zone of a phase transition but that the phase transition has not taken place. The vitreous state is thus unstable compared with the crystalline state. In certain cases it has actually been observed that the glass begins to crystallize, which is called devitrification.

For the sake of completeness attention must here be called to the difference between glasses and amorphous substances which also possess a random molecular structure but which cannot be considered as supercooled liquids. Certain substances, for example, are obtained in the amorphous state by condensing them from the vapour on a cold surface. The atoms deposited then lose their freedom of movement all at once due to the cooling, and remain at the spot where they happened to be deposited.

Very many substances are obtained in the amorphous state when they are precipitated rapidly from a solution and at not too high a temperature (for example metal hydroxides or oxide hydrates). Such an amorphous substance is also formed by the electrolytic oxidation of aluminium in certain solutions, whereby a porous, very adhesive covering layer of Al_2O_3 is formed on the metal (technically known as eloxal film). Upon examining this surface layer with X-rays no crystalline structure is detected. This means that any possible order of the atoms cannot be more than a few atomic distances.

When does the vitreous state occur?

As already stated, the cause of the occurrence of a vitreous state may in general be sought in the fact that the molecules of the liquid are too much hindered in their movements to attain the crystalline state. The fact that the molecules of a substance which solidifies as a glass are actually less mobile than those of other substances may be concluded from the fact that the molten glasses usually possess a high viscosity. It is very possible that the "molecules" of the glasses in the liquid state are already joined to larger aggregates, which, because of their shape or position, are not capable of being combined into a crystalline structure; crystallization then requires a complete regrouping of the atoms, whereby here and there bonds in the "molecules" must be broken. The process therefore requires a very long time and can only take place completely upon very slow cooling.

A striking example is furnished by the behaviour of sulphur, and especially of selenium, which, when the molten substance is not too slowly cooled, is easily obtained in the vitreous form. Since each Se atom can combine with two other Se atoms by means of strong valence forces, at a relatively high temperature in molten selenium, chains

Se - Se - Se - Se - Se - Se

are already formed, which become longer and longer with decreasing temperature. As a result the viscosity of the molten selenium increases rapidly as the temperature is lowered. The forces between neighbouring chain molecules, in contrast to the forces inside a chain molecule, are relatively weak, so that there is little tendency towards crystallization. Only at a fairly low temperature, at which the mass is already extremely viscous, is the state reached where the molecules would take on a regular arrangement. In the meantime, however, the possibility of motion of the chains in the tangled mass has become so slight that this arrangement can only take place very gradually.

If the decrease in temperature takes place fairly rapidly the viscosity quickly reaches such a value that the substance may be considered as solid. The mobility of the molecules has then become practically zero and regular arrangement is then quite impossible.

A type of behaviour similar to that of selenium is encountered in numerous organic substances, for example ethanol (ethyl alcohol), glycerol, glucose, cane sugar and various alkaloids such as brucine and cocaine. In the cases of these substances also the molecules probably join in the liquid state to associated complexes, so that there is little tendency towards crystallization.

In the case of the glasses proper, which are obtained by the fusing of certain oxides or mixtures of oxides, the analogy of their behaviour to that of selenium is less obvious, and for a thorough understanding of the vitreous state it is therefore desirable to investigate the structure of the glass-forming oxides more closely. In the discussion we shall be guided by the work of Zachariasen [1], who showed that the formation of a glass is connected with certain specific properties of the crystal structure. Zachariasen's rules are found to furnish a satisfactory picture of the behaviour of the oxides, and they may with success be used to prophesy whether

[1] W. H. Zachariasen, J. Amer. Chem. Soc. 54, 3841, 1932; J. Chem. Physics 3, 162, 1935.

a given compound possesses a tendency to form a glass.

Structure of vitrifying oxides

The most suitable components of glasses, according to general experience, are the oxides of elements such as boron and silicon, which furnish small atoms with a high charge. We may consider these oxides to be built up in the first instance of negative oxygen ions (O^{2-}) and the very much smaller positive ions of boron or silicon (B^{3+} or Si^{4+}). In these cases the force of attraction of the central ion is so great that the "charge cloud" of the neighbouring negative ions is drawn towards that central ion. A treatment which chooses the centre of the negative ions as the position of the negative charge cannot, therefore, lead to quantitatively correct results.

It is, however, very useful for our purpose, *i.e.* for finding the relation between the structure of the vitrifying oxides and their tendency to form a glass.

Just as in the case of selenium, we shall first examine how with falling temperature the particles will combine to larger aggregates. We shall try to explain the way in which this takes place on the basis of two examples, namely from the structure of crystobalite, the crystalline modification of SiO_2 which is stable at high temperatures, and from that of B_2O_3. The structure of crystobalite, which is well-known from X-ray analysis, is shown in *fig. 1*.

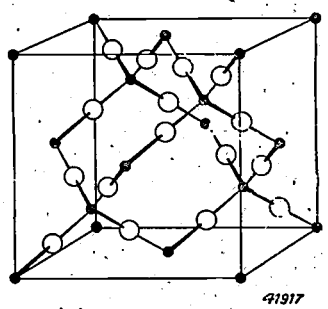


Fig. 1. Crystal structure of crystobalite. Black dots, silicon ions (Si^{4+}); white circles, oxygen ions (0^{2-}). Each silicon is surrounded by a tetrahedron of oxygen ions.

It may be seen that here every Si^{4+} ion is surrounded by four oxygen ions arranged in a tetrahedron. Each of these oxygen ions, however, is not only bound to one positive ion, but at the same time forms a part of the surroundings of a neighbouring positive ion. In this way the structure can be extended in all directions, so that the whole crystal lattice is held together by valence forces.

In the case of boron trioxide the boron ion is surrounded by only three oxygen ions. The structure of boron trioxide and that of silicon dioxide thus contain triangular and tetrahedral groups of ions, respectively, which are shown in *fig. 2*. It may be seen that the space between the oxygen ions only

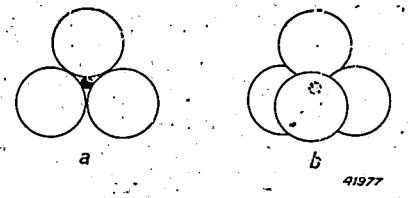


Fig. 2. The most commonly occurring ion groups in the crystal structures of the oxides of glass-forming ions:
a) triply charged positive ion with three surrounding oxygen ions (the case of B_2O_3),
b) quadruply charged positive ion with four surrounding oxygen ions (the case of SiO_2).

offers room for an ion of much smaller dimensions and it is therefore obvious that these structures will only occur in the presence of very small positive ions.

Considered purely geometrically, the silicon ion or the boron ion could be also surrounded by a larger number of negative oxygen ions, for instance six. Since the interstice between 6 negative ions, however, is appreciably larger than the highly charged central ion itself, this coordination is not stable from the point of view of energy; due to the attraction by the positive charge the negative ions are drawn as close as possible to the central ion, and this results in a number of negative ions being pushed out of the first shell around the central ion. The result is the grouping of the oxygen ions in triangles or tetrahedra already mentioned.

The way in which these groups of ions can combine to a crystal is shown in *fig. 3*, in which a number of regular structures — for the sake of simplicity plane structures — are reproduced which are derived from the triangular group.

The distinctive difference between fig. 3*a* on the one hand and figs. 3*b* and 3*c* on the other is that in case *a*) each negative oxygen ion belongs to three adjacent groups, while in cases *b*) and *c*) only two neighbouring groups are joined by each oxygen ion. Which of these two possibilities will be realized depends upon the charge of the positive ion. If one assumes that the positive ion — like the oxygen ions — is doubly charged, it is evident that the negative charge of the ion triangle is three times as large as the positive charge of the central ion. A neutral whole can then only be obtained by having each oxygen ion belong to three different groups, so that a structure like fig. 3*a* will be formed. If, however, the positive ion has a triple charge,

the negative charge of the ion triangle is only twice as large as the positive charge, so that every oxygen ion will be coupled to only two positive ions. Rarefied structures of the nature of figs. 3*b* and *c* then result.

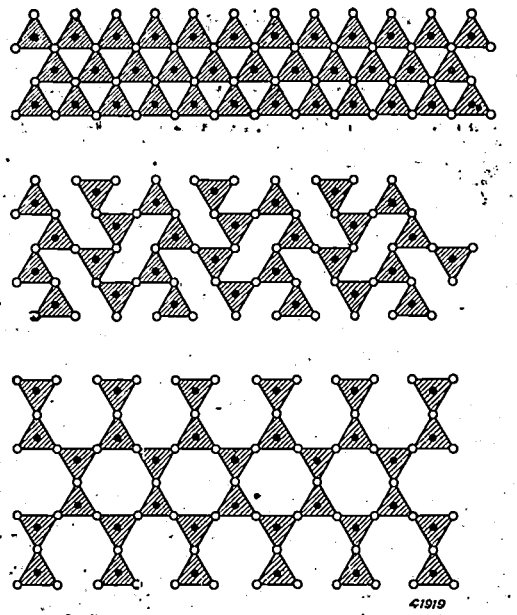


Fig. 3. Diagram of the way in which crystal structures can be formed from triangular ion groups in a flat plane. In case *a*) three ion groups come together at each angle, in the cases *b*) and *c*) only two ion groups. Case *c*) differs from case *b*) by the more symmetrical arrangement of the positive ions and by a more rarified structure.

The fact that the presence of highly-charged positive ions leads to a rarefied structure may seem paradoxical, but it can be explained in principle in the following way. If we are concerned with polyhedra consisting of m doubly-charged negative oxygen ions ($m = 3$ for a triangle, $m = 4$ for a tetrahedron), each of which "belongs" to n positive ions (which bear a charge p), then for the sake of electrical neutrality $2/n = p/m$. Thus for a given kind of polyhedron p, n is a constant, in other words a high charge p on the central positive ion results in a low value of the coordination number [2]) n of negative ions, and thus in a rarefied structure.

It might be imagined that the structure *b*) is more advantageous from the point of view of energy than *c*); it has indeed a greater density and thus yields more to the attractive forces between the particles. This opinion, however, is found to be incorrect. A structure like 3*b* is only possible when each positive ion is surrounded by other positive ions in an only partially symmetrical manner, so that the mutual distances between some of these ions are small. Now it is just these neighbouring positive ions which are not completely screened from each other by the surrounding negative charge. They will therefore repel each other strongly, which makes the structure less stable. The result is that the completely symmetrical configuration according to fig. 3*c* is more favourable than 3*b* from the point of view of energy. This can also be expressed by saying that the repulsive forces between the highly charged positive ions "swell" the lattice until a symmetrical, although also a rarefied, structure is obtained.

The same forces which are active in the crystalline state are also already active in the liquid state of the oxides in question. Perfectly regularly arranged lattice structures are then not yet present, but the first surrounding of every positive ion by negative ions will not differ much from that in a crystal. Here also the electrostatic forces of the ions will result in ion groups joining together to larger aggregates by means of common oxygen ions. As far as the mutual connection of neighbouring groups of ions is concerned, these aggregates will exhibit much similarity with the final crystalline structure; over longer distances, however, the regular arrangement will not be maintained.

For the crystal types of figs. 3*a* and *c* the arrangement in the liquid state is now shown diagrammatically in *fig. 4*. In the attempt to construct such a figure it is immediately obvious that these two cases

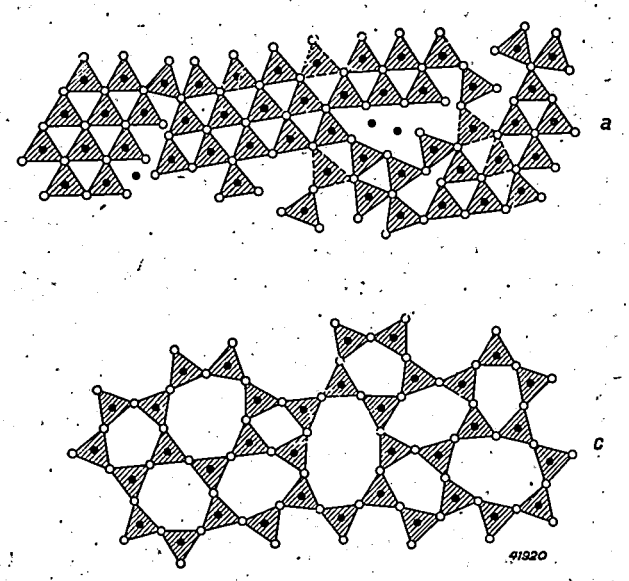


Fig. 4. Diagram of a liquid in which association occurs according to the structures of fig. 3*a*) and *c*), respectively. In case *a*) if there is no completely regular arrangement there must be unsaturated valences; in case *c*) on the other hand the non-regular state differs from the regular state only by slight differences in the valence angles. Therefore in case *c*) there is little gain in energy to be obtained by making a regular arrangement.

[2]) This term is generally used to indicate the number of immediate neighbours (atoms or ions) possessed by a given atom (or ion).

behave quite differently: in the case of lattice type *a*) one has only the choice between the perfect crystal lattice and a structure in which the bonds between neighbouring ion groups are partially broken, whereas in the case of type *c*) quite irregular structures may be formed by slight changes in the valence angles (which always occur due to thermal agitation). Since in the case of *ionic bonds* the valence forces possess no pronounced preferred direction, the irregular structures mentioned are practically as favourable from the point of view of energy as the ideal crystalline structure.

From these considerations it follows that substances with lattices of the type *c*) exhibit a behaviour like that of selenium upon solidifying. Since the binding forces between the ions are very large, association will already occur at a high temperature, and it will lead to a considerable increase of viscosity with decreasing temperature. The associated groups have then indeed no crystalline character as yet, but as far as energy is concerned they are not much less favourable than the crystal structure, so that their tendency towards crystallization is relatively small. Crystallization therefore only begins at such a low temperature that the viscosity of the material very much retards it. When the cooling is not very slow crystallization fails entirely to take place and a glass is formed.

Generalization of the condition for vitrification

In the above the vitrification of the oxides has been explained by the possibility of forming irregular structures in which the valence forces — except for a small change in the valence angles — are just as effective as in the crystal structure. We have already seen that rarified structures best satisfy this condition; in particular it was found that vitrification was impossible in the case of a structure of the type of fig. 3*a*.

The general formulation of the characteristics which the structure of an oxide must possess in order to be able to form a glass is in this way reduced to a geometrical problem. This problem is quite complicated since it must be treated not in a plane but in three dimensions. We shall not go into the details here, but only summarize the practical results in the following four rules.

1) Every oxygen ion must be bound to not more than two positive ions which must be highly charged and small.
2) The number of oxygen ions which surround such a positive ion may not be too large or too small (3 or 4).
3) The ion groups adjacent to each other may have common corners but no common edges or faces.
4) There must be ion groups which have at least three oxygen ions in common with neighbouring ion groups.

Some of these rules (especially the first) already follow from the discussion given by us of the plane case. Others can be verified by a consideration of spatial structures.

These conditions, which were discovered by Zachariasen, are always found to be satisfied in the case of the glass-forming ions, and conversely it may be said that every oxide which satisfies them is able to form a glass. These considerations can also be applied to other compounds related to the oxides. Thus, for example, it follows for the fluorides that only beryllium fluoride can be obtained in the vitreous state, which is actually found to be confirmed. As an illustration the structures of BeF_2 and CaF_2 are shown in *fig. 5*, and it is

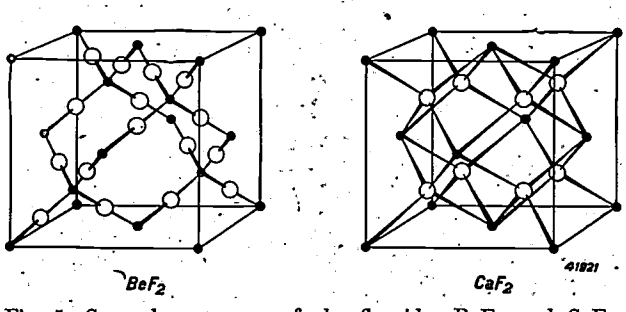


Fig. 5. Crystal structures of the fluorides BeF_2 and CaF_2. White circles F–ions, black dots Be^{2+} and Ca^{2+} ions, respectively. In the first-mentioned structure, which is identical with that of crystobalite, the conditions for the formations of a glass are satisfied, which is not the case with the last-mentioned: the ion groups are in contact not only at corners but over whole surfaces, and each fluorine atom belongs to more than two ion polyhedra.

evident that the last mentioned compound does not satisfy the first three conditions of Zachariasen but satisfies the fourth.

The fluorine ions here play the same part as the oxygen ions in the case of oxides.

Survey of vitrifiable oxides

In the foregoing we have seen that the rarified structures necessary for the formation of a glass can only be obtained when the positive ions have a small diameter and a high charge. In the case of the triangular O^{2-} groups it was found that the positive ion must be at least triply charged; for tetrahedral O^{2-} groups, which occur much more often in three-dimensional structures, a minimum charge of four follows in the same way. We may therefore expect that the elements of the fourth and fifth columns of the periodic system of the elemcnts will be the most suitable as "glass formers".

An experimental investigation has confirmed this expectation. In *fig. 6a* the elements are indicated which form glasses according to experience and also according to Zachariasen's rules. The position of the cations in the periodic system (fig. 6*b*) agrees perfectly with the expectation expressed.

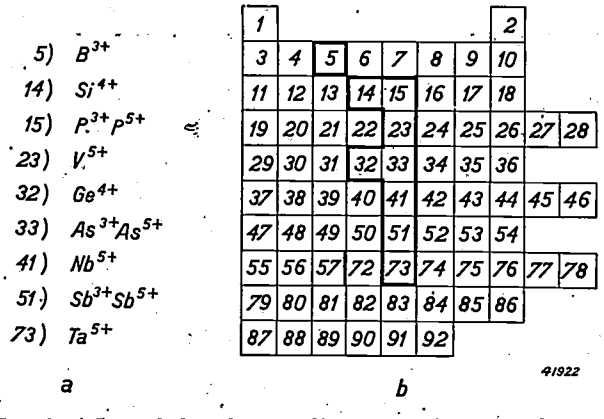


Fig. 6. *a*) List of glass-formers, *b*) position of the glass-formers in the periodic system of the elements. The glass-formers are to be found chiefly in the fourth and fifth columns.

The most important glass-forming element is silicon. Its oxide SiO_2 occurs in four modifications which are stable in different temperature regions. In the order of increasing stability temperature these modifications are β-quartz, α-quartz, tridymite and the already mentioned crystobalite.

In all the modifications the silicon ion is surrounded by a tetrahedron of oxygen ions, but there is a difference in the way in which the tetrahedra are grouped around each other. In addition to the crystalline modifications there is also a vitreous modification, so-called quartz glass. From X-ray diffraction photographs it has been found that this vitreous SiO_2 has a crystobalite-like structure. This structure is actually nothing else than the three-dimensional extension of the rarified plane structure which was built up in fig. 3*c* with the help of triangles. Just as with that structure, it is possible by means of slight changes in the valence angles to obtain an arrangement which, considered over long distances, is without regular order and represents a vitreous state. In the case of other glass-forming ions of the fourth and fifth columns of the periodic system similar structures are obtained consisting of tetrahedral ion groups.

In vitreous B_2O_2 every B^{3+} ion is surrounded by three O^{2-} ions, in other words the structure is composed exclusively of triangles [3]). For the sake of completeness, however, it should be noted that in most borate glasses, in addition to these triangular O^{2-} groups of ions, tetrahedral groups also occur.

Inorganic oxide glasses

If a glass is made up of the oxides of different elements, as is often done in order to obtain a lower melting temperature, only one of the oxides need be a glass-forming oxide. The cations of the other oxides take up positions in the large cavities of the rarified structure and can thereby make solidification more difficult. As a result the temperature of fusion becomes lower and the chance of forming a regular crystalline structure smaller. As an example we may consider the glasses which are formed when positive metallic ions, such as calcium, barium, lead, etc., are taken up in the cavities of the silicon oxide network. Since the whole must remain electrically neutral, the number of negative oxygen ions must of course also be increased. This is done by breaking the connection between neighbouring oxygen tetrahedra at a number of spots, a bridging oxygen ion of two neighbouring groups being replaced in each case by two non-bridging oxygen atoms. The structure which results in this way is shown in *fig.* 7, a plane diagram being given for the sake of simplicity. The grey circles of different sizes represent different metallic ions. By varying quantities and combinations of these ions it is possible to obtain many thousands of different glasses with very divergent properties.

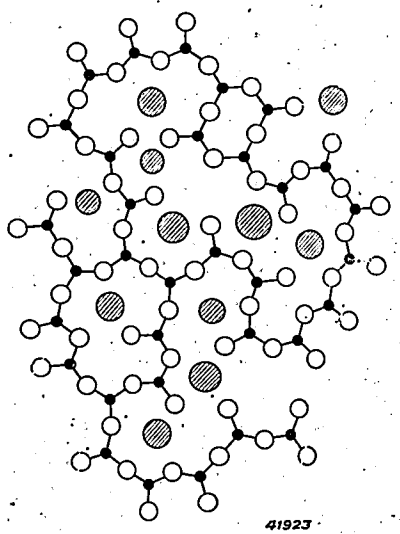


Fig. 7. Schematic representation (in a plane) of the structure of a glass consisting of a glass-forming oxide and different metallic oxides. Black dots positive ions of glass-former, white circles oxygen ions, grey circles different metallic ions. (Borrowed from B. E. Warren and A. D. Loring, J. Amer. Chem. Soc. 18, 269, 1935).

[3]) The spatial network of vitreous B_2O_3 consists of "two-dimensional" structures of the type of fig. 4*c*, which lie practically at random with respect to each other but between which valence bonds are present.

Compared with the glass-forming ions the metal ions possess a smaller positive charge and a much greater diameter. This means, therefore, that the glass-forming ions possess a much stronger electric field on their surface. It is remarkable that this difference is also manifested chemically in a very typical way: the glass-forming oxides behave as acid oxides, while the metallic oxides behave as basic ones. In general it is found possible to consider the field at the position of the centre of the adjacent oxygen ion as a measure of the "acidity" of a positive ion [4]). One may then say that a glass can best be composed of an acid oxide as the glass-former proper and one or more basic oxides as fluxes, the cations of which must fill the cavities in the rarified structure.

What will happen if, as admixture with the SiO_2, metallic oxides are used of increasingly weaker basicity or increasingly stronger acidity? As long as the admixture is limited to the genuinely basic oxides (Na_2O, CaO, PbO, etc.) the vitreous state is tenable. If one continues in the series, however, the cavities of the glass structure will then be more and more strongly drawn together by the forces of attraction of the metallic ions, until finally the glass structure becomes untenable. A demixing of the metal oxide and the SiO_2 then occurs: each of the two kinds of positive ions surrounds itself with oxygen ions according to its own coordination number and forms a corresponding characteristic lattice structure, with as result a strong tendency to devitrification. This case occurs, for example, when titanium oxide (TiO_2) is added to SiO_2. When one passes on to the oxides of still higher acidity, one arrives at the actual glass-forming oxides, such as CeO_2, B_2O_3, P_2O_5. The tendency towards devitrification now disappears and a homogeneous vitreous structure is obtained again. Apparently these glass-forming ions resemble the silicon ion so much in coordination number and valence that they can form a common glass structure with it.

[4]) See A. Dietzel, Naturwiss. 23, 537, 1941; Z. Elektrochem. 48, 8, 1942.

A DIRECTING INSTRUMENT FOR THE OPERATIVE TREATMENT OF FRACTURES OF THE NECK OF THE FEMUR

by G. J. van der PLAATS and A. VERHOEFF. 616.718.42-001.5-089.2

A fracture of the neck of the femur is usually treated operatively, the parts of the broken bone being joined firmly together by hammering in a hollow steel nail after a guiding needle has been bored into the bone. In performing such an operation the surgeon is faced by the problem of directing the guiding needle in such a way as to obtain exactly the desired position of the pin, which has been determined in advance by means of X-ray photographs. Philips has constructed a directing instrument for this purpose which, compared with other apparatus for the same purpose, has the advantage, *inter alia*, that the necessary adjustments and settings indicated by the X-ray photographs can all be made in advance of the actual operation. The operation is thereby considerably shortened, while in addition, except for the carrier of the guiding needle, the instrument need not be sterilized.

Conservative and operative treatment of fractures

In general the surgical treatment of bone fractures may be confined to reduction of the fracture, making every attempt to fit the fragments into their proper places. If this reduction is successful and the position remains correct, which should be checked by X-ray examination, the body itself provides for the further repair by the formation of new connective tissue which later calcifies (callus formation) and makes a firm joint (consolidation).

There are, however, cases where this so-called closed or conservative treatment of a fracture is inadequate, either because there is insufficient callus formation or because this takes place so slowly that the patient has to be confined to bed for months, with the danger, so familiar to doctors, of secondary pulmonary or other disorders. A classic example of this is the fracture of the neck of the femur (fractura colli femoris). This fracture occurs most often with eldery people due to a fall on the side, and since callus formation in the femoral neck generally leaves much to be desired and the fracture in question naturally makes the use of the legs impossible, with the closed treatment healing, if occurring at all, requires a very protracted immobilization of the patient.

Compared with the glass-forming ions the metal ions possess a smaller positive charge and a much greater diameter. This means, therefore, that the glass-forming ions possess a much stronger electric field on their surface. It is remarkable that this difference is also manifested chemically in a very typical way: the glass-forming oxides behave as acid oxides, while the metallic oxides behave as basic ones. In general it is found possible to consider the field at the position of the centre of the adjacent oxygen ion as a measure of the "acidity" of a positive ion [4]). One may then say that a glass can best be composed of an acid oxide as the glass-former proper and one or more basic oxides as fluxes, the cations of which must fill the cavities in the rarified structure.

What will happen if, as admixture with the SiO_2, metallic oxides are used of increasingly weaker basicity or increasingly stronger acidity? As long as the admixture is limited to the genuinely basic oxides (Na_2O, CaO, PbO, etc.) the vitreous state is tenable. If one continues in the series, however, the cavities of the glass structure will then be more and more strongly drawn together by the forces of attraction of the metallic ions, until finally the glass structure becomes untenable. A demixing of the metal oxide and the SiO_2 then occurs: each of the two kinds of positive ions surrounds itself with oxygen ions according to its own coordination number and forms a corresponding characteristic lattice structure, with as result a strong tendency to devitrification. This case occurs, for example, when titanium oxide (TiO_2) is added to SiO_2. When one passes on to the oxides of still higher acidity, one arrives at the actual glass-forming oxides, such as CeO_2, B_2O_3, P_2O_5. The tendency towards devitrification now disappears and a homogeneous vitreous structure is obtained again. Apparently these glass-forming ions resemble the silicon ion so much in coordination number and valence that they can form a common glass structure with it.

[4]) See A. Dietzel, Naturwiss. 23, 537, 1941; Z. Elektrochem. 48, 8, 1942.

A DIRECTING INSTRUMENT FOR THE OPERATIVE TREATMENT OF FRACTURES OF THE NECK OF THE FEMUR

by G. J. van der PLAATS and A. VERHOEFF. 616.718.42-001.5-089.2

A fracture of the neck of the femur is usually treated operatively, the parts of the broken bone being joined firmly together by hammering in a hollow steel nail after a guiding needle has been bored into the bone. In performing such an operation the surgeon is faced by the problem of directing the guiding needle in such a way as to obtain exactly the desired position of the pin, which has been determined in advance by means of X-ray photographs. Philips has constructed a directing instrument for this purpose which, compared with other apparatus for the same purpose, has the advantage, *inter alia*, that the necessary adjustments and settings indicated by the X-ray photographs can all be made in advance of the actual operation. The operation is thereby considerably shortened, while in addition, except for the carrier of the guiding needle, the instrument need not be sterilized.

Conservative and operative treatment of fractures

In general the surgical treatment of bone fractures may be confined to reduction of the fracture, making every attempt to fit the fragments into their proper places. If this reduction is successful and the position remains correct, which should be checked by X-ray examination, the body itself provides for the further repair by the formation of new connective tissue which later calcifies (callus formation) and makes a firm joint (consolidation).

There are, however, cases where this so-called closed or conservative treatment of a fracture is inadequate, either because there is insufficient callus formation or because this takes place so slowly that the patient has to be confined to bed for months, with the danger, so familiar to doctors, of secondary pulmonary or other disorders. A classic example of this is the fracture of the neck of the femur (fractura colli femoris). This fracture occurs most often with eldery people due to a fall on the side, and since callus formation in the femoral neck generally leaves much to be desired and the fracture in question naturally makes the use of the legs impossible, with the closed treatment healing, if occurring at all, requires a very protracted immobilization of the patient.

For this type of fracture, therefore, it has long been the custom to apply a different method of treatment. After reduction the broken bone is exposed operatively and the fragments are joined firmly together by driving in a triradiate nail. The X-ray photograph of *fig. 1* shows an example of

Fig. 1. X-ray photograph of a fractured neck of a femur with nail driven in.

such a "femur neck nailing". When the fracture has been "nailed", the patient can usually be up and about after a few weeks. In cases making favourable progress callus is formed in spite of the leg being used, and consolidation takes place around the pin. The nail, for which nowadays a special stainless steel is mostly used — formerly a piece of fibula was taken from the patient's calf — can subsequently be removed by a simple operation.

Directing the nail in "femur neck-nailing"

The pin has to be driven from the outside of the femur through the neck and the centre of the femoral head (fig. 1). The precise position that the nail should ultimately occupy can be ascertained by the surgeon from the X-ray photographs of the broken bone after reduction .The problem in this method lies in the directing and driving in of the pin so that exactly the desired position will be obtained.

The approximately 8 mm thick nail cannot be driven in without a guide. The pin is therefore made hollow and driven over a thin guiding needle made of knitting-needle steel, which is previously bored into the bone and drawn out after the nail has been driven in. The problem of directing the nail is of course not thereby solved, but it is somewhat facilitated in that it is transferred from the nail to the much thinner guiding needle. Any error in the direction chosen can be corrected by inserting other needles until the photographs show that the needle is in the right position. During recent times in fact the procedure has usually been such that the needle is directed visually with the sole assistance of several reference points and requirements of an anatomical nature, the needle seldom being placed in the correct position at the first attempt. The result was a long operation (with long narcotization) and the accompanying risk of complications.

Attempts have repeatedly been made to improve upon this by constructing directing instruments to make it possible to insert the guiding needle in the correct position at the first attempt. Usually these directing instruments had to be aligned with respect to the patient's body by fixing them to a separate supporting pin driven at some point into the femur, while of course the greater part of the instrument had to be sterilized and all the adjustments had to be made during the actual operation (with the bone exposed). It is not surprising that most surgeons have abandoned such aids, for their attention was too much distracted from the operation itself, which leaves no time and attention to be devoted to complicated adjustments; the surgeon can barely spare the time taken in waiting for and studying the X-ray photographs.

The new directing instrument

The drawbacks mentioned are avoided in a new directing instrument for "hip-nailing" which has been constructed by Philips[1]). In order to reduce appreciably the time taken for the operation this instrument has been designed so that all the manipulations of fixing and adjusting can take place prior to the operation proper. This instrument is not fixed to the bone itself. Instead, bone and instrument are both fixed with respect to a third point, the operating table, *viz.* a stand fastened to it. Furthermore the directing of the needle is reduced to three adjustments, which can be derived in a simple manner from the X-ray photographs made in advance and necessary in any case for checking the success of the reposition.

The principle of the instrument is clearly illustrated in *figs. 2a* and *b*. The needle to be bored into the bone is placed in a needle conductor, which can be moved around a horizontal axis (by sliding along the arc *C*) and around a vertical axis (by turning the arm *B* about a hinge *S*), in such a way that the needle is always directed towards the

[1]) The practical application of the method has been tried under the direction of the first named author in the Hospital "Calvariënberg" near Maastricht (Netherlands) and has already jielded good results.

same point (*O* in fig. 2*a*). This point lies at a fixed distance vertically under the hinge. Since, however, the hinge is fastened to an arm on a stand (*A* in fig. 2*b*), which in turn can be rotated about the vertical stand, moved up or down the latter and also varied in length, the point *O* can be determined at any arbitrary point with respect to the operating table. The surgeon has to make the point *O* coincide with the centre of the femoral head while the patient is fixed on the operating table. Then the guiding needle always points to that centre.

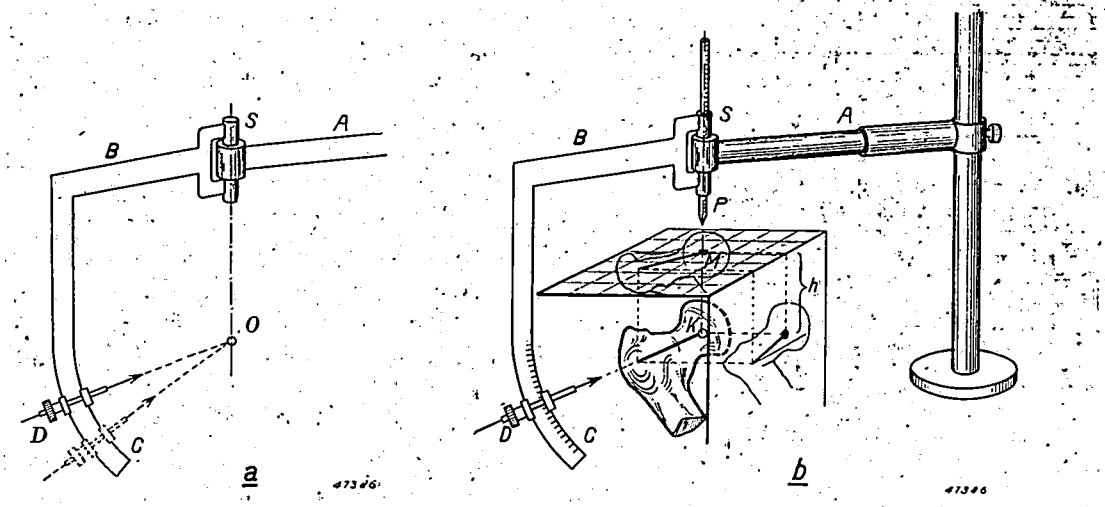


Fig. 2. Principle of the new directing instrument.
a) *D* needle conductor with the guiding needle, *C* arc with graduated scale, *B* arm rotatable about the hinge *S*. Upon the rotation of *B* and the sliding of *D* along *C*, the guiding needle always remains directed towards the point *O* lying at a given distance vertically below *S*.
b) Bij means of radiographs the fractured femoral neck is projected upon a horizontal and a vertical plane. *K* centre of head of the femur, *M* projection of this on the horizontal plane, *h* depth of *K* under *M*. *P* vertical hinge-pin graduated and moveable in the hinge *S*. *A* telescopic arm of stand. When *P* is set at a scale division corresponding to *h* and the extremity of *P* brought to the point *M*, the fixed point *O* to which the guding needle is pointed lies at the centre of the head *K*. In both projections the desired position of the nail is drawn in, and the needle conductor can then be set at the angles found.

After that it only remains to adjust the angle in the horizontal plane and that in the vertical plane.

In order to carry this out a horizontal and a vertical projection of the femur is made by means of an antero-posterior (with respect to the patient) and a lateral X-ray photograph. For the sake of simplicity it is assumed that the two planes of projection are the two planes indicated in fig. 2*b*, which are in a fixed relation to the operating table and which contain the axes of a three-dimensional rectangular system of coordinates. The projection of the centre of the femoral head on the horizontal plane is indicated by *M*. The projection of the centre on the vertical plane lies a distance *h* below the horizontal plane of projection. The arm *A* is now so adjusted that the extremity of the pin *P* of the hinge, which slides vertically in the hinge *S* and is provided with a scale, stands at the point *M*. The hinge is at the zero point of the scale when the end of the hinge-pin coincides with point *O*, that is the fixed point towards which the guiding needle points. When the hinge-pin has previously been moved up so far that the hinge stands at the scale mark *h*, the fixed point *O* for the guiding needle then lies a distance *h* vertically below point *M*. The first adjustment is thus completed. On the two X-ray projections of the femur, on which the desired position of the nail is sketched in, the angles can be found at which the needle conductor has to lie. The latter can then be adjusted to these two angles by means of a graduated scale on the arc *C* and by setting the rotating arm *B* parallel to the horizontal projection of the desired direction [2]).

Practical application

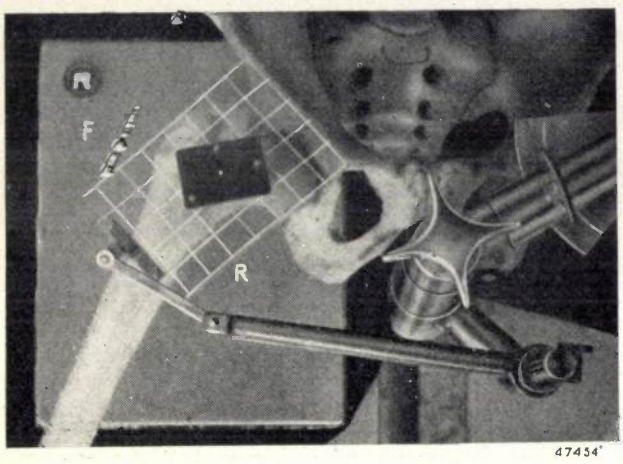
The horizontal plane with system of coordinates on which the femur must first be projected is realized by a horizontal wire grid, which can be swung around

[2]) The angle measured in the lateral X-ray photograph is exactly equal to the required angle on the arc *C* only when the vertical plane of projection is parallel to the plane of *C* (*i.e.* parallel to the guiding needle) and the centre of projection (*i.e.* the focus of the X-ray tube) lies at the same height as the centre of the head of the femur. In the case of small deviations from these conditions, however, the difference is very slight, so that it is sufficient to adjust the X-ray film for the lateral exposure parallel to the femur visually and the X-ray focus approximately at the desired height.

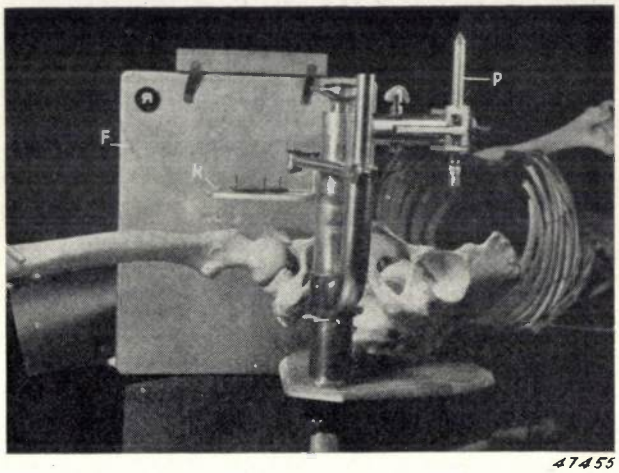
the stand column of the instrument and which is photographed together with the neck and head of the femur in a position just above these. *Figs. 3a* and *b* show the positions for taking the two radiographs, the patient being replaced by a skeleton for the sake of clarity. On the antero-posterior radiograph the point of projection M of the centre of the femoral head is drawn in in the grid; the depth of this centre below the grid is determined from the lateral radiograph.

In the case of the lateral radiograph it is necessary to make sure, with the help of a spirit level on the film cassette, that the upper edge of the film is exactly horizontal (fig. 3*b*). The angle of the nail measured with respect to this edge gives the adjustment of the needle conductor on the arc.

Having thus obtained all the necessary data, the directing instrument can be adjusted. In *figs. 5* and *6* two stages of this process are shown. The hinge-pin is moved along to the correct scale mark



a



b

Fig. 3. Situation for taking the antero-posterior (*a*) and the lateral (*b*) X-ray photographs. The patient is represented by a skeleton. The film cassette F with the X-ray film and the wire grid R serving as horizontal plane of projection may be seen. On this grid is a plate with various lead rods as markers for the corrections to be made. In the foreground in photograph (*b*) is the column of the stand, which is fixed to the operating table and with respect to which the patient is secured immovably. On the arm of the stand to the right is the hinge with hinge-pin P. On the film cassette is a spirit level for setting the upper edge of the film exactly horizontal.

Each of these two figures needs a correction. Let us first consider the lateral aspect. Due to the unavoidable enlargement obtained in making an X-ray shadow image because of the divergence of the rays, the depth (h') measured on the radiograph is too large (see *fig. 4a*). For the correction three lead rods 2 cm long are set up vertically on the grid (clearly visible in fig. 3*b*), which rods are also shown enlarged on the photograph, but enlarged to different degrees according to their distance. On the antero-posterior radiograph it is ascertained which rod lies closest to point M, and the enlargement of that rod then furnishes the correction factor for deriving from h' the true depth h.

The correction of the other figure, the projection of the centre of the femoral head on the wire grid in the antero-posterior radiograph, is less important. This projection (M_1) is only identical with the desired point M vertically above the centre of the head when the X-ray focus was also directly above the centre during the exposure; see fig. 4*b*. The usually very small and often negligible correction is found by placing a mark on the grid directly under the focus, for instance a small hole in a horizontal lead rod. If this hole is found to be projected at a distance d from the centre of the head on the radiograph, it follows from the figure that point M on the grid must be located a distance $x = d \cdot h/f$ in the opposite direction from M_1 (f is the measured distance from focus to film).

and set on the correct point (M) of the wire grid. By means of a pointer inserted in the hinge-pin underneath, the arm B is turned to the correct

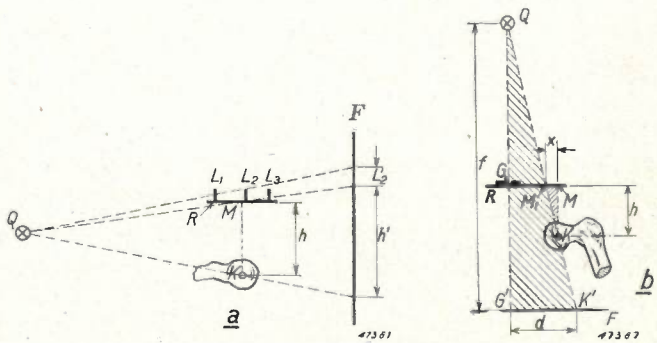


Fig. 4. Explanation of the required corrections.
a) In the lateral radiograph. Q focus of the X-ray tube, K centre of the head of the femur, F film, R wire grid. L_1, L_2, L_3 lead rods. The enlargement of the rod L_2 gives the correction factor for determining the true depth h from h'.
b) In the antero posterior photograph. Q, K, F and R as in *a*). G small hole in horizontal lead rod as marker for point directly below Q. The distance x from the desired point M to the erroneous projection M_1 follows from $x : h = d : f$, due to the similarity of the cross-hatched triangles.

position according to the horizontal projection of the position required for the nail [3]). In fig. 5 it will be seen that the arc *C* and the needle conductor are not shown. Since these parts will be close to or

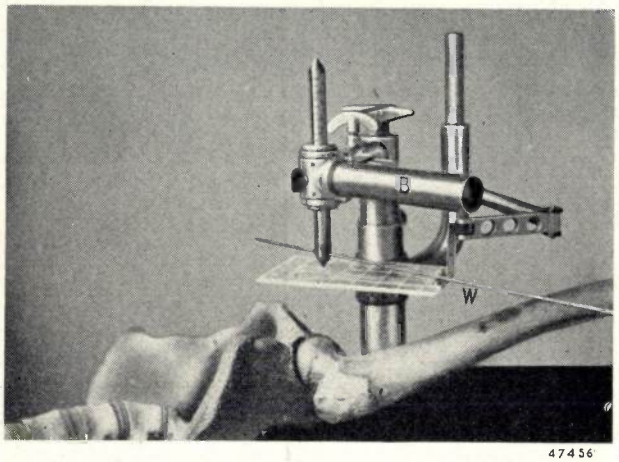


Fig. 5. After the end of the hinge-pin has been brought to the correct point on the wire grid, by means of a pointer *W* inserted in the hinge-pin, the latter, which is made to slide in a slot and therefore always carries the arm *B* with it, is turned to the correct direction, *i.e.* according to the desired direction of the nail drawn on the wire grid in the horizontal projection. The arc with the needle conductor, which have to be sterilized, are not shown.

even in contact with the area of the operation, they have to be previously sterilized together with the other instruments necessary for the operation. Only after all the adjustments are completed and the area of the operation is disinfected and covered sterilely, are they passed over the arm *B* and fixed as shown in fig. 6. The needle is then fixed at the correct scale division of the arc.

After this the actual operation begins. By sliding the needle conductor as far as the surface of the skin of the thigh the surgeon determines the correct position for the incision to be made, so that the wound can be kept as small as possible. The arc with the needle conductor can then be turned upwards 180° in order not to be in the surgeon's way. After the bone has been exposed and the hard superficial layer has been chiselled or bored away, the arc is definitely fixed in the directing position, the needle conductor, which is provided with two sharp points, is tapped a few mm into the bone and the guiding needle proper is bored in to the required depth — which has also been deduced in the obvious manner from the X-ray photographs. The needle conductor and the whole directing instrument are then removed, the hollow nail is passed over the guiding needle and hammered into the bone with a hollow punch.

Before this last step is taken two new X-ray photographs could, of course, be made to make sure that the guiding needle has taken the desired direction in the bone. Experience has shown, however, that there is really no need for this, unless meanwhile the position of the patient should have unexpectedly changed, which is not at all likely. After the nail has been driven in, however,

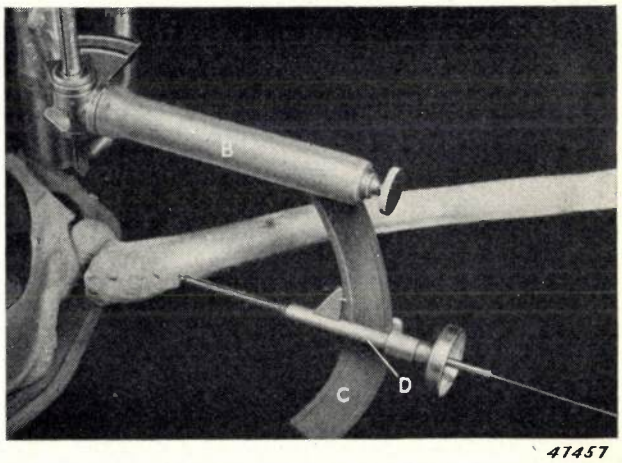


Fig. 6. Arc *C* and needle conductor *D* have been put in place, the needle conductor has been tapped into the bone and the guiding needle can now be bored in.

another couple of radiographs is always taken to check whether the nail has been driven in to exactly the desired depth ; if the photograph shows it to be necessary the surgeon can drive the nail slightly farther in.

[3]) This direction, too, only agrees exactly when the focus happens to be exactly vertically above the centre of the head of the femur. The correction required for the deviation from this condition, however, is so small as to be practically negligible for our purpose.

ILLUMINATION INTENSITY IN OFFICES AND HOMES

by A. A. KRUITHOF and H. ZIJL.

535.241.46:628.972

An attempt is made in this article to solve the problem of the most suitable illumination intensity for the activities of office and home. Use is made not only of data about visual acuity and contrast sensitivity of the eye, but also of tests of the fatigue resulting from long-continued reading. The conclusion is that upon increasing the illumination intensity from 10 to 100 lux a very considerable improvement can be obtained, from 100 to 1000 lux still an appreciable improvement and above that only a slight improvement. Consideration is given to a method for finding the optimum illumination level for practical office lighting, while it is pointed out that experience has shown 100 lux to be the necessary minimum illumination intensity. In conclusion the reasons are discussed for the fact that in homes almost invariably much lower illumination intensities are used than are advisable for ordinary household tasks; possibilities of improvement are suggested.

The human eye possesses a remarkable power of adaptation: it can discern objects in the brightest sunlight under an illumination intensity of maybe 100,000 lux and on moonless starry nights when the illumination intensity is only 0·0003 lux. Vision under such extreme conditions, however, is of course difficult and defective. Man, unwilling blindly to accept unfavourable natural conditions, looks for means of correction: in bright sunlight he puts on dark glasses, at night he uses a lantern, in order in both cases to adjust the amount of light available for the eye to more suitable values.

The region of the "more suitable" values of illumination intensity, where the eye usually works without extra assistance, is still extremely wide: it extends from a few lux, the level employed in street lighting — motor car drivers ordinarily switch on their head-lights as soon as the illumination intensity on the road falls below 3 to 4 lux — to the more than 10,000 lux necessary in a film studio. These "more suitable" values are certainly not the "most suitable" ones: if only the requirement of light for the eye is considered, higher intensities would certainly be employed in the first case and lower intensities in the second. In the first case considerations of economy chiefly determine the upper limit, while in the second case it is the light required by photographic material which determines the lower limit.

In all cases where illumination is necessary such limitations will be manifest. But let us suppose that for a while these limitations were absent. For each type of activity that has to be carried out there would be a more or less accurately determined value of the illumination intensity at which that activity can "best" be carried out. This value would probably differ somewhat for different persons.

There can be no doubt that it is important to find out what are the most suitable values of the illumination intensity.

On the basis of a number of investigations, part of which have been published during recent years, we shall in this article attempt to answer this question. We shall confine ourselves to the activities ordinarily performed in offices and homes. At the same time the economic, technical and other (chiefly esthetic) considerations will be discussed which are the cause that the illumination levels actually used in practice remain very far below the "most suitable" ones.

However, before we go farther into the effect of the illumination level, it must be emphasized that illumination intensity alone is of course not sufficient to characterize an illumination. The quality of the illumination is also of very great importance, and this factor is determined by such features as the degree of diffuseness of the light, the distribution of the illumination over ceiling, walls and working surface and the spectral distribution of the light. We shall discuss the influence of these factors only incidentally and very briefly. For our discussion we shall simply start from the assumption that all the factors mentioned have been chosen as favourable as possible; in particular we shall limit the discussion of practical illumination to those cases in which the fixtures used are constructed in such a way that the disturbing and harmful phenomenon of glare cannot occur.

Factors characterizing the required performance of the eye.

In every human activity the task of the eye comes down to distinguishing some detail against a given background under the given illumination intensity. In reading printed matter, for example, the details are formed by the printed letters, in drawing by the pencilled line, in sewing by the peculiarities of the weave of the material, while in the first two cases the background is a sheet of paper and in the third it is the material as a whole.

The requirements made upon the eye in the different activities can be characterized by the following quantities:

1) the size of the detail to be observed, or more precisely: the angle of vision subtended by the detail under observation;
2) the contrast between the detail and the background given by the ratio of the coefficients of reflection [1]);
3) the time available for the observation (for instance, in sewing the time necessary to find the spot to insert the needle);
4) the length of time during which the activity is continuously performed.

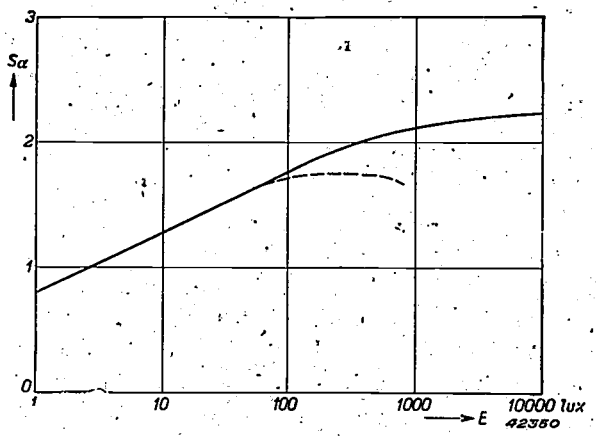


Fig. 1. Visual acuity S_a of an average observer as a function of the illumination intensity E of the background. The background is "white" (reflection coefficient $\varrho = 0{\cdot}75$), the detail observed "black" ($\varrho = 0{\cdot}04$). In the case of the continuous-line curve the surroundings are just as bright as the background; in that of the broken-line curve the surroundings are quite dark. The curves are obtained by a combination of data of Eguchi, Lythgoe, Schober, König. For lower reflection coefficients of the background the experimental results are also valid when account is taken of the fact that it is not the illumination intensity but the brightness which determines the threshold sensitivity of the eye.

Each of the first three quantities, angle of vision, contrast and time of observation, must exceed a certain (minimum) value (depending on the other circumstances) in order to make an observation possible. If the angle of vision must be at least α minutes, the contrast at least $c = B_1 : B_2$ (B_1 = the brightness of the detail, B_2 = that of the background), the time of observation τ seconds, then $S_a = 1/\alpha$ is called the visual acuity, $S_c = {}^1/_2\,(1+c)/(1-c)$ the contrast sensitivity and $S_w = 1/\tau$ the speed of observation of the eye.

[1]) The surfaces which occur in offices and living-rooms often exhibit approximately diffuse reflection. When this is not the case it is impossible to speak of one reflection coefficient, and in order to indicate the contrast direct use must be made of the ratio of brightness of detail and of background. The same is also true, of course, when detail and background do not receive the same illumination intensity.

The threshold values mentioned depend very much upon the intensity of illumination. A consideration of this relation will furnish us with a first indication about the desired illumination levels. Meanwhile the effect will become evident of the other characteristics of the illumination, such as the colour and distribution of the light and especially of the brightness of the surroundings, for instance of the walls of the room.

In *fig. 1* the visual acuity S_a of an average observer, determined by measurement under the circumstances given in the text for the figure, is shown as a function of illumination intensity. It may be seen that up to about 1000 lux visual acuity increases rapidly, while above 1000 lux it increases only slowly. The broken-line curve shows how the performance of the eye decreases when the brightness of the surroundings is too low.

In *fig. 2* analogous curves are drawn for the contrast sensitivity S_c. Let us first consider curve *1*, which is the most important for us. Here also a rapid increase in the performance of the eye is observed when the illumination intensity is increased to the region of 1000 lux, while at still higher illumination intensities the increase becomes very slight. Curves *2* and *3* prove again the harmful influence of too dark as well as too bright surroundings; in the latter case this is a consequence of glare. Curve *4* is recorded with a larger test

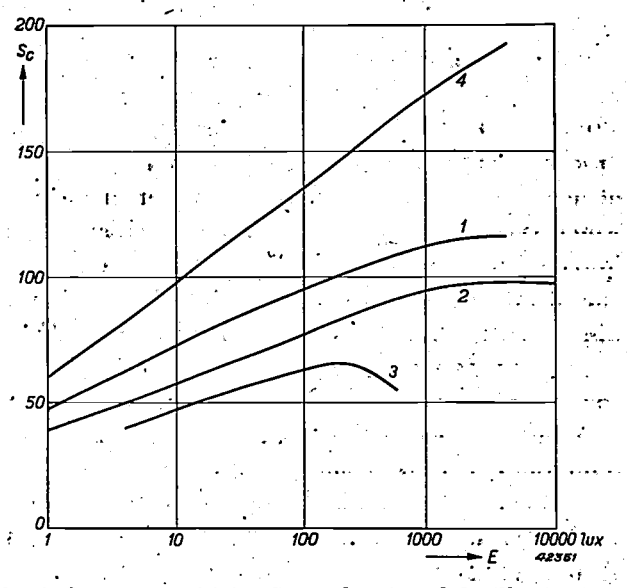


Fig. 2. Contrast sensitivity S_c as a function of the illumination intensity of the background. In the case of curves *1*, *2* and *3* the test object has a diameter of 1°, while the surroundings are, respectively, about equally as bright as the background, completely dark, and three times as bright as the background. In the case of curve *4*, the surroundings are just as bright as the background, but the test object has a size of about 4°. Data have been taken from Stiles and Crawford, Hoppe, and Siedentopf, Schumacher.

object than the others; the contrast sensitivity in this case is higher over the whole range.

Similar behaviour as for visual acuity and contrast sensitivity is found for the speed of observation S_w, but here the increase with illumination intensity is only worthwhile up to 100 lux, above which it is only slight.

From all this information conclusions can now be drawn as to the improvement that can be obtained by an increase of the illumination intensity. Two objections must, however, be taken into account. The first is that this information naturally relates only to the threshold of observation. At a few lux already the angles of vision and contrasts occurring in the practical work to be performed lie in general far above this threshold [2]). The increase in the performance of the eye with increasing illumination intensity, as far as the threshold sensitivities are concerned, is thus of itself of no importance to us.

The other objection is that the visual acuity, the speed of observation and the contrast sensitivity are quantities of such a strongly conventionalized character that they cannot express the practical performance of the human eye.

An attempt to meet these two difficultics may be found in a recent investigation by Weston [3]). He used test objects presenting details of the size of 1 to 6 minutes and defined the performance of the eye with a given size of detail and contrast as the quotient of the relative number of correctly judged test objects and the time necessary for the observation of the detail. His conclusion is that for normal work (size of detail about 3′, corresponding to the reading of printed matter in Bodoni 10 points, see footnote [2])) the performance distinctly increases up to about 40 lux. At this intensity it already amounts to about 90% of that at 5000 lux. For fine work (size of detail 1′) there is a marked increase in the performance up to about 1000 lux, where again about 90% of that at 5000 lux is attained. These results constitute a further specification and a good confirmation of the conclusions to be drawn directly from figs. 1 and 2.

[2]) In reading printed matter, for example, the size of a normal letter type (such as Bodoni 10 points, which is used in the text of this article) is about 1·5 mm; the detail of each letter which makes it recognizable is usually about 5 times as small. At a distance of 30 cm from the eye the angle of vision is then 3·5. According to fig. 1 at 1 lux details down to about 1·2 can still be observed. Furthermore the reflection coefficient of the background lies between 0·8 and 0·6, that of the type between 0·04 and 0·10; the contrast thus amounts to from 1 : 6 to 1 : 20, while at 1 lux and 3·5 angle of vision contrasts of 1 : 2 can still be observed.

[3]) H. C. Weston, Industrial Health Research Report No. 87, H. M. Stationery Office, London 1945.

We now arrive at the last point in the list of characteristics of an activity, namely the duration. The effect of the duration manifests itself in the fatigue phenomena. It is remarkable that this fatigue is scarcely manifested in the performance of the eye as far as visual acuity and contrast sensitivity are concerned. The performance of an activity considered as a whole, however, is felt to be "more difficult". Parallel with certain physiological phenomena, such as more frequent blinking and a slowing up of the heartbeat [4]), a general decline in the performance of work is observed, a decrease in the rate of reading, in the accuracy of drawing, etc.

Fatigue tests.

The increase in the performance of the eye with increasing illumination intensity as discussed in the foregoing may be related to the more or less trivial fact that with increasing illumination intensity observation becomes "easier". From the shape of the curves of figs. 1 and 2 and especially from Weston's results it is possible, to some extent, to read off the degree of this increase in ease. The fact that an increase in illumination intensity makes observation "easier", while we have just ascertained that upon long duration of the activity observation of work becomes "more difficult", makes it reasonable to expect that illumination intensity will have an effect on the occurrence of fatigue. It is, however, obvious that it will not be so simple to determine this effect quantitatively as the effect on the observation thresholds, on account of the difficulty of expressing the effect of becoming fatigued in numbers.

Direct tests of fatigue have been made by Luckiesh and Moss [5]), among others. For a large number of test persons they investigated the fatigue occurring upon reading under accurately determined optical conditions which may be considered as an idealization of the conditions occurring in practice. They used various criteria of fatigue. The most obvious, *i.e.* the criterion which is most closely connected with the act of reading itself, is the decline in reading speed. Remarkably enough, however, this decline is always found to be only slight and consequently to exhibit only small differences for very divergent illumination intensities. According to Luckiesh and Moss [6]) this may be

[4]) M. Luckiesh and F. K. Moss, Trans. III Eng. Soc. 34, 571, 1939.

[5]) M. Luckiesh and F. K. Moss, Trans. III Eng. Soc. 35, 19, 1940; 35, 703, 1940.

[6]) M. Luckiesh and F. K. Moss, Reading as a Visual Task, D. van Nostrand Co., New York 1942.

understood to mean that under worse visual conditions the eyes are stimulated by the brain to correspondingly greater effort. The relation between "fatigue" and speed of reading is thus lost for the greater part, so that speed of reading is not a very suitable criterion of fatigue.

Another criterion used by Luckiesh and Moss in their tests is the frequency of the involuntary blinking during reading. This criterion is less directly connected with the act of reading than the speed of reading; but there appears to be a satisfactory correlation between greater difficulty in reading and more rapid blinking. Measurement of the number of blinks per unit of time at the beginning of reading and after 1 hour of reading gave the following results:

at 10 lux an increase after one hour of 72%,
at 100 lux an increase after one hour of 31%,
at 1000 lux an increase after one hour of 8%.

It is clear that under the higher illumination intensities less fatigue occurs; between 10 and 1000 lux a steady improvement can be observed with increasing intensity, although the improvement per lux increase in intensity rapidly becomes less at higher levels.

These results are quite parallel with our conclusions from figs. 1 and 2 about the increasing "ease" of observation, and we may therefore accept it as established that an increase in illumination intensity from about 10 lux to 100 lux is accompanied by a considerable improvement as far as fatigue is concerned, while an increase above 100 to 1000 lux is still accompanied by an appreciable improvement. Above 1000 lux no direct fatigue tests have been carried out, but on the basis of the parallelism already mentioned we may expect here also a slight improvement similar to the curves of figs. 1 and 2.

In the tests of Luckiesh and Moss use was made of very clearly legible printed matter. It may be expected that for less easily legible printed matter (smaller letters, little contrast with the background, as in carbon copies of typed material) more pronounced fatigue phenomena will be found. We may assume that in these cases the decrease in fatigue upon increase in illumination intensity will be at least just as great as was found in the tests with good printed matter. This conclusion also holds, of course, with respect to other activities which are more difficult than reading, for instance drawing, mending of clothing, etc.

In conclusion it must further be mentioned that in the tests of Luckiesh and Moss the influence of the brightness of the surroundings (at an illumination intensity of 135 lux) and of the spectral composition of the light was also studied. Equal brightness of surroundings and background were found to be most favourable. Incandescent lamps and luminescent lamps of daylight colour gave no appreciable difference in case of reading.

Illumination in practice.

Let us now consider the practical application of the information obtained. The high levels of illumination found to be so favourable for the eye will seldom be found in offices and never in a living room. In the introduction we have already suggested the reason for this: it is the influence of a combination of other factors besides the behaviour of the eye. If we are to describe these influences in more detail we must consider offices and homes separately.

Offices.

For office illumination the utility factor is most important. Both for the initial installation and for possible alterations, it is primarily the economic advantages and disadvantages that are weighed.

What are the advantages? According to the foregoing an increase in illumination intensity will result in an increase in the production of the employees, which represents a certain money value. This increase will be considerable at low levels of illumination but will become less and less at higher levels, until finally there is no advantage at all. We thus obtain a variation of the advantages as indicated qualitatively by the continuous-line curve in *fig. 3* [7]). The region of illumination intensities where the curve is practically horizontal will lie higher according as the work performed in the office is more difficult.

Costs show an opposite trend. The additional expense entailed, for example, by doubling the illumination intensity is less at a low level of illumination than at a high level. The curvature of curve *1* in fig. 3, which shows qualitatively the variation of costs, is thus just the opposite of that of the continuous line curve representing the ensuing advantages. If the two curves are compared in two points pertaining to a given value of the illumination a certain value of the illumination intensity is found where the tangents are parallel. This is the point where the best economy is attained, because at a still higher illumination level the increase of expense is greater than that of advantages.

Due to advances in technique in the course of

[7]) It is obvious that such a curve can in practice only be based on very rough estimations and often cannot even approximately be drawn. The general shape should, however, be like that indicated.

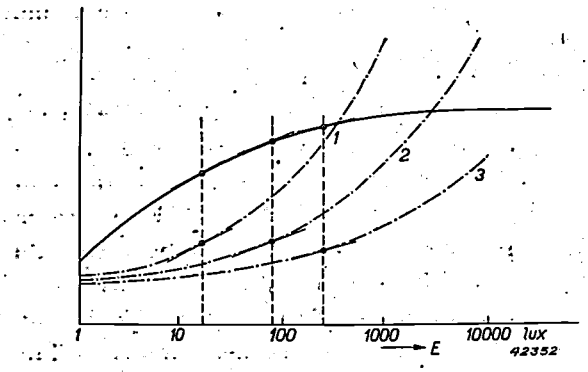


Fig. 3. Schematic variation of the advantages of an office illumination (full-line curve) and of the costs (broken-line curves) as a function of the illumination intensity. The three curves *1*, *2*, *3* correspond to increasingly higher stages of technical development. The most favourable illumination intensity lies at the value for E where the curves for advantages and disadvantages are parallel. With advancing technique this point shifts towards higher illumination levels.

time the curve of costs will fall with respect to the curve of advantages. The points showing parallel tangents will therefore be shifted toward the right, *i.e.* the most economical illumination level rises. Some indication of the height of that level at the present stage of technical development may be found in the values which are recommended by the lighting societies in various countries, see the *table* below [8]). The above-mentioned advance in technique is clearly manifested in the U.S.A. values for different periods.

Type of work	U.S.A.			Gt. Britain	Switzerland
	1930	1939	1942	1937	before 1940
Reading and writing with interruptions	80-100	200	250	100-150	150
Long-continued calculation or study	100-150	300-500	500	150-250	300

Since some time always elapses before practice adapts itself to technical progress, in most offices one encounters illumination levels which are lower than what are to be considered the most economical. As long as the level is not too low the difference is only a question of utility, as discussed above. It is, however, clear that at too low intensities phenomena must occur which can no longer be expressed purely in money value but which already belong to the sphere of hygiene. There will then be complaints by the employees of fatigue, headache, etc. General experience in the practice of factory and office lighting is that the limit above which these phenomena are no longer frequent for activities such as reading lies at about 100 lux.

This value is also found on the basis of the following considerations. The avoidance of complaints means that each individual worker must be satisfied with the illumination. The preceding discussion, however, referred to the average behaviour of a large number of persons. The effectiveness of vision for such a collection of persons, even though they are all "normal", will show a certain spread: The curve showing the spread of the contrast sensitivity and the visual acuity for a large number of persons has been determined (*cf.* for example the article referred to in footnote [4])). In both cases the curves deviate from the normal distribution curve of Gauss, but in the following we shall approximate them by normal distribution curves since that does not alter the order of magnitude of the results. In *fig. 4* curve *a* represents

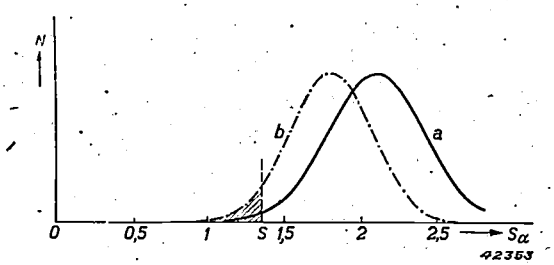


Fig. 4. Schematic representation of the spread of the visual acuity S_a of a large number of test persons. A Gauss distribution is assumed; the actual distribution deviates somewhat from that, but the deviation may be ignored for our purposes. Curve *a* refers to an illumination intensity of 1000 lux. At a lower illumination intensity the whole curve is displaced towards smaller values of S_a; thus for example at 100 lux curve *b* is found.

schematically the distribution curve, *i.e.* the number of persons N as a function of visual acuity, for an illumination intensity of 1000 lux. With each value of visual acuity, no matter what combination of conditions (which may be characteristics of the person or of the light) is responsible for it, we may correlate a definite value of the "ease of working". At what value of visual acuity is the ease of working unsatisfactory? From the fact that at 1000 lux practically none of all the test persons experiences any difficulty, we may conclude that the critical value of visual acuity will lie to the left of the point S in fig. 4. The part of the distribution curve *a* lying to the left of S contains fewer than 1% of all test persons. If we construct the corresponding distribution curve for a lower intensity than 1000 lux we obtain again, schematically, curve *b*, which is shifted to the left with respect to *a*. Now a larger percentage of all test persons falls below the limit S. The shift for an illumination intensity

[8]) It must not be forgotten that in establishing these illumination levels other considerations besides utility, such as, for example, that of the ease of the observation itself, have played a part. As to the values recommended in America, in the report in Trans. III Eng. Soc. 34, 371, 1939, the preliminary studies for the recommendations of 1939 are included. These preliminary studies contain a discussion of economic advantages and disadvantages; the conclusion is that the increase in the advantages at the levels given more than balances the increase in expense. The most recent American recommendations are to be found in Trans. III Eng. Soc. 37, 275 and 449, 1942; 40, 339, 1942. A new proposal for Gt. Britain is in preparation; Trans. Ill. Eng. Soc. (London), 8, 17, 1943.

of 100 lux, however, according to the experiments is found to be still so small (about 40% of the distance between S and the maximum of curve a) that again the total percentage of persons whose vision lies below S is still small, namely 5%. Analogous arguments hold for the contrast sensitivity, the same qualitative conclusion being reached: at 100 lux 3% of the persons fall below the limit S. The probability that a person possessing "normal" vision will encounter difficulty from too little light is therefore also small at 100 lux. With less than 100 lux, however, this probability increases rapidly.

Similar considerations can be found in the article by Weston[3]) already referred to. At high illumination intensities the performance values found for different persons show only a small spread. At lower illumination intensities the relative spread increases, this being ascribed to the fact that the conditions of seeing become too unfavourable for some persons. The limit depends upon the kind of work, but always lies at about the illumination intensity at which the "performance" reaches 90% of the maximum value, thus, according to the values given above, at 40 lux for ordinary work and at 100 lux for fine work.

The living-room

In the case of the illumination of the living-room the situation is quite different from that of the illumination of offices, because the living-room is used not only for certain activities such as reading, mending, etc. but is alternatively also the cosy gathering place of the family, the festively lighted reception room or the quiet surroundings where rest is enjoyed. In order to obtain suitable light distribution in all these cases, different illuminations will have to be installed which can be used separately or in combination.

We shall confine ourselves here to a consideration of the illumination which will be employed when the living-room is used as a work room. From the discussion of office lighting it followed that for activities such as long-continued reading, intensities of at least 100 lux must be employed. This result was not reached on the basis of considerations of utility — which are of less importance in the case of the home — but on the basis of considerations of hygiene. The minimum limit can therefore also be considered valid for the living-room. Having regard to the poorer vision of older persons and the greater difficulty of some types of household tasks, embroidery for instance, as compared with reading, it is better to choose a 50 or 100% higher illumination intensity.

What is now the situation as regards the realization of these illumination levels in the living-room? The solutions arrived at for the office would not in general be acceptable in the living-room, since, even though it is used as a work room, a domestic sphere is desired which cannot be combined with the technical austerity of utility illumination. Esthetic considerations here occupy the foreground: the lighting ornament must harmonize with the interior, and that not only when lighted but also in the daytime. The shape and colour which it exhibits in the daytime are indeed usually the criteria which decide the choice of a given ornament, while the illuminating technical characteristics are usually given too little consideration. The illumination intensity is often quite insufficient, even when using the largest lamp permissible in connection with the heat development, to say nothing of the distribution of the light flux and the precautions against glare.

It is no wonder that in almost all living-rooms illumination intensities are found which lie far below what is desirable in a work room. If the eye protests too strongly recourse is usually had to extra local illumination, which, however, when several persons are working at the same time cannot be considered a satisfactory solution.

Having pointed out this rather disappointing state of affairs, we need not resign ourselves to it. At the present day it is without doubt technically possible to install an esthetically satisfactory living-room illumination of sufficiently high level, and in this respect two points have been found to be of great importance.

1) The esthetic acceptability of an illumination depends very much on the relation between the brightnesses of the surfaces occurring in the room (table cover, ceiling, walls). With a very light table covering the surroundings should also be light, since it is found that with increasing brightness large contrasts make a more and more unpleasant impression. Dark walls already form of themselves a great contrast with the working surface; from economical considerations they also hinder the employment of indirect lighting, which is so useful for reducing contrasts, and thus quickly set a limit to the illumination intensity. In interiors with light walls and not too dark upholstering, illumination intensities of 150 to 200 lux are much more easily attainable than in darker surroundings. The correct technical construction of the fixture to distribute the light in the desired manner over the working surface and surroundings no longer presents any problems with the present-day technical development.

2) The level of the illumination considered as being "cosy" or "pleasant" is related to the colour of the light. For the above-mentioned illumination intensities of 150 to 200 lux no unplea-

sant effects [9]) are, it is true, experienced with the reddish light of incandescent lamps, but a whiter light, more nearly approaching daylight, may in some cases give the whole interior a more agreeable aspect. The possibility of such an improvement in colour is now offered by the development of luminescent lamps [10]). At the same time, with the smaller heat development of these lamps there are no difficulties with the fixtures in that respect.

It may be hoped that these new possibilities will ultimately remove the great difference between the desirable and the actual illumination intensity in the living-room when it is used as a work room.

[9]) See Philips Techn. Rev. **6**, 69, 1941.
[10]) Philips Tcchn. Rev. **6**, 65, 1941.

MAINTENANCE MEASUREMENTS ON CARRIER TELEPHONY EQUIPMENT *)

by J. de JONG.

621.395.44:621.317.3

The maintenance of telephony equipment involves many complex and very accurate measurements, the most important of which are measurements of level, for ascertaining the strength of a signal at different points in a line of communication. The equipment employed for this purpose, as well as for other measurements, is described in the following, which deals in particular with the influence of the special requirements appertaining to telephony, upon the construction of the equipment and the manner in which these tests are carried out. For example, accuracy and stability of the test gear are of the utmost importance: again, it must be possible to perform the tests while the system is in operation and, moreover, at many different points in the equipment, possibly even some hundreds of miles apart. In view of the comprehensive nature of the tests, the apparatus must be at once adaptable to any special contingency (impedance, levels, earthing etc), while all the necessary adjustments and readings must be quickly and readily accomplished.

Few branches of technology necessitate such far-reaching test-schedules, or place such high demands on accuracy of measurement as Telephony, and this applies not only to the development of new equipment and components, but equally to the many special tests carried out in the manufacture and installation of this type of equipment. We are more especially concerned, however, with what may be called maintenance measurements, by which is meant those measurements essential to the proper upkeep of the equipment and the communications which it maintains. The importance of such measurements will be manifest when it is recalled that every channel of communication involves the use of the most complex types of equipment, whilst, secondly, the risk of interruption in the service must be minimized as much as possible.

Maintenance measurements may be divided roughly into two categories, the location of faults and purely preventive inspection. The object of the first is to detect and localise sudden causes of failure, but by far the greater part of these maintenance tests is of a preventive character, since constant checks on the performance of the equipment will reveal latent causes of breakdown, for instance ageing of valves, variations in cable-losses due to fluctuations in temperature etc., all of which may be detected and remedied before an actual failure occurs. Such preventive measures are of particular importance in carrier communication, for which far more elaborate equipment is necessary than in the case of low-frequency telephony. Not only are the chances of faults therefore much greater, but any single fault may lead to an interruption of various channels of communication. In the case of the earlier system developed by Philips, employing 17 speech-channels, the failure of one of the repeaters results in the interruption of all 17 calls carried by any single pair of conductors in the cable. All the channels, including those of other pairs, are interrupted if the carrier supply common to these pairs becomes defective.

The introduction of carrier telephony has brought with it a great increase in the extent and nature of the tests to be carried out; for example in the 17-channel system alluded to the frequency range to be covered by the measurements has been increased from about 3000 to roughly 100 000 c/s, whilst, apart from cross-talk between the different pairs, it is now also necessary to measure the mutual effects among the 17 channels of any single pair.

Of those maintenance measurements which are of special importance in the upkeep of carrier-telephony communication we now propose to discuss the following:

1) Measurements of level, including the current strengths and levels of speech of signals at various points in the channels. (The level is defined as the ratio of the strength at any given point to that at the origin of the connection, converted to the same impedance.)
2) Cross-talk measurements between the various channels.
3) Checking of the working of the amplifier valves.
4) Location and identification of interfering frequencies and measurement of distortion.

The very extensive measurements of level will be dealt with in some detail; the other points will be examined more briefly. Further tests, such as

a) Editor's note. This article is the last of a series on carrier telephony commenced in 1941 and based on the 17-channel system. It is hoped that shortly a description will be forthcoming of a new carrier system developed in recent years.

those applied to signal and selector equipment, the inspection of telephone cables, subscribers, instruments and connections, will not be touched upon.

Measurement of level

In the construction of present-day telephone networks it is stipulated that attenuation of speech between one subscriber and another, within the boundaries of the country, shall not exceed 28.5 db, as this ensures reasonably intelligible conversation. Of this tolerance, twice 6.3 db is reserved to cover the relatively short connections between subscriber and exchange and this margin is sufficiently wide to dispense with regular tests which would otherwise be far too costly. On the other hand, the total losses in trunk connections that may be some hundreds of miles apart might easily exceed the remaining maximum of 15.9 db, due to temperature variations and so on. Again, such variations may even produce "negative losses" to such an extent as to cause instability (manifested by a howl in the telephone at each end of the line, so shrill as to drown ordinary speech and render conversation quite impossible).

In this form of communication, in which the working pressure is so great, the costs may be correspondingly higher and a continuous check on the losses must therefore be maintained.

This means that at the terminal station and at each of the repeater stations — of which there may be 10, 20 or even more on a long distance run — the levels of the incoming and outgoing signals have to be measured regularly and amplification readjusted to maintain the levels within the prescribed limits. These tests need to be carried out at several frequencies individually and the losses, which differ to a lesser or greater degree between one frequency and another, are subsequently equalised by means of small correction networks. The measurements necessitate the use of a test-oscillator capable of supplying an alternating voltage of variable frequency (within the whole range of speech frequencies to be transmitted), that is, in ordinary low frequency telephony, roughly from 300 to 3400 c/s. This oscillator is connected to the input end of the connection under test and an A.C. voltmeter is applied to the output end to measure the strength of the signals received.

Measurements in connection with carrier telephony are of a more complex type, the frequency range being much wider; in the case of the 17-channel system the appropriate oscillator and measuring instrument need to be suitable for the frequencies already mentioned, namely up to about 100 000 c/s. Another complication arises from the transformation which the speech frequencies undergo and in this connection reference is made to the diagram of a carrier channel depicted in *fig. 1*.

A speech frequency q is impressed by the modulator upon a carrier wave of frequency p. This furnishes two side-band frequencies $p \pm q$, as well as others, such as $2p$, $3p \pm q$ and so on, and that side-band whose frequency is $p + q$ is employed for transmission, the others being suppressed by the

[1]) See Philips Techn. Rev. 7, 83, 1942.

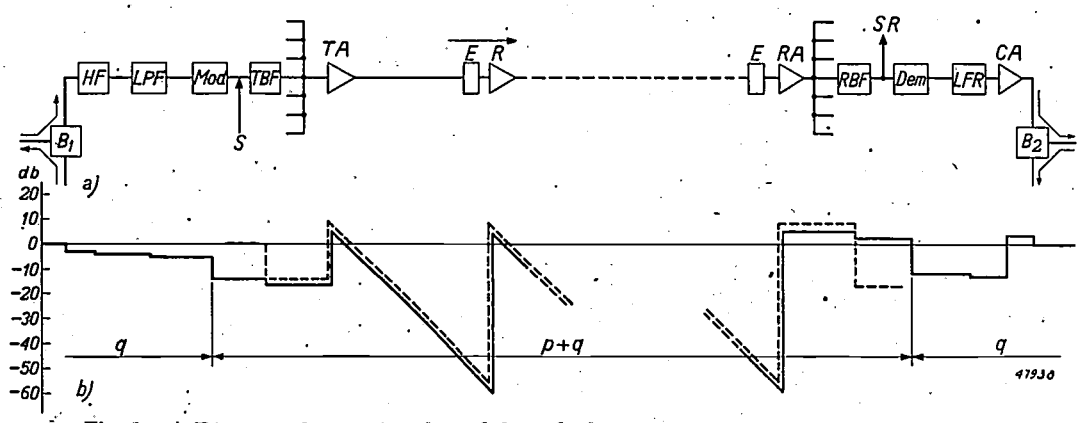


Fig. 1. *a*) Diagram of a carrier channel for telephony. B_1 = branch; HF = high-pass filter; LPF = low-pass filter; Mod = modulator; TBF = transmission band filter; TA = transmitter amplifier; E = equaliser; R = repeater; RA = receiver amplifier; RBF = receiving band filter; Dem = demodulator; LFR = low-pass receiving filter; CA = channel amplifier; B_2 = branch. Further, for carrier signalling: S = carrier wave injection; SR = connection for signal receiver.

b) The relative level diagram in respect of all the components under normal working conditions. The full lines show the level of speech at the various points in the system illustrated in the upper figure. It is assumed that the output and input levels are exactly equalised. At a speech frequency q and carrier frequency p, the signal under test will have the frequencies as shown. The broken line refers to the carrier used for signalling purposes.

transmission-band filter. Now, if it is desired to check the losses in the modulator, a frequency q is applied to the input and the voltage of $p + q$ is measured at the output. To ensure the exclusion of other frequencies at this point, a very selective measuring instrument, known as an analyser, is employed, this being also used for measurements of level following the transmitting band filter, for example, somewhere in the cable. It is true that it would be quite possible to apply only frequency $p + q$ by means of the oscillator, to the input of the transmitting band filter, but it must be remembered that at the other end of the band filter we are also concerned with the frequencies of all the other channels on the same pair of conductors.

In telephone work, in general, it is usually laid down that it must be possible for tests to be carried out on any one channel while the others are in normal operation, and the analyser is therefore essential for the elimination of the frequencies of the other channels.

When levels are measured in this way at the input and output sides of all the components of a given channel, a diagram is obtained of the kind depicted in fig. 1*b*. The level of the carrier transmitted for signalling [2]) is also shown, in broken lines.

The losses in the transmitting band filter with respect to the carrier frequency tend to vary very considerably with the temperature, seeing that this frequency is located just at the edge of the pass band; therefore, measurement of the level and subsequent correction are of particular importance.

Speaking generally, these remarks also apply to the highest frequency channels, since the absolute width of the frequency range is actually the same in all the channels (3 100 c/s, corresponding to speech frequencies of 300—3 400 c/s); the relative bandwidth in the highest channels is therefore much smaller than in the lower, so the effect of temperature variations in the coils and condensers upon the losses of the bandfilters, at the edges of the ranges passed by them, is greater in the higher than in the lower channels.

Description of the equipment required

The equipment by means of which the measurement of level is carried out will now be considered in greater detail, and our description is based on the types of instrument that would be suitable for the maintenance of the 17-channel system to which reference has already been made.

[2]) See Philips Techn. Rev. 8, 168, 1946.

The Oscillator

The oscillator is constructed along the same lines as the audio-frequency instrument described in a previous issue of this Review [3]), the required frequency in this case being the heterodyne of two alternating voltages of higher frequency, generated by two oscillator valves. In order to provide beat frequencies up to 100 000 c/s, the frequency of one of the oscillators is designed to give 400 000 c/s, whilst that of the other is variable between 400 000 and 300 000 c/s. The beat frequency is filtered out and amplified to the desired level, which, to meet the required accuracy of measurement, must be the same over the whole frequency range, to within 0.1 db. Since it is not an easy matter to construct an output transformer, such as is needed for matching purposes, that will give a sufficiently straight characteristic over the entire frequency range of 30 to 100 000 c/s, two output transformers are provided for switching in as required. One of these is employed for the frequencies from 30 to 10 000 c/s and the other from 4000 to 100 000 c/s. The transmitting level is adjusted to the required value by means of a calibrated attenuator.

Fig. 2 illustrates the arrangement of the oscillator.

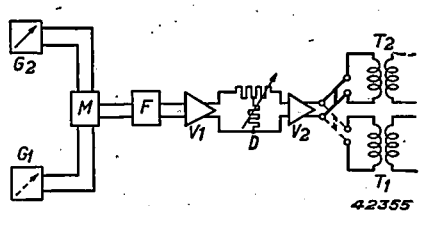


Fig. 2. Simplified diagram of the oscillator. G_1 = fixed frequency oscillator; G_2 = variable frequency oscillator (the frequency of G_1 can also be varied to a certain extent for trimming purposes); M = mixer stage; F = filter for suppressing all voltages of which the frequency is over 100 000 c/s; V_1, V_2 = amplifiers; D = calibrated variable attenuator; T_1, T_2 = output transformers.

The oscillator has to meet very stringent requirements from the aspect of "reproducibility" of the frequency setting, in connection with the previously mentioned frequency drift to which the speech frequencies in carrier telephony are subject. The absolute error in the test frequency in the higher channels (72 kc/s) must not be greater than in the lower, *i.e.*, a maximum of about 20 c/s, which means that the higher oscillator frequencies must be capable of adjustment and indication within an accuracy of 0.03% and that, despite variations in

[3]) "A Tone Generator" by L. Blok, Philips Techn. Rev. 5, 276, 1940. The instrument described therein is not required to meet such stringent requirements as those arising in telephony.

temperature etc. during testing, they must not drift more than a fraction of this 0.03%. Due attention to the temperature coefficients and careful disposition of the components in the chassis, with respect to sources of heat, such as the valves and power pack, have made it possible to satisfy these requirements.

The maximum output is 850 mW, but the sizes of the components, especially the amplifier valves, are very much more generous than would be normally expected for such a relatively low output power, and the reason for this is to be found in the special demands made upon them; in telephony all the valves, including those in the measuring equipment, are almost continuously in operation, so that a much longer guaranteed life is necessary than for the average radio valve, which is in use only for a few hours each day. For this reason the valves in the oscillator are of a relatively larger type.

The level-meter

This unit (*fig. 3*) comprises in the main an amplifier suitable for frequencies of 30-100 000 c/s and a double-diode rectifier, followed by a milliammeter

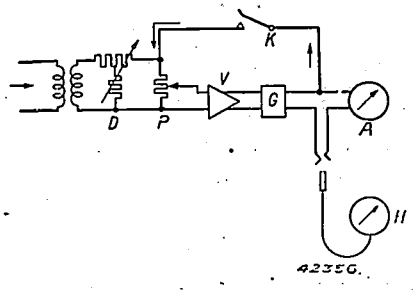


Fig. 3. Diagram of the level-meter. D = calibrated attenuator; P = potentiometer; V = amplifier; G = rectifier; A = meter; H = auxiliary, plug-in hand meter, for use if required. When switch K is closed the feed-back produces oscillation, for use in "internal calibration".

of which the scale corresponds to a range of 14 db, giving easy reading to within an accuracy of 0.1 db. In practice the levels actually differ by much more than 14 db, for, in the diagram of levels *fig. 1*, one channel alone reveals differences of from 5 db above to 60 db below the so-called zero level (1 mW).

In some instances the line repeaters provide amplification to the extent of 65-70 db and the level-meter is therefore also provided with a calibrated, variable attenuator for reduction of the sensitivity of the unit to the required value. The weakest signals to be measured lie 72 db below the zero level (potentials of about 30 μV) and the strongest are 30 db above it (potentials of 70 V); the attenuator is very carefully designed to avoid the possibility that unusually high input voltages might in themselves affect the output and disturb the reading of the instrument. With the same object in view, care is also taken with the wiring, whilst due consideration is given to the reduction of stray impedances, in order to ensure the required flat characteristic over the whole of the very wide frequency range of 30 to 100 000 c/s.

The dial calibration of the milliammeter is correct only in respect of a certain definite degree of amplification from the built-in amplifier, and this setting can be readily checked by means of the feed-back which causes the amplifier to oscillate. The deflection of the meter is governed by the degree of amplification, which is made to correspond to the calibration line on the dial; should the setting be incorrect, the amplification is adjusted by means of the potentiometer P in fig. 3 (internal calibration).

The Analyser

This instrument must be capable of filtering from a composite signal a very carefully determined and readable frequency and of measuring the strength of this frequency. A filter is used that will pass only a very narrow band of frequencies in the region of 370 c/s and the signal under test is converted to 370 c/s by modulation with a carrier wave of variable frequency, see *fig. 4*. The circuit may therefore be regarded as the reception channel of a carrier system in which the carrier wave employed for modulation is variable and of which the channel amplifier admits only one frequency, of 370 c/s, this having been chosen to avoid confusion with possible harmonics of the main frequency.

The fact that an analyser of this type is definitely essential for the measurement of high values of filter-loss can best be illustrated by means of an example. Suppose it is required to measure at 2000

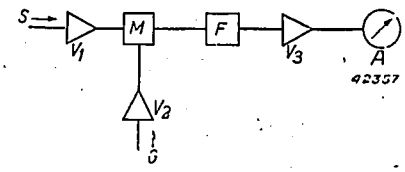


Fig. 4. Diagram of the analyser. S = input for signal under test, of unknown frequency f; G = input for an indicated frequency, variable within 1 or 2 c/s; M = modulator; F = sharp cut-off filter for 370 c/s; V_1, V_2, V_3 = amplifiers. A = measuring instrument to deflect only when G supplies a frequency of $f + 370$ or $f - 370$ c/s.

c/s a high-pass filter of which the cut-off frequency is 3000 c/s, using an oscillator of which the inherent distortion is 0.1%, *i.e.*, the sum of all the harmonics is 60 db below the fundamental. Now,

if a straight-characteristic test-amplifier were used instead of the analyser, it would not be possible to measure a greater quantity of loss than 60 db, seeing that all harmonics of 2000 c/s would pass the filter without any attenuation and produce a signal just as strong as the 60 db damped fundamental frequency. The analyser, on the other hand, first eliminates all harmonics and heavy background noises, after which only the strength of the fundamental frequency of 2000 c/s is measured, this being passed by the filter and exactly reproducing the filter losses at that frequency.

Construction of the equipment

For the daily testing of a large station, the necessary measuring equipment is mounted on a special panel as depicted in *fig. 5*; this is located at a central point in the station and is connected by permanent wiring to all the panels at which regular measurements of level are taken. On each of the latter panels, for instance that of the channel amplifier for the 17-channel system, a jack *A* is provided, connected to the output voltage of the oscillator and a similar jack *B*, connected to the input of the level-indicator. Further, the inputs and outputs of the components on that panel (in this case the channel amplifiers) are all taken to a jack-panel, so that, using only two leads with plugs at the points *A* and *B*, it is possible to check all the components in quick succession.

Measurements are carried out by two inspectors, one of whom makes the necessary adjustments to and takes readings from the test panel, while the other completes the successive connections to the different components at the panel to be tested. The inspector at the latter can also take readings, if desired, by means of a hand meter connected in series with the instrument on the test panel. This is convenient if any adjustments to the panels are to be made, *e.g.* matching of modulators [4]) or readjustment of the amplifiers.

The great rapidity of measurement made possible by this division of labour is of the utmost importance in view of the very numerous measurements that have to be made regularly, especially in the larger stations, where sometimes 10 or 20 carrier systems are mounted side by side.

In smaller stations no test equipment is normally available and this has to be brought to the spot as and when required. It is understood in all measurement work that a call taking place over the channel under measurement shall not be interrupted and, during the time taken for such measurement, the calls are transferred to special spare channels.

Fig. 6 shows a photograph of the analyser, designed as a readily transportable instrument for

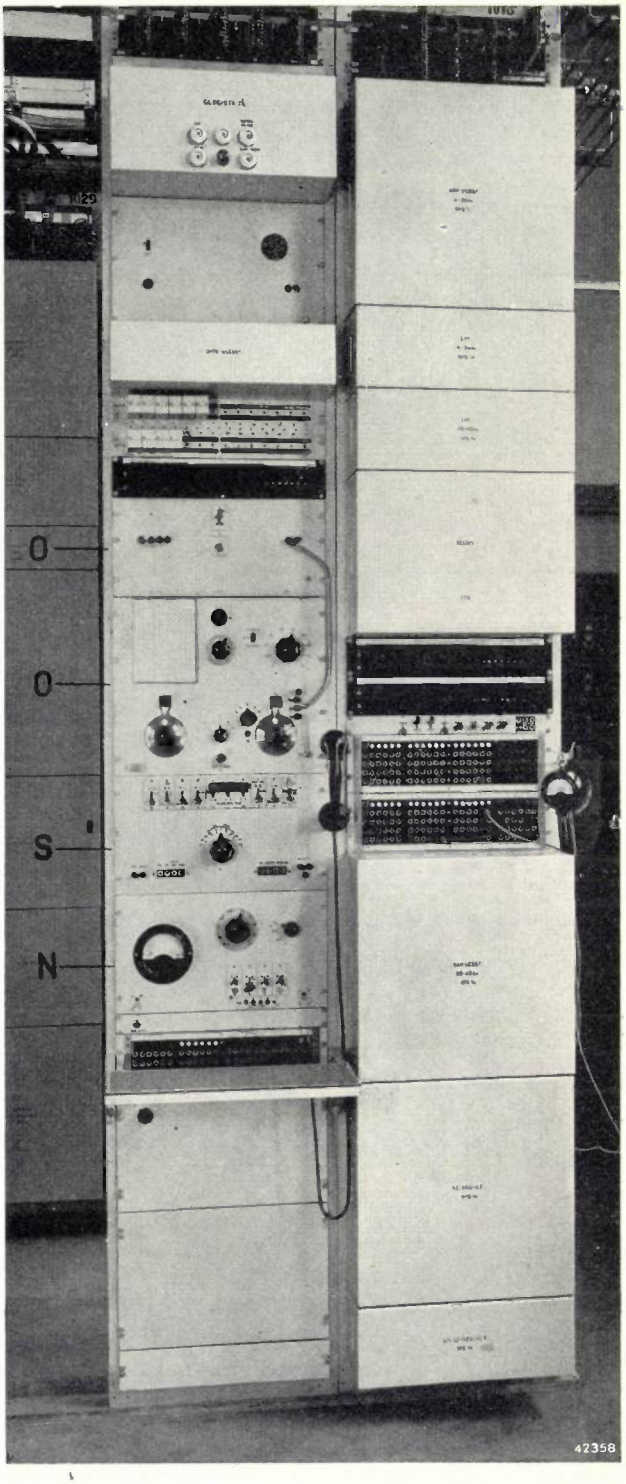


Fig. 5. The test panels, as mounted in a Philips 17-channel carrier station. *O* = oscillator; *N* = level indicator; *S* = switch panel. The right-hand panel contains the channel amplifiers and the illustration shows the jack-panel to which the input and output jacks of all the amplifiers, as well as tappings from the anode circuits, are grouped. On the right-hand side, by the jack-panel, there is a hand-meter for checking the amplifier valves.

[4]) Philip v. Res Techn. 7, 83. 1942,

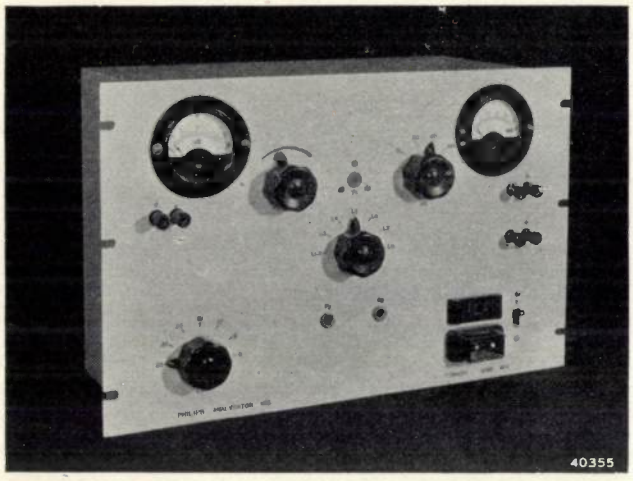


Fig. 6. The Analyser. One of the meters is for checking the symmetrical setting of the balanced modulator, whilst the other indicates in db the strength of the signal to be tested, a calibrated attenuator being provided for this purpose.

mounting in a panel or wooden box, the latter being preferable for locating faults, as it can be connected to any desired point in the system or used for measurements in unattended terminal or repeater stations. The oscillator and level-indicator described above can be similarly built into standard wooden boxes and the latter, mounted in this manner, is shown in *fig. 7*.

Ancillary apparatus used in normal operation

For the transmission of speech over telephony channels embodying a large number of elements (four-wire circuits) it is essential that losses due to reflection in these circuits be reduced to the greatest possible extent, which means, in effect, that each four-wire circuit must be terminated by its so-called characteristic impedance and, to simplify matters, a certain amount of standardisation has been introduced. In low frequency telephony equipment (up to 10 000 c/s) each of the elements in the connection must have an impedance of 600 Ohms, or, in the case of higher frequencies (up to 200 kc/s), 150 Ohms. Then, in order to determine the impedance of any one component this must be similarly terminated, exactly as under working conditions. If the oscillator *0* (*fig. 5*) is substituted for the preceding components and the level-indicator *N* for the subsequent elements, the output impedance of the oscillator and the input impedance of the level indicator must conform to the prescribed value of 600 or 150 Ohms. Again, it may be desired to measure the level of a line whilst in normal operation: the level indicator, or, if required, the analyser is then connected in parallel with that line, in which event the introduction of the meter into the circuit must not produce any alteration in the speech circuit, *i.e.* the level must not be changed more than 0.1 db. The input impedance of the level indicator is then set to a very high value, *i.e.* 30 000 Ohms for a line of 600 Ohms.

Fig. 7. Level indicator, mounted in a robust wooden box. The construction, as indeed that of all the equipment described, must be extremely rigid to avoid distortion of the chassis, buckling of the front panel etc. and thus exclude any possibility of consequent changes in the electrical properties.

The switch panel S shown in *fig. 5* incorporates the necessary equipment for the adjustment of the oscillator output and level indicator input, as required, to the impedance values mentioned (which must be constant throughout the whole frequency range of 30 to 100 000 c/s) [5].

Further, by means of a calibrated attenuator it is possible to reduce by 60 db the transmitting level, normally of 1 mW, but which for purposes of checking the performance of the limiters and certain other components may be 6 mW. This facility is desirable notwithstanding the availability of the calibrated attenuator in the oscillator itself, since in some instances the carrier feed-voltage of the whole unit is employed as test signal instead of the

[5]) The impedance of 4-wire circuits is seldom exactly 600 or 150 Ohms over the whole frequency range, due amongst other things to the presence of unavoidable capacities and inductances. It is only in certain parts of the range, *e.g.* in the case of the bandfilters, in the centre of the channel that the impedance exactly meets the required value. Nevertheless, the measurements are carried out with impedances of 600 or 150 Ohms, as these not only always furnish comparable results but the arrangement also approximates working conditions, in so far as attenuators for balancing or adjustment are often used, of which the impedance is actually quite independent of frequency.

oscillator voltage, and the 17 jacks seen on the small panel in the lower part of the equipment (fig. 5) serve this purpose.

The transmitting and receiving cables forming the connections between the switch panel and the various other panels will in some cases carry signals of widely different levels, and checking of the amplifiers may reveal differences of as much as 70 db; in order, therefore, to ensure the required accuracy of measurement, the transmitting and receiving cables must be very carefully screened from each other. The method of earthing the screening, too, is of great importance. According to whether the input or output of a component to be tested is balanced or not with respect to earth, the screening, which is provided in each case with separate leads, is connected to different earthing points: the jacks referred to above are then so arranged that when the plugs are inserted the screens on the relative wiring are automatically inter-connected.

All the adjustments described in the foregoing, such as the impedance of input or output, the transmitting level, or earthing, are accomplished by means of a group of switches on the switch-panel; the use of keys or special tumbler switches, to permit of rapid action and to provide a simple means of reversing numerous contacts simultaneously, is a typical feature of telephone equipment construction. In the design of these keys, however, due consideration must be given to the fact that at higher frequencies, in the region of, say, 100 000 c/s, they may tend to form a shunt across the 150 Ohm line-impedance and upset the matching.

In conclusion, it should be noted that the two inspectors, working together on the measurements in the manner described, may well be some hundreds of miles apart, for example when measuring the losses in the cables or inter-station connections.

To meet such contingencies, the switch panel includes the necessary connections to permit of the transmission of instructions and reports telephonically over the actual line under test.

Other measurements

A few additional maintenance tests mentioned in the opening paragraphs will now be briefly discussed.

Cross-talk measurements

Measurements of cross-talk are virtually measurements of level, with this difference that the signal is applied to the input of one speech channel, whilst the level of the resultant interference is measured at the output end of another channel. For this purpose it is necessary to select different frequencies according to the degree of annoyance caused to the listener, and this process of selection is made possible by a filter possessing a very definite characteristic; the equipment incorporating an instrument of this type is known as a psophometer [6]).

In general, cross-talk measurements only become necessary in the event of an actual defect, such as damage to a cable, which would disturb the balance between the pairs of conductors. In these cases the

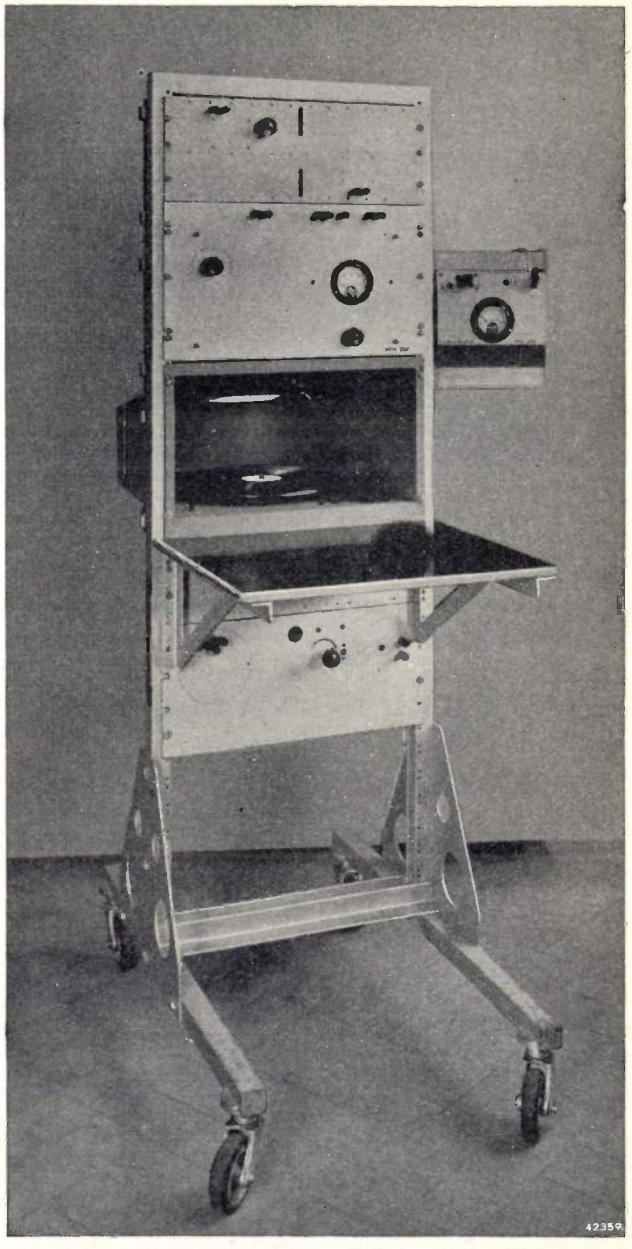


Fig. 8. Transportable equipment for cross-talk measurement. Top: the psophometer [6]); below: the play-back apparatus for the speech records. On the right: the volume-meter.

[6]) See Philips Techn. Rev. 7, 108, 1942.

number of measurements to be taken is extremely large, 552 individual measurements being necessary for the checking of mutual cross-talk between 24 pairs of conductors in a carrier cable.

Fig. 8 depicts the equipment used for this kind of measurement. The outgoing signal usually consists of speech, and to provide speech of a readily reproducible character and level specially selected gramophone records are employed, of which the play-back strength is controlled by means of a specially adapted volt (volume) meter. The play-back unit and psophometer are mounted in a transportable panel, to enable cross-talk measurements to be made for the purpose of localising faults at various points in the system.

Measurements of the strength of hum and background noises are also carried out with this instrument.

Checking the amplifier valves

In a terminal station serving a number of carrier telephony lines many hundreds of amplifier valves are used, each of which in the course of time deteriorates and requires replacement before an actual interruption in the line occurs. These valves, then, are inspected at regular intervals, for instance once per week, and it is sufficient in the case of most of the valve types to measure the anode current under normal working conditions. So that this may be done rapidly, two tappings are taken from the anode resistance of each valve to a jack and all the jacks associated with each panel of valves are grouped on the common board to which reference has already been made. The different anode currents are read direct from a portable milliammeter, plugged successively into the appropriate jacks. For varying types of valve the anode tappings in question have to be arranged differently, in such a manner that the full deflection of the meter can be used in each case, and this would normally mean a special scale for each different type of valve. Since the Philips Carrier Equipment under review employs only two distinct types, however, it is only necessary for the meter to be fitted with two scales. To facilitate reading these scales are printed in different colours and each jack is marked with its appropriate colour.

The anode current of some types of valve does not necessarily furnish adequate evidence of its proper working, a more important feature being the slope, and this is obtained by measuring the anode current at two different values of grid bias. The jacks relating to such valves are therefore provided with a key, by means of which the bias can be set to a value other than the normal and this second measurement can be taken.

Interfering frequencies; distortion measurements

As we have already shown, the modulators in a carrier system not only supply the desired sidebands but numerous other, undesired frequencies as well, and in certain circumstances the latter may tend to pass into the particular channel or even into other channels, causing troublesome whistles. The highest degree of accuracy is often needed to identify such frequencies and trace them to their source, and this work can be very conveniently carried out by means of the analyser described in the preceding paragraphs.

The same instrument is also used for the measurement of non-linear distortion; a particular frequency is applied to the suspected component and the strength of the resultant second and third harmonics etc. is then measured. Special equipment can, of course, be constructed for the direct measurement of distortion, but as these tests lie virtually on the border line of normal maintenance measurements in carrier telephony it has not been found necessary to develop such special instruments for the particular system mentioned above.

VOL. 8 No. 9, p. 257-288 SEPTEMBER 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

THE MULTIREFLECTION TUBE
A NEW OSCILLATOR FOR VERY SHORT WAVES

by F. COETERIER.

621.385.83

The multi-reflection tube is a reflex oscillator differing essentially in construction and functioning from the ordinary reflex oscillators. With the latter effective use can be made of only one reflection in the space between the modulator system and the repeller electrode. In the multi-reflection tube, on the other hand, the electrons execute a pendulum motion about the modulator system with a constant periodic time, so that each time they pass the modulator system they induce a current of the correct phase, which results in a much higher efficiency than that of the usual reflex oscillators. This constancy of the periodic time is achieved by a suitable choice of the potential gradient between the modulator system and the repeller electrode and cathode respectively. The multi-reflection tube can be used for pulsating operation as well as for continuous wave operation.

Among the modern oscillators the magnetron and velocity-modulated tubes occupy the most important places in the field of decimetre and centimetre waves. Velocity modulated tubes have recently been discussed in detail in this periodical[1]); their action is based on the electrons first being modulated in velocity and the velocity-modulation then being converted into a density modulation due to the electrons overtaking each other in a field-free space. In the so-called reflex oscillators the effect of the electrons overtaking each other also plays an essential part, but the conversion of velocity-modulation into density-modulation no longer takes place in a field-free space, the electrons being "reflected" by an electrostatic field back to the modulating system that modulated them in velocity. The principle of the reflex oscillator will be compared with that of the "ordinary" velocity modulated tube in more detail farther on. Here it is to be remarked that one great objection to the reflex oscillators hitherto known was their low efficiency. A new reflex oscillator, the "multi-reflection tube", has now been developed by Philips, the principle and construction of which will be described in this article. This new reflex oscillator has a high efficiency, mainly due to a suitable choice of the above-mentioned electrostatic field.

Principle of the reflex oscillator

In order to show clearly the fundamental difference between the action of the reflex oscillator and that of the ordinary velocity-modulated tube we shall outline the action of the latter with reference to

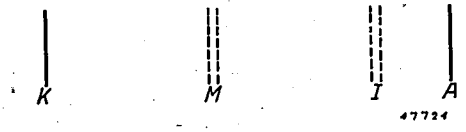


Fig. 1. Diagram showing the principle of a velocity modulated tube. K cathode, M grids of the modulator between which the electrons are modulated in velocity by an A.C. voltage. Between M and I is a field-free space for overtaking, in which velocity-modulation is converted into density-modulation; I inductor to which density-modulated electrons give off their high-frequency energy.

fig. 1; for all details we refer to the article already mentioned [1]). The electrons of a beam emitted from the cathode are first brought to a high velocity, which is the same for all. Between the grids of the "modulator" there is an A.C. voltage, and here the electrons undergo small variations in velocity, the magnitude of which depends on the moment at which the electrons enter the modulator. Next the electrons enter a space that is free of field, the "drift space". Because of the variations

[1]) F. M. Penning, Velocity-modulation Valves, Philips Techn. Rev. 8, 214, 1946.

in velocity the electrons which started later will be able to overtake in this space those which started sooner, this resulting in variations in density along the beam. At a point where the density reaches a high value there are the grids of the induction system, the "inductor". Here, owing to the fluctuations of density, high-frequency alternating currents are induced [2]). By applying a feedback from the inductor to the modulator an oscillating system is obtained.

The conversion of velocity-modulation into density-modulation can be brought about otherwise than by causing the electrons leaving the modulator to overtake each other in a field-free space. In the reflex oscillators this is accomplished, as already stated, by causing the electrons to strike up against a retarding electrostatic field, so that they are ultimately thrown back upon the modulator (*cf. fig. 2*), obviously resulting in density modulation.

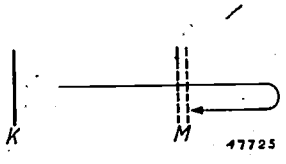


Fig. 2. Diagram showing the principle of a reflex oscillator. K cathode; M modulator which here also plays the part of an inductor. To the right of M is an electrostatic retarding field which causes the electrons to reverse their direction. (How the retarding field is obtained is not indicated in the figure.) Since the transit time of the electrons in the retarding field depends upon their initial velocity, and this is modulated by M, the electron current returning to M is modulated in density. By a proper choice of the phase difference between this current and the A.C. voltage, the high-frequency energy can be made to be given off to M.

The transit time of an electron in the retarding field — *i.e.* the time that the electron takes to travel from and back to the modulator — depends not only on the properties of the retarding field but also on the velocity with which the electron leaves the modulator. The electrons which leave the modulator at equal intervals but at different velocities will therefore in general arrive back at the modulator at unequal intervals, which means that the originally constant current density is in fact converted into a time-dependent current density, and if the modulating voltage is periodic that current density will be a periodic function of the time.

Now if it is desired to use this method of converting velocity-modulation into density-modulation to obtain an oscillator, it is only necessary to arrange for the modulator to act the part of inductor for the returning density-modulated electrons. This is accomplished by a suitable choice of the phase difference between the modulating voltage and the current density of the electrons returning to the modulator.

The fundamental similarity with the ordinary velocity-modulated tubes thus consists in the fact that, by means of a relatively long transit time, either in a field-free space or in a retarding field, the electrons are given the opportunity of converting their velocity-modulation into density-modulation. The fundamental difference consists in the modulator and the inductor of an ordinary velocity-modulated tube having become identical. Henceforward this arrangement, which in reflex oscillators provides both for the velocity-modulation of the electrons and for the induction of their high-frequency energy, will be called the "modulator system" (*cf.* footnote [2])).

Reflex oscillator with a linear retarding field

In the foregoing our remarks have been confined to generalities and no precise definition has yet been given of the kind of retarding field used in the existing reflex oscillators. In point of fact, what has been used almost exclusively so far is a linear retarding field, by which is meant the simple situation shown diagrammatically in *fig. 3*. Here the retarding field is formed by introducing at some distance behind the modulator system a "repeller electrode" at a potential a few hundred volts lower than the cathode potential: the potential variation between the modulator system and the repeller electrode is then linear, at least if space-charge effects are to be disregarded, which will usually be the case here.

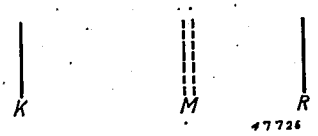


Fig. 3. Diagram showing the principle of a reflex oscillator with linear retarding field. K cathode; M grids of the modulator system (which plays both the part of modulator and of inductor); R repeller electrode to which a voltage is applied lower than that of the cathode. The variation of potential between M and R is linear, except for possible space-charge effects.

In a reflex oscillator with a linear retarding field an electron moves in the retarding field like a ball thrown vertically into the air (we disregard

[2]) In itself the action of the inductor is nothing else than that of a second modulator, the only difference being that the density of the electron beam when passing through the first modulator is homogeneous and inhomogeneous when passing the second.
As a consequence the change in the kinetic energy of the electrons in the first case is on an average zero per cycle, whilst in the second case this variation per cycle is negative, so that a surplus of energy is taken up by the inductor.

space charge and air resistance). If v is the velocity with which the electron coming from the cathode reaches the modulator system and Δv is the (positive or negative) velocity modulation which the electron thereby undergoes, its transit time in the retarding field is

$$t_l \text{ proportional to } v + \Delta v.$$

Therefore the accelerated electrons ($\Delta v > 0$) are overtaken by the retarded electrons ($\Delta v < 0$) which started later, thus just the reverse of the case with the continuously forward-moving electrons in the ordinary velocity-modulated tube, where the transit time from the modulator to the inductor is

$$t_l \text{ proportional to } \frac{1}{v + \Delta v}$$

When $|\Delta v/v| \ll 1$, *i.e.* at small depths of modulation, however, there is a far-reaching analogy between the two types of velocity-modulated tubes, considering that then

$$\frac{1}{v \pm |\Delta v|} \approx v \mp |\Delta v|.$$

This analogy will be more clearly shown in the discussion of the efficiency.

Efficiency of the reflex oscillators

Let us suppose that a reflex oscillator (not necessarily with a linear retarding field) is oscillating. The periodically varying voltage on the modulator system and the current which is induced in it by the density-modulated electron beam will in general contain a number of harmonic components. We assume, however, that the oscillator circuit (for instance a cavity resonator or a Lecher system), of which the modulator system forms a part, is tuned to the fundamental frequency, so that we need only consider the fundamental components of the voltage and of the induced current — with the amplitudes V_1 and I_1 respectively. Further we denote by V_0 the electrostatic voltage of the modulator system with respect to the cathode, and by i_0 the unmodulated electron current emitted by the cathode. Assuming, further, that the induced current is in phase with the voltage, which is the most favourable situation for efficiency, the following expression holds for the efficiency η:

$$\eta = \frac{{}^1/_2\, I_1\, V_1}{i_0\, V_0} \quad \ldots\ldots \quad (1)$$

The same formula also holds for ordinary velocity-modulated tubes, it being understood, of course, that V_1 and I_1 are then the voltage on the inductor and the current induced in it. [3])

In order to obtain a high efficiency it is therefore necessary, according to equation (1), to ensure that the ratios I_1/i_0 and V_1/V_0 are as large as possible.

The maximum value of I_1/i_0 at a small depth of modulation is found to be exactly the same for an ordinary velocity-modulated tube as for the reflex oscillator with a linear retarding field (*cf.* the remark at the end of the previous section), being in both cases equal to twice the maximum of the Bessel function $J_1(x)$, which amounts to 0.58 [1]). Now with an ordinary velocity-modulated tube V_1 can be chosen equal to V_0 (it cannot be chosen larger because the electrons would then be thrown back before they had reached the inductor), so that the maximum efficiency in that case becomes $\eta_{\max} = {}^1/_2 \cdot 1.16 \times 1 = 0.58$.

On the other hand, in the case of the reflex oscillators in general, and thus also those with a linear retarding field, the ratio V_1/V_0 must be smaller than unity, having regard to the starting of the oscillation. The maximum efficiency of a reflex oscillator with a linear retarding field is therefore considerably less than 0.58.

Due to all kinds of incidental factors, however, which have been disregarded in the derivation of these theoretical maximum efficiencies, in practice the latter are not even approximated. The efficiencies actually attained in practice with ordinary and with reflex velocity-modulated tubes are of the order of 0.2 and 0.1 respectively.

The multi-reflection tube: its principle and efficiency

The objection, just mentioned, of the low value of V_1/V_0 in reflex velocity-modulated tubes can, however, be overcome indirectly, thanks to a factor which so far has been left out of consideration here, namely what happens to the electrons which after reflection have given off part of their kinetic energy to the modulator system. These electrons, of course, first travel farther on in the direction of the cathode, but ultimately they are driven back again to the modulator system. The question then arose whether it might not be possible to improve efficiency by causing the electrons to swing back and forth around the modulator system in such a way that each time they pass through they give off as much of their energy as possible. This was indeed found to be possible and in the multi-reflection tube we

[3]) The efficiency of the ordinary velocity-modulated tube is discussed in detail in the article already referred to [1]). The readers' attention is called to the fact that in that article our quantities V_1 and I_1 are denoted by V_2 and I_2 respectively.

have a practical realization of just such a possibility.

Let us consider this question somewhat more closely. Each time it passes through an electron gives off to the modulator system an amount of energy equal to $eV_1 \sin(2\pi t_1/T)$, where e is its charge, T the periodic time of the voltage and t_1 the moment of passage. The factor V_1/V_0 in formula (1) for the efficiency is thus equal to the ratio of the maximum energy which an electron can give off to the modulator system in one passage, namely eV_1, to its original energy eV_0. If, now, the electrons are reflected n times by the retarding electrode and the cathode alternately before they disappear from the beam for some reason or other (for instance through divergence), the situation from the point of view of an electron is as if in equation (1) nV_1/V_0 took the place of V_1/V_0. Therefore, if from now on we take I_1 as indicating the amplitude of the induced current when in a reflex tube an electron is reflected only once, then with n reflections the same tube will have an efficiency n times as great:

$$\eta_n = {}^1/_2 \frac{I_1}{i_0} \cdot \frac{nV_1}{V_0} \quad . \quad . \quad . \quad . \quad . \quad (2)$$

provided the induced current is in phase with the voltage at every passage of the electrons through the modulator system, and its amplitude I_1 remains unaltered. We shall return to this requirement shortly. First we will consider the situation "from the point of view of the modulator system". From that point of view the amplitude of the A.C. voltage does not depend upon the number of reflections, but it is the amplitude of the induced current which becomes n times as large — still under the proviso just mentioned, which might be expressed by writing equation (2) as follows:

$$\eta_n = {}^1/_2 \frac{nI_1}{i_0} \cdot \frac{V_1}{V_0} \quad . \quad . \quad . \quad . \quad . \quad (2a)$$

Now what is implied by the above-mentioned requirement for equation (2) to be valid? Obviously it means in the first place that the periodic time of all the electrons must be the same (in that case once the density modulation has been obtained it will not change during the pendulum motion), and in the second place, that the periodic time must remain constant, in spite of the energy of the electrons diminishing each time they pass through the modulator system (the induced current will then always remain in phase with the voltage). In other words, the condition in question will only be satisfied when the electrons execute a (damped) harmonic pendulum movement about the modulator system with a periodic time that is the same for all electrons. Now it is known that such a motion can only take place when the particle is attracted by a centre of force with a force proportional to the distance; the potential then changes as the square of that distance and the transit time of a particle does not then depend upon its velocity at the position of the centre of force. Thus if the potential to the left and right of the modulator system had a parabolic slope, a density maximum once obtained would travel back and forth without being smoothed out, and would always pass the modulator system in the correct phase. With any other retarding field — especially with the above-mentioned linear retarding field — a density maximum becomes more and more smoothed out during the motion, so that I_1/i_0 soon becomes very small, and there can be no question of an improvement in efficiency.

By introducing the parabolic retarding field, however, we should be rejecting the good with the bad. We wanted to introduce the parabolic field because in that field the density modulation of the electrons dit not change during the motion. In particular, therefore, an electron beam originally unmodulated in density would always remain unmodulated in a parabolic field; in other words I_1/i_0 would not only be small, but it would be exactly zero.

A solution of this dilemma could be sought by dividing the problem into two, namely by first modulating the electron beam in density in some way or other as well as possible, *i.e.* by providing for the largest possible value of I_1/i_0 in equation (2), and then causing the modulated beam to travel back and forth in a parabolic field, so that the factor nV_1/V_0 in equation (2) also assumes a favourable value without the first mentioned factor being changed.

This procedure is now employed in the multi-reflection tube in a remarkable manner, as will be explained in the following.

Let us return for a moment to the case of a single reflection in some retarding field or other; for the present, therefore, we will consider only the variation of the potential to the right of the modulator system (to the left of it is the cathode). For every retarding field there is in general a different maximum value of I_1/i_0, so that two questions arise:

a) with what retarding field do we get the greatest of all those maximum values?
b) How great is that greatest value?

The key to the situation — and on this is based the principle of the multi-reflection tube — is the

following: The optimum retarding field referred to in question a) provides — as we shall soon see — not only that the electrons receive the greatest possible density modulation upon the first reflection, but also that this automatically remains unchanged in the following reflections, provided the field to the left of the modulator system is parabolic.

We shall now investigate question a) more closely. We call i_1 the amplitude of the fundamental component of the modulated electron current at the position of the modulator system after the first reflection, while I_1 was the amplitude of the fundamental component of the induced current. Since $i_1 = I_1$, we may consider i_1/i_0 instead of I_1/i_0. The ratio i_1/i, *i.e.* the density modulation of the electron current, — as we have already stated — is determined entirely by the transit times of the electrons in the retarding field, *i.e.* the time taken by the electrons to travel once back and forth from and to the modulator system. In a given retarding field the transit time is again determined by the velocity with which the electron enters that field. Since this velocity varies periodically with the time t_1 at which the electron enters the retarding field, the transit time is also a periodic function $f(t_1)$ with the same period. The form of the function $f(t_1)$ thereby depends on the retarding field, each $f(t_1)$ corresponding, in general, to a different retarding field. Therefore, for question a) we may substitute the two following questions:

(a_1) With what $f(t_1)$ is i_1/i_0 a maximum?

(a_2) What must be the form of the retarding field in order that the transit time will vary according to the optimum $f(t_1)$ referred to in question (a_1)?

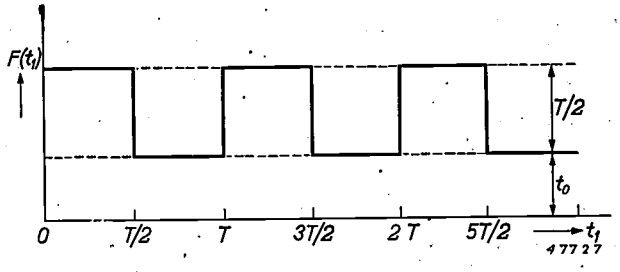


Fig. 4. The transit time $F(t_1)$ of an electron in the optimum retarding field as a function of the time t_1 at which it leaves the grids of the modulator system. The transit time of the electrons retarded by the modulator system is indicated by t_0. The transit time of the accelerated electrons is then $t_0 + T/2$ (T is the period of the A.C. voltage on the modulator system).

We shall not go into the (not very difficult) mathematical considerations which furnish an answer to question (a_1) but will confine ourselves to stating the result. The desired function, which we shall denote by $F(t_1)$, is represented in *fig. 4* [4]). Except for a constant it is given by a so-called square sine with an amplitude amounting to one quarter of its period T (the function $F(t_1)$ having the dimension of a time).

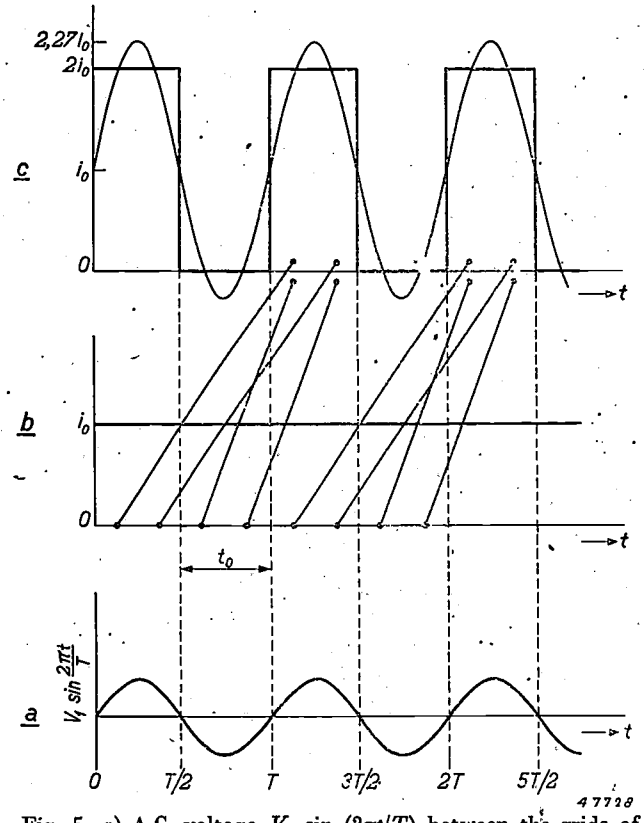


Fig. 5. *a*) A.C. voltage $V_1 \sin(2\pi t/T)$ between the grids of the modulator system as a function of the time.

b) The unmodulated electron current $i = i_0$ as a function of the time; the black dots along the t axis represent the electrons coming from the cathode and reaching the modulator system at equal intervals.

c) The modulated current of the electrons returning to the modulator system as a function of the time, where the transit time function $F(t_1)$ is given by the curve of fig. 4. The heavy line is the total current, the thin line the sum of the time-independent component i_0 and the fundamental component $(4i_0/\pi) \sin(2\pi t/T)$. The black dots above and below the t-axis represent respectively the accelerated and the retarded electrons. It may be seen that at equal intervals during one half of the cycle one accelerated and one retarded electron return simultaneously to the modulator system, and during the other half of the cycle no electrons return.

This has led to the discovery of a very important property of the optimum retarding field. This retarding field must evidently be such that 1) all electrons which are retarded by the modulator system have the same transit time t_0 and 2) all the accelerated electrons also have the same, longer, transit time $t_0 + T/2$. As a result the current density of the electrons returning to the modulator system, as a function of the time, will also be a

[4]) Strictly speaking, this is not the most general form which $F(t_1)$ can have. For our purpose, however, the consideration of the $F(t_1)$ represented in fig. 4 is sufficient.

square sine; it is alternately zero for one half cycle and during the other half cycle equal to twice the unmodulated current density.

This can easily be seen from *fig. 5*. The phase difference between the current density and the A.C. voltage is equal to $2\pi/T$ times the transit time t_0 of the retarded electrons. In order that the energy given off by the returning electrons to the modulating system may be as large as possible, the following equation must be satisfied:

$$t_0 = (l + 1/2)\,T \quad (l = 0, 1, 2, ..) \quad . \quad . \quad (3)$$

Given $F(t_1)$, it is possible to calculate the maximum value of $I_1/i_0 = i_1/i_0$ referred to in question b), and then is found to be $4/\pi = 1.27$ (*cf.* fig. 5).

In itself this is only a slight improvement compared with the case where the retarding field is linear: $(1.27 - 1.16)/1.16 = 9.5\%$.

What we are interested in, however, is the "optimum" retarding field that corresponds to the specific transit-time function $F(t_1)$, mainly because, as already stated, it allows of the effective employment of the pendulum motion of the electrons.

What, then, is this optimum retarding field that is realized in the multi-reflection tube?

If v_0 is the velocity of an unmodulated electron, all the retarded electrons will reverse their direction to the left of point R (see *fig. 6*), for which

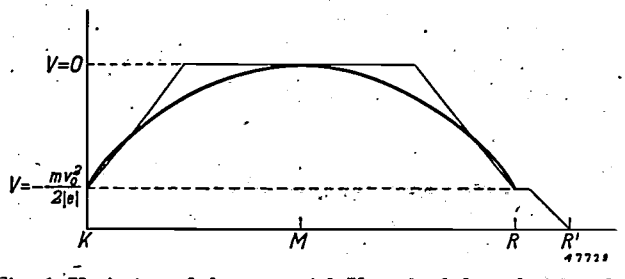


Fig. 6. Variation of the potential V to the left and right of the modulator system M in a multi-reflection tube; K cathode. The parabola is approximated by the three sections of straight line. The potential slope to the right of R (the broken line RR') provides for the accelerated electrons to have a transit time $T/2$ longer than the retarded electrons. The potential difference between M and K (or R) multiplied by the charge e of an electron is equal to the kinetic energy $mv_0^2/2$ of an unmodulated electron.

the voltage, with respect to the modulator system, is $-mv_0^2/2e$ (m and e being the mass and charge of the electron respectively).

It being essential that all retarded electrons shall have the same transit time, the variaton of potential between the modulator system and point R, in accordance with what has alread been stated, must be parabolic.

If the variation of potential to the left of the modulator system is given by the second symmetrical half of the same parabola, then upon subsequent reflections a retarded electron will pass the modulator system at equal intervals $(k + 1/2)T$ (see equation (3)), in spite of the fact that its energy decreases each time. With a multi-reflection valve the parabola in question is approximated by the tripartite line in fig. 6.

The accelerated electrons, *i.e.* the electrons which on leaving the modulator system have a velocity greater than v_0, will on the other hand continue to the right of R. Thanks to this circumstance, the requirement that the accelerated electrons must all have a transit time $T/2$ longer than the retarded electrons can be satisfied in a simple way by choosing a suitable variation for the potential to the right of R: the accelerated electrons are made to "wait" at the right of R. A potential variation which causes the electrons to wait in approximately the desired way — and which is realized in the multi-reflection tube — is represented in fig. 6. The transit time of an accelerated electron in the retarding field to the right of the modulator system is thus equal to

$$(l + 1/2)T + 1/2\,T = (l + 1)T,$$

so that such an electron, upon passing the modulator system after the first reflection, loses exactly as much energy as it took up when passing in the opposite direction: thus it enters the space to the left of the modulator system with the velocity v_0. From that moment onwards the originally accelerated electron will behave exactly like an originally retarded electron: upon subsequent reflections it will never be able to enter the region to the right of R. With reference to fig. 5, we have seen that each time after the first reflection one accelerated and one retarded electron return simultaneously to the modulator sytem subsequently. Such a pair of electrons will therefore always pass the modulator system together, and this means that the density modulation obtained upon the first reflection — the square sine of fig. 5 — will actually remain unchanged for all subsequent reflections.

What efficiency is to be expected from a multi-reflection tube on the basis of the theory outlined above? In a multi-reflection tube the optimum retarding field is only approximated. After a certain number of oscillations even in a multi-reflection tube the electrons will become out of phase, or will be removed from the beam by other causes. Actually this falling out of phase occurs quite suddenly: it can be demonstrated that the electrons oscillate with a constant periodic time until their velocity has fallen to about one half, and thus their energy

to a fourth, after which they give off practically no more energy; unless special precautions are taken they might even begin to consume energy. Practically speaking, the current induced in the modulator system is therefore formed — in agreement with equation (2a) — by the superposition of n currents, as represented graphically in *fig. 7*; n here depends upon the value of V_1/V_0.

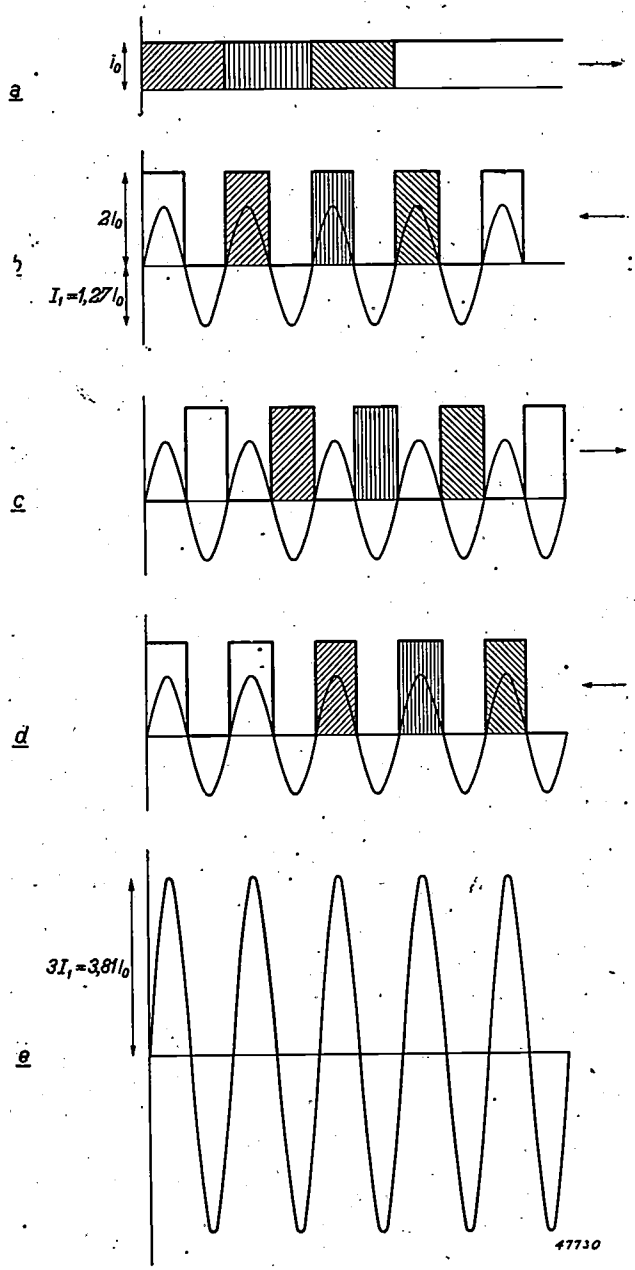


Fig. 7. How the current induced by the electrons is formed. *a*) represents the non-density-modulated current of the electrons passing the modulator system for the first time after leaving the cathode. At each subsequent passage of the electrons in one direction or the other (indicated by arrows) between the grids an A.C. is induced the variation of which with time is represented in *b*), *c*) and *d*); the fundamental component of the induced current is drawn separately in each case. The sum of these currents, *i.e.* the fundamental component of the total induced current, is shown in *e*). Similarly cross-hatched areas represent the currents induced by the same group of electrons.

For a calculation of the efficiency it is to be noted that:

$$\frac{nV_1}{V_0} = \frac{E' - E''}{E'}; \quad \ldots\ldots \quad (4)$$

where E'' is the value of the kinetic energy of the electrons at the moment it falls out of phase, and E' the original kinetic energy. Since $E''/E' = 1/4$ and $I_1 = i_1$, we find for the efficiency, by combining equations (2) and (4):

$$\eta = {}^1/_2 \cdot \frac{i_1}{i_0} \cdot \left(1 - \frac{E''}{E'}\right) = {}^1/_2 \cdot \frac{4}{\pi} \cdot \frac{3}{4} = 0.48.$$

It is found, however, that some multi-reflection tubes can attain efficiencies up to 50%. Even if the error of measurement were of the order of 4% the result would still be doubtful, because the circuit losses and radiation losses certainly cannot be ignored. We shall revert to this surprising result later.

Construction of the multi-reflection tube

On the basis of the above theoretical considerations we shall now describe a practical form of construction of the multi-reflection tube. In the first place there must be a parabolic field or something closely resembling it. As already stated in the foregoing, the parabolic field can be sufficiently approximated for our purpose in a simple way by the combination of an equipotential field and a field with a linearly decreasing potential. Expressed in a different way this means the combination of

1) a space free of field in which the retarded electrons have a longer transit time than the unmodulated electrons, with
2) a linear retarding field in which they have a shorter transit time.

By a suitable choice of the dimensions of the two spaces it is possible to make the two effects compensate each other, so that we actually obtain a practically equal transit time for electrons of divergent velocities, provided these are smaller than the unmodulated velocity.

In *fig. 8* it is shown schematically how the modulator system MM' is surrounded by two grids A and A', which are electrostatically at the same potential as the modulator system. Except for the relatively slight high-frequency A.C. voltage on the electrodes of the modulator system, which causes the retardations and accelerations of the electrons, the two grids A and A' bound an equi-potential space, which we shall call the anode space. Beyond A and A' are the spaces with linearly decreasing potential, respectively towards the cathode K and

the reflection electrode R, which is also at cathode potential. A second point for consideration is how, by simple means, the "waiting time" of $T/2$ of the accelerated electrons is to be attained. Here again the combination of an equipotential space with a field of linearly decreasing potential furnishes a

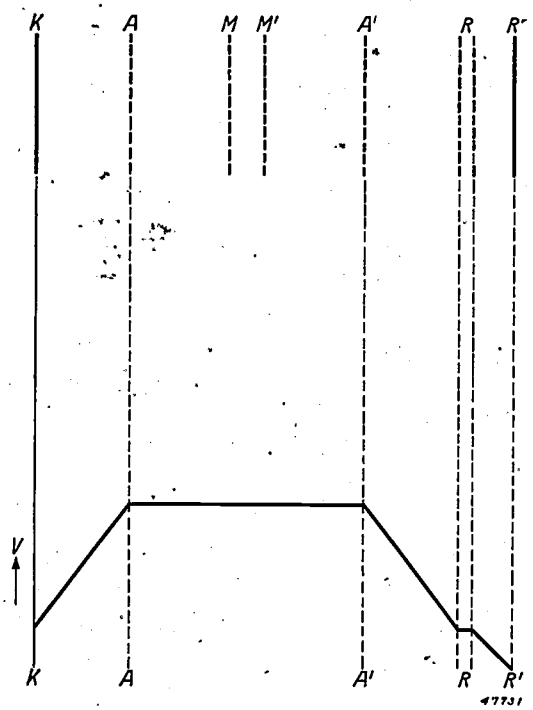


Fig. 8. *a*) Diagram showing the principle of the multi-reflection tube. K cathode; A and A' grids forming the anode; M and M' grids of the modulator system; R' repeller electrode; R grids providing for the desired variation of the potential between the anode and the repeller electrode. *b*) Static potential variation in a multi-reflection tube.

solution. To that end, the electrode R is constructed as a double grid, while behind R there is a second electrode R' to which a variable negative voltage is applied. The short field-free space between the grids of R in combination with the matched linear retarding field between R and R' gives the desired more or less constant waiting time of $T/2$. The electrons that are only slightly accelerated are mainly held up between the grids R, while the strongly accelerated electrons shoot quickly through the double grid but are held up longer in the retarding space RR'. The distances between the electrodes are determined by the choice of the wave length of the oscillations to be excited and of the voltage to be applied between the modulator system and the cathode. The wave length determines the periodic time T of the oscillations, and the voltage, in combination with the distances in question, determines the transit time t_0 of the electrons for one movement to and fro, while the previously derived expression (3)

$$t_0 = (l + {}^1/_2)\, T$$

must be satisfied.

In the diagram of fig. 8 the various electrodes are all represented by grids. The advantage of a grid lies in the practically constant potential over the whole cross-section of the electron beam. A disadvantage, however, is that the grids take up some of the active electrons, and in the case of electrodes with a low potential this is usually accompanied by secondary emission, resulting in electrons with prohibitive velocities getting mixed with the others. Moreover, there is a considerable loss of energy, and the part that is lost is apt to heat the grids to incandescence and cause them to fuse, especially in the case of the electrodes with a high potential. In the practical model the grids are replaced by plates with circular openings. The theory as developed in the foregoing therefore applies actually only for the electrons on the outside of the beam. For the electrons on and near the axis the potential ratios and thus also the transit times are somewhat altered.

We shall not enter into all the mathematical considerations in this paper. Suffice it to say that as far as the axial electrons are concerned the requirement of equal transit times for all retarded electrons and the same transit time $+ {}^1/_2\, T$ for all accelerated electrons is not satisfied. Here the mechanism of overtaking resembles more that of the reflex oscillators with linear retarding field. At the same time it is found that the density distribution of the axial electrons in the returning beam is in phase with the density distribution of the outer electrons, so that the replacing of the grids by plates with circular openings does not, at least, give rise to any fundamental disturbances. It is even not impossible that the action of the axial electrons is of predominating importance in the starting of the high-frequency oscillations in the modulator system, while the outer electrons only become of importance at the higher oscillation voltages and thus determine the efficiency.

Figs. 9a and *b* represent a multi-reflection tube and its components. In *fig. 9c* the construction of the valve is represented diagrammatically. The cathode K is an indirectly heated oxide cathode. In front of the flat emitting side of the cathode is the drilled plate r, which is at a weak negative potential with respect to the cathode and serves to regulate the cathode current. The anode system A, which surrounds the modulator system L, is a rectangular box open at the top and with two circular holes in opposite sides for the passage of the electron beam. The modulator system L con-

sists of two metal strips, likewise provided with holes, forming a quarter-wavelength Lecher system for the high-frequency oscillations to be excited.

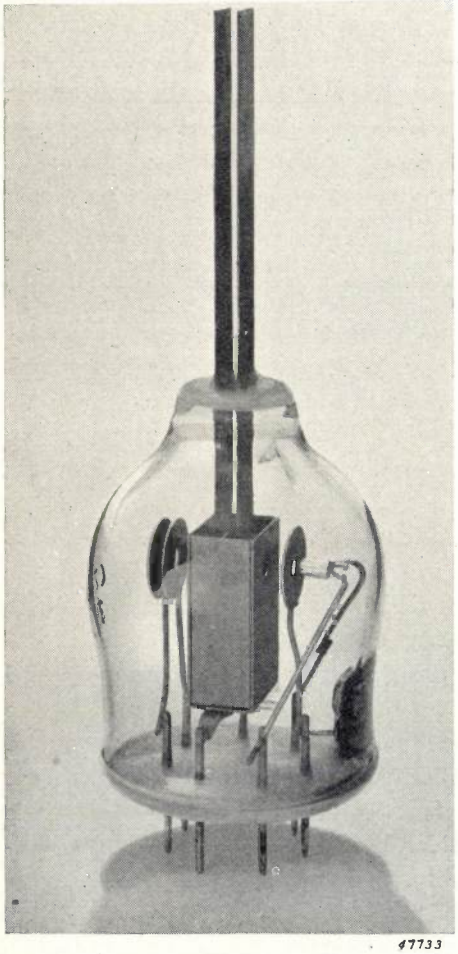
The plates R and R', the first containing an opening, form the reflection system. The plate R is at cathode potential and according to the theory should be constructed as a double grid. The equipotential space enclosed by the double grid, in combination with the retarding field between R and R', has to provide for the constant delay of $T/2$ for all the accelerated electrons. If, however, it is desired to dispense with any form of grid and to deal only with the outer electrons, the plate R will have to be of a certain thickness.

With the arrangement described in the foregoing the Lecher system begins to oscillate with a voltage maximum at the ends of the strips. Opposite these ends there are two other strips of the same width, which take off the high-frequency energy capacitatively and conduct it out of the valve. These strips are passed through the glass with the help of the so-called sintered-glass technique [5]).

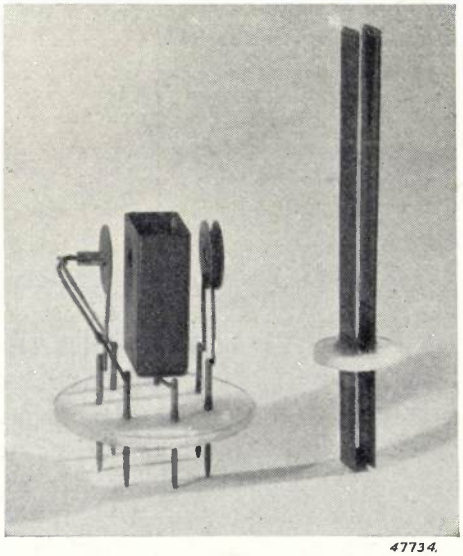
Influence of a focussing magnetic field on the efficiency

For a satisfactory functioning of the multi-reflection tube care must be taken that the electrons are not driven to the electrodes by the lateral forces of the space charge before they have given off the greater part of their high-frequency energy to the modulator system. To that end a magnetic field in the direction of the electron beam is introduced, which opposes the divergence of the beam. At first sight it might be thought that the intensity of this magnetic field could be chosen quite arbitrarily, but that is not so. It is found that the power output of the tube has decidedly certain maxima for certain critical values of the magnetic field. In the discussion of the efficiency we have already stated that with some multi-reflection tubes the efficiency was found to be greater than was to be expected according to the simple theory, where no account is taken of a possible effect of the magnetic field. It will be seen that the explanation of this is closely connected with the critical values of the magnetic field in question.

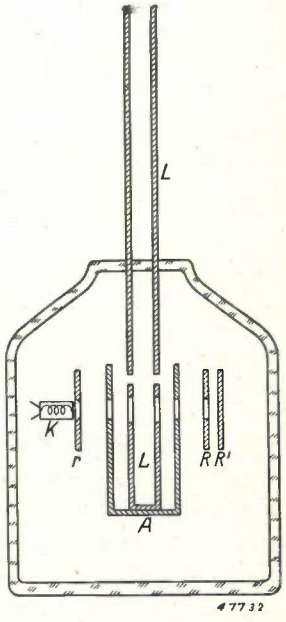
[5]) See E. G. Dorgelo, Sintered Glass, Philips Techn. Rev. 8, 1, 1946.



a)



b)



c)

Fig. 9. *a*) A multi-reflection tube. Diameter of the glass envelope 55 mm. *b*) components *c*) diagram of the multi-reflection tube of *a*); K indirectly heated cathode; r control electrode; A anode; L Lecher system (modulator system); R and R' repeller electrodes.

Let us consider the following experiment:

The Lecher system is detuned, so that the valve cannot oscillate. If the anode voltage is now applied and the current between anode and cathode measured with no magnetic field, we find a certain value. Upon switching on the magnetic field the current falls. This fall is greatest for the above-mentioned critical values of the magnetic field. From a measurement of the magnetic field strength it is found that the critical values are just those at which the cathode is projected electron-optically upon itself, *i.e.* at which the electrons leaving the cathode are brought back to the identical points of the cathode after reflection. The resulting increased space charge in front of the cathode is capable of decreasing the emission of the cathode to a considerable extent, which explains the fact that the fall in the current is a maximum for the critical values of the magnetic field.

What happens now when the tube oscillates at such a critical value of the magnetic field? Owing to their loss of velocity, the originally retarded electrons no longer return to the cathode. After reflection, the originally accelerated electrons lose their excess velocity and consequently return to the cathode, but a closer investigation shows that because of their longer transit time they cannot contribute to the magnetic projection of the cathode on itself. Only the practically unmodulated electrons return to the cathode in a concentrated form and then momentarily suppress the cathode current. This means that the emission is no longer constant, but falls to almost zero twice per cycle. It can be shown that the optimum retarding field then remains the same as for the constant emission. The ratio i_1/i_0, however, becomes greater and consequently also the efficiency. It is not, therefore, so surprising that the efficiency of some multi-reflection valves is found to be higher than 48%, the value found theoretically when disregarding the part played by the magnetic field.

Practical details of the multi-reflection tube

The power that can be yielded by a multi-reflection tube is limited by two factors: the maximum dissipation permissible for the anode system and the maximum cathode current that the oxide cathode can furnish per cm² without detriment to its durability. If 3000 volts is taken for the anode voltage, the area of the cathode can be so chosen that the maximum dissipation for the optimum efficiency is just reached. For the type of multi-reflection tube shown in fig. 9 the effective power is then 15 to 20 watts at a wave-length of 12 cm.

The wave length is fixed for each multi-reflection tube, because of the built-in $\lambda/4$ Lecher system. If, however, it is desired to work with a certain wave-length band it is possible to lead the Lecher system out through the glass wall and to tune it from the outside with a moveable bridge.

The best results are obtained when the transit time t_0 of the retarded electrons for one return movement from and to the modulator system amounts to $3T/2$ or $5T/2$ (this corresponds to $l = 1$ and $l = 2$ in equation (3)). With the same voltages and the same magnetic field a multi-reflection tube can thus work on two wave-lengths, for instance 12 and 20 cm. With a Lecher system that can be tuned oscillations can be set up on these two wave-lengths, and, moreover, the wave-lengths can be varied continuously for a few centimeters above and below; in this case, however, the voltages and the magnetic field must be re-matched each time.

The multi-reflection tube can be used very well as a transmitting valve for all kinds of purposes in the region of very short waves. In particular large types of multi-reflection tubes can be used for the (capacitative) high-frequency heating of all kinds of substances having a strong absorbing power for very short waves.

The multi-reflection tube can be used not only for this ordinary continuous wave operation, but it can be made suitable also for pulsating operation, *i.e.* for the transmission of short wave trains of high intensity. In that case the distances between the electrodes must be adapted to the generally much higher values of the anode voltage used in pulsating operation.

The multi-reflection tube can also be employed as local oscillator in receiving sets working on the superheterodyne principle. Since power plays no great role in receivers, it is of advantage to omit the magnetic field, which of course reduces the efficiency considerably (to about 4-5%) but simplifies the assembly.

LIVING ROOM LIGHTING WITH TUBULAR FLUORESCENT LAMPS

by L. C. KALFF and J. VOOGD. 628.972.6.033

Because of the high efficiency of tubular fluorescent lamps (more than 40 lumen/watt compared with not more than 15 lumen/watt of the incandescent lamp) as well as their good colour-rendering properties, these lamps are suitable in all respects for use in living rooms. The large light flux obtained for a given power consumption makes it possible to realize a more general illumination of the living room, in contrast to the local lighting to which we are accustomed with incandescent lamps. A much larger area of the room can be used for the various activities of the family — reading, writing, mending, etc. At the same time, by placing the tubular lamps against the ceiling and walls instead of in the middle of the room, the whole aspect of the room can be made more spacious. In this article arguments are advanced for the utilization of these possibilities. A series of measurements have been taken in a model room, with a total power input of 150 to 200 W, in which a fixture built into the ceiling and a cornice were used. The merits of different possible light distributions are discussed with reference to isolux diagrams.

Artificial lighting enables us to spend dark evenings more or less according to our own desires. The lighting of homes and especially of the living room has therefore always been of primary importance in practical living as well as in cultural life. It is therefore understandable that eager use has always been made of every new technical possibility for improved interior illumination. The succession of oil lamp and candle, kerosene lamp, gas mantle and electric lamp have marked the most important steps in this development.

Gaseous discharge lamps, which have been developed within the last few decades and among which especially the sodium and mercury-vapour lamps have found many uses because of their high efficiency, were not at first suitable for use in the living room. For that purpose their colour-rendering properties left much to be desired. By the employment of fluorescent substances, however, new possibilities were presented. The result was the familiar tubular fluorescent lamp with an efficiency of about 40 lm/watt and offering a choice in the spectral composition of its radiation by selecting suitable substances for the fluorescent layer [1]). This meant an important advance with respect to colour rendering, even surpassing the modern incandescent lamp; whereas the incandescent lamp must be considered defective in the blue and bluish-green part of the spectrum, with the tubular fluorescent lamp very good colour rendering can be obtained also in those parts.

It is therefore now possible to profit by the above mentioned high efficiency of the tubular fluorescent lamp also for living-room illumination.

In the past developments from oil lamp and candle *via* kerosene lamp and gas mantle to incandescent electric lamps the only thought in each case was to make use of the new invention for "more and easier light", without fundamentally changing the system of illumination. With the introduction of the tubular fluorescent lamp, however, it seems to us that a change in the principles of the lighting system is called for. The jump in efficiency from 15 lm/W for the modern incandescent lamp to 40 lm/W for the tubular fluorescent lamp is so large that an entirely different system of living-room illumination can now be considered. In this article we shall bring forward some conclusions based upon experience and go more deeply into several possibilities.

Localized versus general lighting

When looking around attentively and objectively in all kinds of living rooms one is soon struck by the fact that with the present lighting by means of incandescent lamps much space and freedom of movement is lost. A central lamp with shade hanging at eye level and one or more table or standard lamps form the often voluminous attributes of the usual lighting system, and the result is in most cases a localized lighting which though perhaps more intense is not fundamentally different from that obtained in the times of the candle and the kerosene lamp.

This localized lighting creates a sphere of seclusion which is felt and appreciated by many as an essential element in the peaceful evening hours of home life. On the other hand there are others who regret that the effect of their carefully arranged interiors is to a large extent lost in the evening because of insufficient general illumination, who consider the

[1]) A. A. Kruithof, Tubular luminescent lamps for general lighting purposes, Philips Techn. Rev. **6**, 65, 1941. Since the word "luminescent" is apt to lead to misunderstanding, it is better to speak of "fluorescent lamps".

fixtures at eye level to be a hindrance to the appreciation of the spaciousness of a room or who regard the system of localized lighting as an undesirable restriction of the utilitarian possibilities of the room. The last point of view will be of especial significance when several members of a family are in the living room at the same time and need for their various occupations a light level of 100 to 200 lux, which with localized illumination is only obtained in a very limited part of the room.

Thus on esthetic as well as on practical grounds the need of a good general illumination of the living room is felt by many. The obvious question, therefore, is why such a general illumination is only very exceptionally found.

In order to obtain some insight into this question we shall begin with the fact that in most countries not more than 150 watts of electric power is used for living-room lighting, a limit which is determined, among other factors, by the ratio of the cost of electrical energy to the standard of living of the population. With these 150 watts when incandescent lamps are used a light flux of at most 2200 lumens is available under the most favourable conditions. Suppose that a living room of ordinary size, for instance 4×5 m^2, has to be lighted. We shall consider especially the illumination of the "working surface", *i.e.* a surface at table height above the whole floor area of the room. With fixtures placed close above this working surface one can count on a light efficiency of 0.5, so that in the case in question 1100 lumen are available on the working surface. The lighting will be approximately so arranged that an area of at least $2^1/_2$ m^2 under the central lamp receives an average intensity of 200 lux, while at another spot in the room under a standard lamp about 1 m^2 receives the same intensity. For the remaining $16^1/_2$ m^2 of the working surface there is only $1100 - (200 \times 3^1/_2) = 400$ lumen available. The greater part of the room thus has less than 25 lux at table height. The walls are no better off. One may thus rightly speak of "localized lighting".

What light flux, then, would be needed for an efficient system of general illumination?

From considerations of lighting technology, for such a system the fixtures would be placed high, perhaps even against the ceiling. In that way it is easier to obtain a uniform illumination of the working surface and a stronger illumination of the walls. This, it is true, implies that in general the lighting efficiency will be lower; a value of 0.33 can be counted on. The light flux available on the working surface will thus amount to about one-third of the total light flux employed. We shall now assume once more that a level of 200 lux is required on $3^1/_2$ m^2 of the working surface and that for the remaining $16^1/_2$ m^2 an average of 80 lux is desired. This, however, still does not leave enough light to allow of work involving eye strain to be done in any arbitrary part of the room. But the usefulness of the room as a whole, compared with the case first described, will be very much improved. With these minimum requirements for a general illumination we arrive at a light flux of about 2000 lumen on the working surface, *i.e.*, according to the above, a total of 6000 lumen.

In addition to the increased usefulness of the room with this illumination, the disappearance of fixtures at eye level and the good illumination of the walls should be noted — advantages which, in the light of the arguments already mentioned, contribute towards making general lighting attractive to many.

To change over from the 2200 lumen for localized lighting to 6000 lumen for general illumination would considerably increase the consumer's light bill if he had only incandescent lamps at his disposal, and this is presumably the reason why the general public has never taken such a step. It is now reasonable to assume that the development of the tubular fluorescent lamp will be able to change this. Because of the much higher efficiency of these lamps the replacement of incandescent lamps for 2200 lumen by fluorescent lamps for 6000 lumen will in the end mean a much smaller increase in the cost. The practical and esthetic advantages of general illumination will then be appreciated more and also the advantage of fluorescent lighting with respect to colour rendering will become indispensable.

These considerations have led us to investigate more closely the practicability of a general lighting system with tubular fluorescent lamps, and here a brief account is given of the results obtained.

Some experiments with tubular fluorescent lamps for general interior lighting.

The name already expresses the fact that the tubular fluorescent lamp is not to be regarded as a point source of light like the incandescent lamp. Therefore the position in which the lamps are to be installed in a room has to be considered from a new angle which was unknown with separate incandescent lamps, namely that of mounting the lamps parallel to lines or planes in the room. For practical purposes this means that these lamps, being straight-lined, are especially suitable for mounting

against walls or ceiling, the very position most suitable for the above-described general illumination. Even when these tubular lamps are placed in fittings, which must be the case in living rooms in spite of their relatively low brightness, this still applies to a certain extent.

The next question is how the lamps required to give the total light flux should be distributed in the room. The 6000 lumen calculated above as the minimum requirement for a general illumination corresponds in the case of fluorescent lamps to a power of about 150 watts, which is just the figure taken above as basis for the present customary lighting with incandescent lamps. Since the fluorescent lamps now available are in units of approximately 40 watts net, three or perhaps four such units can be used to obtain the desired total light flux.

In order to give some idea of the distribution of the light when the necessary lamps are placed in different positions in the room, a number of measurements have been taken with an experimental installation in a test room with a floor area of 5.5×3.5 m^2 and height 2.75 m in which two fittings were installed as shown in *fig. 1*. One is a cornice fitting

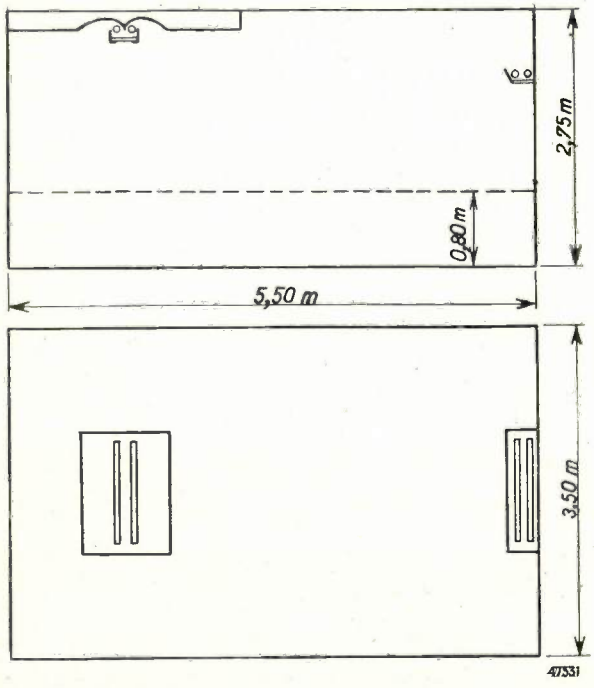


Fig. 1. Plan and elevation of the test room with the two fittings: one let into the ceiling (which was somewhat lower over one half of the room) and one in a cornice on the narrow wall of the room.

placed over a window in one of the shorter walls of the room and shut off from underneath by a plate of frosted glass; this cornice contains two tubular fluorescent lamps of 40 W each. The other fitting is set into the slightly lower ceiling of the other half of the room and consists of two mat-white curved surfaces reflecting into the room the light from two tubular lamps, likewise of 40 W

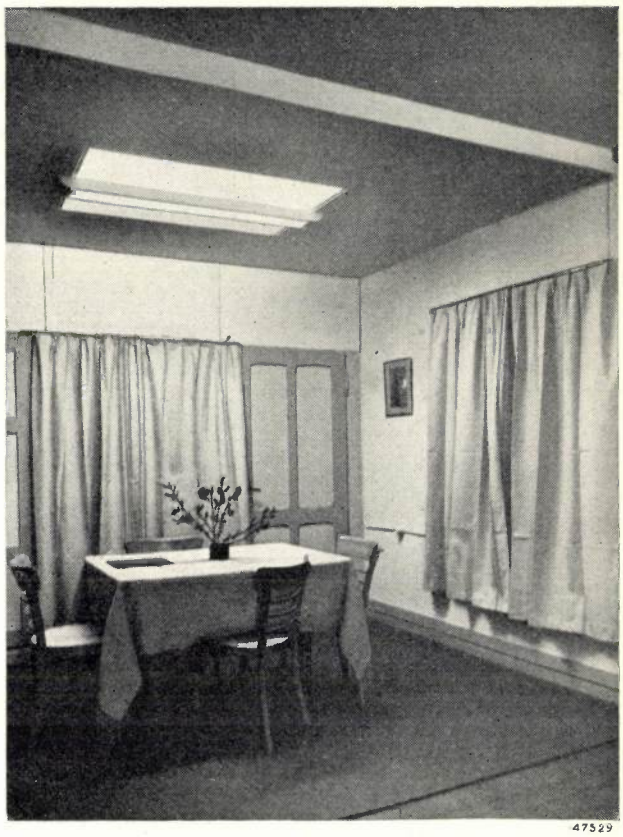


Fig. 2. Half of the test room with the fitting let into the ceiling.

each; this fitting, too, is shut off underneath by a glass plate which disperses the light. (In actual practice with such a fitting it may be found preferable to replace this horizontal plate by vertical diffusing partitions, so as to avoid loss of light owing to the collection of dust.)

Figs. 2 and *3*, which are photographs of the room with the two fittings, give an impression of the possibilities with such a lighting system as far as the general aspect of the room is concerned.

By switching on a different number of lamps in each fitting, different light distributions were obtained. The horizontal intensity of illumination at table height was measured at a large number of points and the results were plotted in diagrams of isolux lines. (These diagrams apply for lamps yielding exactly 200 Dlm; in practice lamps deviate somewhat from this figure.) In *figs. 4a-d* the floor area of the test room is represented in each case by a series of isolux lines for the cases 2+0, 2+1 and 2+2, respectively, the first number referring to the number of lamps burning in the ceiling fitting and

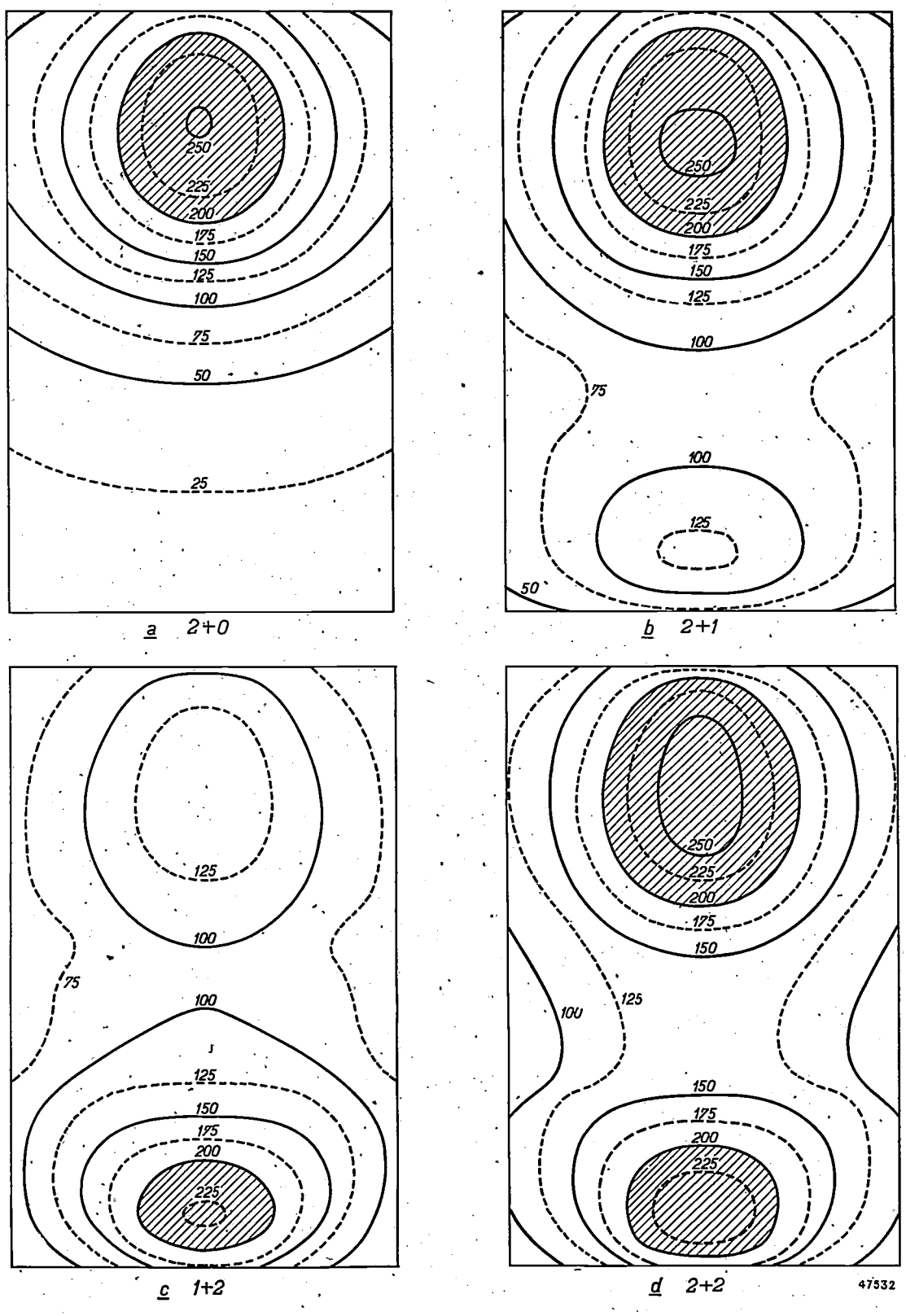


Fig. 4*a-d*. Different distributions of light in the test room obtained with different combinations of the lamps in the ceiling and cornice fittings. Each figure represents a plan of the room with isolux lines of the light distribution on the working surface. The numbers indicate the light intensity in lux. The areas which receive more than 200 lux are cross-hatched. *a*) Refers to the case 2+0 (2 lamps burning in the ceiling fitting, 0 in the cornice), the other figures, as indicated, referring to the cases 2+1, 1+2 and 2+2.

the second to the number burning in the cornice. In the first case (2+0) a large part of the room receives less than 25 lux. The second case (2+1) shows a quite useful distribution with nowhere less than 50 lux, more than 200 lux over an area of 2.5 m^2 and more than 100 lux over 11.3 m^2. In the third case (1+2) the general illumination of the room is still better, being scarcely anywhere below 75 lux, but on the other hand the area with the highest illumination (200 lux) is rather too small, only 0.9 m^2. Finally there is the case 2+2, the isolux diagram of which speaks for itself: there is nowhere less than 100 lux, whilst there are two areas of 3.0 m^2 and 1.1 m^2 with more than 200 lux. It is true that in the last case the total light flux is somewhat greater than was premised, namely 8000 lumen, but it goes to show how with such a relatively small additional power a rather good general illumination (according to our present conceptions) is attained, whereas if incandescent lamps were used this would only be possible by increasing the consumption to two or three times 150 W. Moreover, with the 2+1 illumination (fig. 4b), where the power consumption budget is certainly not exceeded, the result is satisfactory in most respects.

Fig. 3. Other half of the test room with cornice fitting over the window.

In conclusion we may say that, thanks to the high efficiency of the tubular fluorescent lamps, our present day lighting system for the living room can not only be improved but actually revised. Especially for the smaller homes, where more intensive use is made of the living room, the system here described, with its high level of general illumination and inconspicuous fittings, is most appealing. Lamp manufacturers, lighting architects and occupants will have to cooperate if the possibilities indicated here are to be brought to full development.

A VOLTAGE STABILIZING TUBE FOR VERY CONSTANT VOLTAGE

by T. JURRIAANSE.

621.316.722.1:621.384.5

Glow-discharge tubes can be so constructed that the working voltage is independent of the current within a fairly wide region of currents. Such tubes may therefore be used for the stabilization of voltages. However, even the best stabilizer tubes at present available have two drawbacks. In the first place they are not stable, the working voltage, *i.e.* the stabilization voltage, varying considerably for different specimens of the same type of tube, giving variations of 10 to 15 volts at a nominal voltage of, for instance, 100 volts. In the second place the working voltage varies with time: a variation of 10 to 15 volts during the life of the tube is quite common. Stabilizing tubes have now been developed by Philips which are practically free of these drawbacks, the variation of the working voltage for different specimens varying by not more than a few volts, while the variation with time is not more than $^1/_2$ volt per 1000 working hours. This great improvement has been attained by using a carefully prepared molybdenum cathode and depositing a thick layer of molybdenum on the walls of the tube by sputtering the cathode in a gas discharge. With the new stabilising tube ambient temperature has but little effect on the voltage. This temperature effect, which was completely overshadowed by the above-mentioned large voltage variations in other stabilising tubes, is discussed at the end of the article.

Introduction

In an electrical apparatus it is often necessary that the D.C. voltage should be independent of variations in the supply voltage. The grid voltage of an amplifier valve, for example, should be very constant, since any fluctuations may cause a variation of the amplification. It may also be desired to smooth the voltage ripple which is retained after rectification of A.C. voltage.

This stabilization of voltage can be achieved in a simple way by means of gas-discharge tubes [1]). The glow discharge is particularly well adapted for such purposes, as will appear from the following.

A glow discharge occurs between two electrodes in an atmosphere, for instance, of a rare gas at pressures of 0.5 to 40 mm of mercury when the cathode is cold and thus gives no thermionic emission of electrons. At voltages of 100 to 200 volts a luminous aura is seen around the cathode — the glow — which is separated from the cathode by a dark layer, Crookes dark space. If the current through the tube is small the surface of the cathode is not completely covered by the glow, the latter appearing as a sharply defined spot on the anode. When the current is increased the spot spreads, while both the current density, which is of the order of 1 mA/cm^2, and the voltage on the tube remain constant, until the glow has completely covered the surface of the cathode.

In this current region the so-called normal glow discharge takes place (*cf. fig. 1*). Once the whole cathode is covered by the glow, a further increase in current necessarily increases the glow-current density and the voltage over the tube also increases. This is then the region of the anomalous glow discharge.

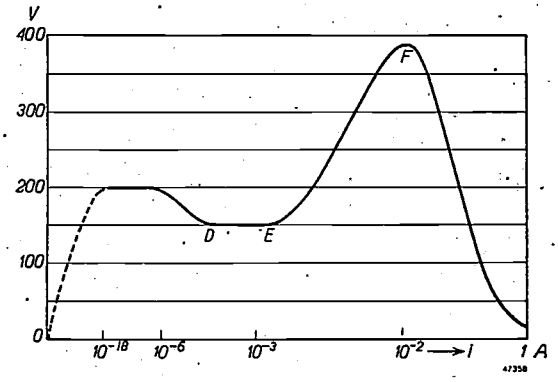


Fig. 1. The voltage V on a gas-discharge tube as a function of the current i (represented diagrammatically). Several values of the current are indicated, although not to scale, in order to show the order of magnitude of the current at which a given type of discharge occurs. To the left of E is the region in which breakdown occurs. The region between E and D corresponds to the "normal glow discharge". To the right of D the glow discharge is anomalous. To the right of F is the region of arc discharge.

In the region of the normal glow discharge, therefore, the voltage on the glow discharge tube is independent of the current. Upon this fact is based the possibility of using a glow-discharge tube for stabilization of voltage.

The change of potential between the electrodes of the tube is not linear. In *fig. 2* it may be seen that the voltage drop takes place mainly near the cathode; this is called the cathode drop. The distance covered by this potential difference corresponds quite well to the depth of the Crookes

[1]) Another method of stabilization is described in Philips Techn. Rev. 4, 54, 1941.

dark space. There is sometimes also a voltage drop close to the anode, which is called the anode drop and which is visible on the anode as a luminous film or globule. This voltage drop amounts to 10 to 20 volts, much less than the cathode drop, which

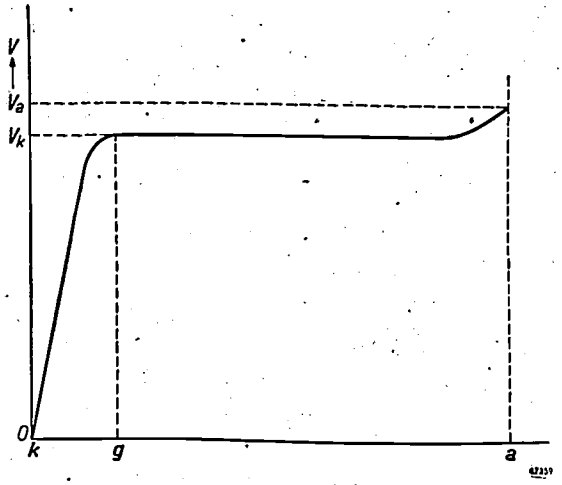


Fig. 2. The voltage distribution between the cathode and the anode of a glow discharge tube. V_k is the cathode drop; V_a is the voltage of the anode with respect to the cathode, *i.e.* the working voltage; $V_k - V_a$ is the anode drop. At k, a and g, respectively, are the cathode, the anode and the boundary between the Crookes dark space and the glow.

is 100 to 200 volts. Between these two regions of drop in potential the voltage is fairly constant. In the following we shall speak only of tubes with no anode drop, unless the contrary is especially stated, so that the working voltage measured is equal to the cathode drop. Thus for the tubes in question the cathode drop is equal to the working voltage. When the tube burns in the region of the normal glow discharge the cathode drop is also called normal.

The value V_n of the normal cathode drop is determined by the material of the cathode and the nature of the gas. *Table I* gives a survey of several values of that voltage as given in the literature for different combinations of gases and cathode materials.

Table I,

The normal cathode drop V_n (in volts) for different combinations of gases and cathode materials

	neon	argon	nitrogen
barium	—	95	155
graphite	200	—	—
iron	150	165	215
potassium	70	65	170
molybdenum	115	—	—
nickel	140	130	195

The glow-discharge tube as voltage-stabilizing tube

For purposes of stabilization the discharge tube is connected with the voltage source *via* a series resistance (see *fig. 3*). The load B, which is here shown as a resistance, is connected across the glow-discharge tube. The essential point is that a voltage variation of the source shall be taken up completely in the series resistance, since the variation of the current through the tube causes no change in the working voltage.

The load thus experiences no fluctuations in voltage.

The series resistance and the load will usually be so chosen that at the average value of the voltage from the source the glow-discharge tube carries a current at which the glow half covers the cathode, so that the variations of the current through the tube in both directions can be a maximum.

Now in practice the cathode surface will not be uniform physically. Very slight contaminations of the surface have a large influence on V_n, so that the values of table I give only a rough impression. A variation of 30 volts in the values given in the literature is no exception. The contaminations mentioned result in different parts of the cathode having different normal cathode drops. In that case current variations do indeed cause voltage changes. The

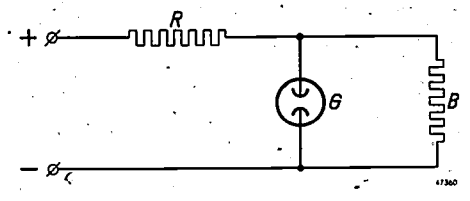


Fig. 3. The circuit diagram of a voltage stabilizing tube. G glow-discharge tube, R series resistance, B load.

differential resistance dV/di, also called internal resistance R_i of the tube, will then differ from zero, since the glow discharge will first pick out those parts of the cathode where V_n is lowest. Now the quality of the tube is higher the lower the internal resistance. It is thus clear that some care must be taken in preparing the surface of the anode.

The position of electrodes with respect to the wall and to each other can also affect V_n and thus also R_i. In particular V_n and R_i will increase when the glass wall is in the immediate vicinity of the glow-discharge, due to the fact that the positive and negative carriers or charge can reunite on the wall.

For many purposes satisfactory values of R_i

have been successfully obtained and a number of tube types for various current ranges and voltages have been manufactured by Philips for many years. The different voltages are obtained by using different combinations of gases and cathode materials (table I). Some types for 100 volts have an iron cathode covered with barium, in combination with a mixture of rare gases, usually containing neon with several tenths of a percent of argon. By the addition of this small quantity of argon the fairly high breakdown voltage of neon is reduced to a practical value approximating to the working voltage. Other tubes for 100 volts have an iron cathode covered with magnesium and a gas filling of a different neon-argon mixture.

A very low internal resistance is obtained in the case of the type 100 *E1* by not only preparing the cathode with great care, but also by choosing the most favourable geometrical arrangement of the electrodes. The tube has three concentric cylinders with several millimeters intermediate space, in which the discharge takes place. The inner and outer cylinders are the anode, the middle one the cathode, so that both sides of the surface can be covered by the glow-discharge up to a maximum current of 200 mA. The glow-discharge, no matter what its extent, thus always burns under geometrically almost equivalent conditions, while the above-mentioned recombination of the ions on the walls is out of the question.

The voltage stabilizing tube for very constant voltage

From the above it is clear that it has long been possible to construct glow-discharge tubes with low internal resistance. For many applications, however, there is the drawback that the working voltages of different tubes of the same type vary so much, and, moreover, that in the course of time the working voltage of a given tube may vary considerably. The figures given for the voltage tolerance of the usual stabilisation tubes are sufficiently proof of this. The deviations in the values of the working voltage and the variation during the lifetime are usually of the same order of magnitude and may amount from 10 to 15 volts.

We have now, however, succeeded in developing a tube which, in addition to the normal good smoothing with a low differential resistance, also has a working voltage which does not vary more than 0.5 volt in 1000 hours and the tolerance of which does not exceed 2 to 3 volts as between the different experimental specimens. From the following it will be seen that the essential factors responsible for this are the use of molybdenum as cathode material, the cleaning of the cathode surface by sputtering in the glow discharge and finally the deposition of a layer of atomized cathode material on the walls of the tube; this atomization is accomplished by continuing the sputtering for a long time after the cleaning of the cathode.

The phenomena observed when a glow discharge, for instance in pure neon at a pressure of 20 mm of mercury, is allowed to burn on a chemically well cleaned molybdenum surface are as follows [2]).

When the current is so large that the glow-discharge entirely covers the cathode, after burning for five or ten minutes the cathode glow will be seen to contract to a small part of the cathode surface, the working voltage meanwhile falling from about 300 to 150 volts. The glow-current density rises from about 0.5 mA/cm^2, corresponding approximately to the normal current density, to one hundred times that value. If the current through the tube is now increased slightly the intensity of the light in the cathode spot increases, while the glow at the same time begins slowly to spread out, the glow discharge, however, remaining entirely anomalous.

This continues until the whole cathode is again covered by the glow. If after some time the current through the tube is diminished until the discharge is normal, the low normal current density of about 0.5 mA/cm^2 is recovered, which initially prevailed at the high voltage of 200 volts but is now found at about 108 volts.

If the glow is now allowed to continue to burn at this low current, after several minutes the working voltage is observed to rise, at first only little, but ultimately by several tenths of a volt. This rise in the working voltage can also be brought about for instance by adding to the neon several hundredths of a percent of oxygen after the low working voltage has been reached. The conclusion is that the low working voltage corresponds to a molybdenum surface which has been cleaned by sputtering in the discharge with high current density, and that during the burning with a low current density the contaminations which have been liberated spoil the clean surface again.

This latter effect can now be avoided by continuing the sputtering after the cleaning of the cathode surface, until a good thick layer of atomized molybdenum is deposited on the walls of the tube. It therefore seemed obvious to assume that the

---

[2]) *Cf.* F. M. Penning and J. H. A. Moubis, Philips Res. Rep. 1, 119, 1946.

contaminations in question came out of the glass walls, and this was confirmed by further investigation. The sputtered molybdenum can act as a getter to take up these small quantities of gas.

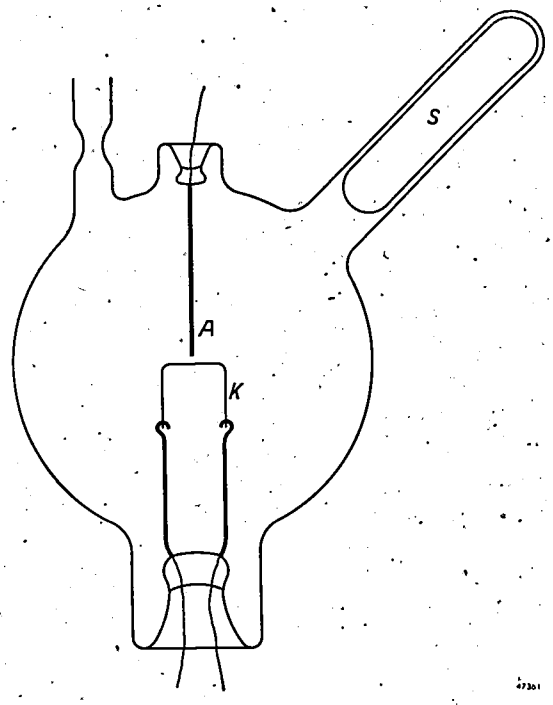


Fig. 4. Arrangement for demonstrating the liberation of contaminations from the wall by the glow discharge. *A* anode, *K* cathode, *S* moveable glass rod.

In addition, however, it is found that the discharge itself frees gases from the glass walls, so that the sputtered layer serves as a shield between the discharge and the glass wall. Only after the deposition of this layer could a voltage variation of less than 0.5 volt, often only 0.1 volt, per 1000 hours be obtained.

The following simple experiment clearly demonstrates that the discharge does actually free the contaminations from the glass. In a side tube of a discharge tube with a molybdenum cathode (*cf. fig. 4*) a glass rod *S* is placed in such a way that by shaking the tube it can be moved toward or away from the cathode. When the tube is evacuated and the cathode sputtered the rod lies at the back of the side tube, so that no sputtered molybdenum reaches it. When the cathode has been sputtered long enough for so much molybdenum to be deposited on the glass walls that the working voltage, reckoned over many days, remains constant, then upon the glass rod being brought within a distance of say 1 cm of the cathode, the working voltage will rise at a rate of about 5 mV per minute.

After some time the cathode even becomes so contaminated that the cathode spot begins to contract. When the rod is again shaken into the back of the side tube the variation of the working voltage ceases immediately and even reverses its direction, due to the fact that the cathode is restored under the influence of the glow-discharge. The gaseous contaminations are removed from the cathode and taken up in the sputtered layer of molybdenum on the walls. This porous layer of metal thus acts not only as a getter [3]) but also as a shield between the discharge and the walls.

Due to the measures described, the value of the normal cathode drop is now very satisfactorily reproducible and varies only a few volts for different tubes. In pure neon at a pressure of 40 mm the average value of V_n is about 106.5 volts; in 10 mm argon about 103.5 volts and in neon with $^1/_2$ percent argon 84.5 volts. *Fig. 5* shows a practical model of the tube for 8 mA and 85 volts stabilization voltage.

Many other metals can also be treated in this way and give reasonably reproducible values of V_n, but molybdenum and zirconium give the best results

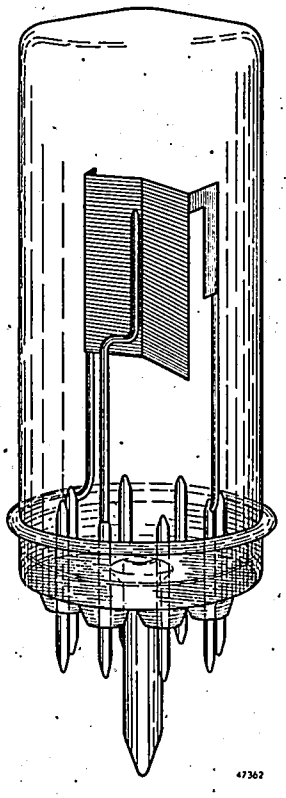


Fig. 5. Practical model of the new stabilizing tube for very constant voltage. The rod-shaped electrode is the anode, the plate is the molybdenum cathode. Actually the interior of the tube is scarcely visible through the black layer of sputtered molybdenum of the walls.

[3]) *Cf.* T. Jurriaanse, F. M. Penning and J. H. A. Moubis, Philips Res. Rep. **1**, 225, 1946.

and also offer the advantage in manufacture that, notwithstanding the high current density during the sputtering, the glow discharge contracts entirely on to these cathode materials. The current leads on which the molybdenum or zirconium are mounted are usually made of materials which have a much higher normal cathode drop (for nickel, for instance, $V_n \approx 140$ volts).

The temperature coefficient of voltage stabilizating tubes

Now that it is possible to make tubes with a voltage which is independent of the current and time, another variation of the working voltage becomes of importance, to which previously no attention had been paid. This is the variation of the working voltage according to the ambient temperature. With the various older types of stabilizing tubes this variation amounts to 20 to 30 mV per degree centrigrade, the value being a negative one in the tubes which have no anode drop. The new tube with molybdenum cathode for 8 mA and 85 V has a still smaller temperature coefficient, lying below —10 mV per degree. This slight dependence on temperature cannot be entirely avoided, as will appear from the following.

According to the generally accepted view, the value of the normal cathode drop is independent of the density of the gas, so that no effect from the ambient temperature should be expected. The actual situation, however, does not appear to be so simple.

In the first place, from measurements which were made possible by the fact that the value of V_n is very well reproducible when the cathodes are prepared according to the above-described method of procedure, it appeared that the normal cathode drop V_n is indeed dependent on the gas density, as *fig. 6* shows for the gases neon and argon.

In the second place, the gas density varies somewhat in the vicinity of the cathode when the temperature of the surroundings of the tube changes. The glow heats the cathode with a power approximately equal to the voltage consumed by the tube, since the whole voltage drop is taken up in a gas layer several tenths of a millimeter thick around the cathode. The cathode is thus at a higher temperature than the walls of the tube, so that the density of the gas at the cathode is less than at the walls. Approximately the densities referred to are inversely proportional to the absolute temperatures at those points.

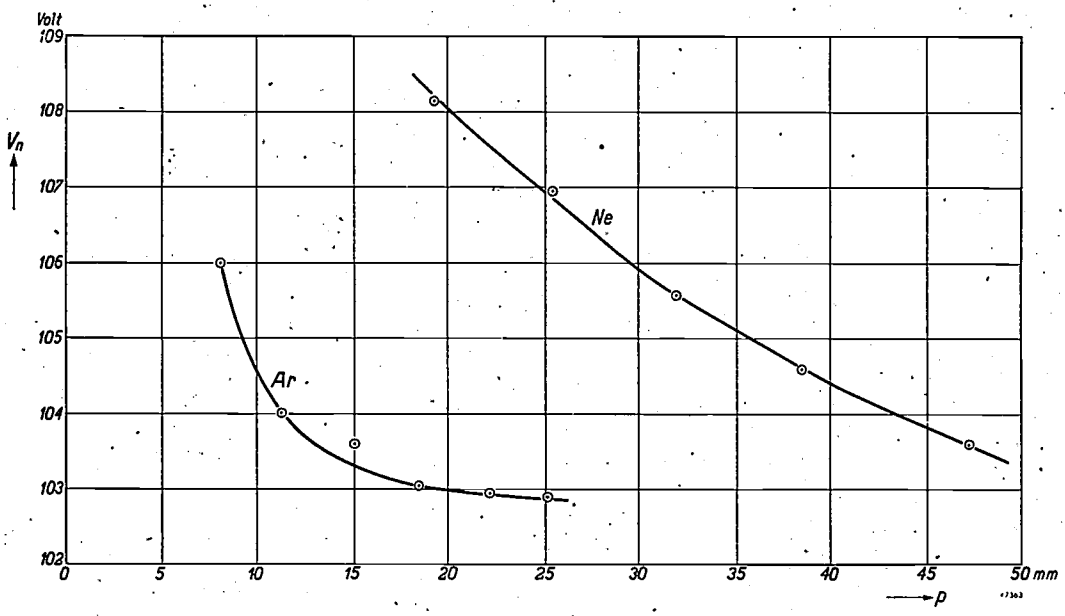


Fig. 6. The normal cathode drop V_n as a function of the gas pressure p for argon (Ar) and neon (Ne).

Now if at a given current through the tube, and thus with a given power consumed, the ambient temperature T_1 is increased by t degrees, in the first instance the temperature T_2 of the cathode will also rise by t degrees, at least if the coefficient of heat conductivity is independent of the temperature. Now, however,

$$\frac{T_2+t}{T_1+t} < \frac{T_2}{T_1},$$

in other words the ratio of the temperatures and thus

that of the densities more nearly approaches unity. Thus if the ambient temperature becomes higher the density near the cathode rises, and according to fig. 6 the cathode drop decreases. It is clear that a negative temperature coefficient of the working voltage is the result, as was indeed found experimentally. Actually, however, the coefficient of heat conductivity of the gas is by no means independent of the temperature, and, moreover, the transfer of heat by radiation also plays an important part.

Due to these causes, as T_1 increases so the quotient T_2/T_1 approaches unity even more closely than was the case before [4]).

The coefficient of heat conductivity is about proportional to $\sqrt{T}$, so that for the same transfer of heat at a higher temperature a smaller temperature gradient is required than at a lower temperature. And, finally, the difference in radiation between cathode and wall of the tube, which is necessary for a given transfer of heat and is approximately proportional to $T_2^4-T_1^4$, is reached with a smaller temperature difference according as T_1 is higher.

For certain applications which have become possible due to the new construction of stabilizing tube the unavoidable temperature effect here described will therefore have to be taken into account.

---

[4]) Dr. T. Jurriaanse, Philips Res. Rep., to appear shortly.

IMPEDANCE MEASUREMENTS WITH A NON-TUNED LECHER SYSTEM

by J. M. van HOFWEEGEN.

621.317.33.029.63

A method is described by which impedances can be measured whose resistance component is of such a magnitude that measurement by a previously described method is practically impossible. The characteristic feature of this method is the use of a Lecher system which is not tuned to the measuring frequency. By measuring the variation of voltage occurring along the Lecher system when it is loaded by the impedance to be measured, the so-called reflection factor can be determined. From this reflection factor the unknown impedance can be calculated, whereby a diagram can be used to advantage. The instrument used for measuring the voltage must be calibrated only relatively, just as in the measurements with a tuned Lecher system. A method of measurement is finally described in which the measuring instrument need not be calibrated at all.

In a previous article [1]) a method was described by which impedances can be measured at wave lengths shorter than 1 metre. There use is made of a Lecher system tuned to the measuring frequency. Attention was drawn to the fact that with the use of a short-circuited Lecher system of a quarter wave length this method is only suitable for measuring impedances the resistance component of which is at least several times larger than the characteristic resistance ζ. By using a short-circuited Lecher system of a half wave length impedances can be measured of which the resistance component is at least several times smaller than the characteristic resistance. Impedances having a resistance component of the same order of magnitude as the characteristic resistance (*i.e.* from 10 ζ to $\zeta/10$) cannot, therefore, be measured by the method described. Since the wave resistance of practically usable Lecher systems always lies within rather narrow limits (100 to 300 ohms), there is a rather large region of resistance, from about 30 to about 1000 ohms, which cannot be practically measured by the method in the article mentioned. It is, however, still possible to measure resistances of this order of magnitude with the help of a Lecher system but a different principle has to be applied. The characteristic feature of the method is that use is made of a Lecher system which is not tuned to the measuring frequency. By measuring the variation in voltage which occurs along such a Lecher system when the system is terminated by the impedance to be measured, the so-called reflection factor can be determined, from which the unknown impedance can be calculated. In the following we shall describe the way in which this principle is worked out.

Principle

In order to explain the method of measurement followed here, we will first deal briefly with the mathematical considerations. We begin with the familiar differential equations for voltage and current in the two conductors of a Lecher system [2]).

When a sinusoidal A.C. voltage E with the angular frequency ω (*fig. 1*) is applied to one of the extremities of a Lecher system we may write for the complex representations of voltage and current at an arbitrary point on the Lecher system $Ve^{j\omega t}$ and $Ie^{j\omega t}$, where V and I are complex quantities depending exclusively on the point on the Lecher system, and thus not on the time. The following differential equations now hold for V and I: [3])

$$\left.\begin{aligned}\frac{dV}{dy} &= I\,(r^I + j\omega L^I)\\ \frac{dI}{dy} &= V(g^I + j\omega C^I)\end{aligned}\right\} \quad \ldots \quad (1)$$

In these equations y is the distance from the point in question on the Lecher system to a certain point chosen as zero point (here the right-hand end,

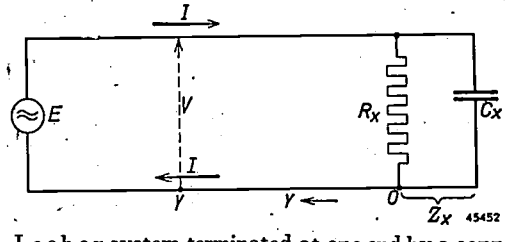


Fig. 1. Lecher system terminated at one end by a connection in parallel of a resistance R_x and a capacity C_x and at the other end connected with a source of high-frequency voltage E. At a distance y from the right hand end the voltage and current are, respectively, V and I.

see fig. 1); r^I en L^I are respectively the resistance and the self-induction per unit of length of the two

[1]) Philips Techn. Rev. **8**, 1946.

[2]) *Cf.* also: Philips Techn. Rev. **6**, 240, 1941.

[3]) These equations are not given in the more usual form, in which one of the members has the opposite sign, because here the coordinate y is reckoned from the end to which the source of voltage is *not* connected (see fig. 1).

conductors together, while g^I and C^I represent respectively the shunt conductance and the capacity between the two conductors, also per unit of length. The following equations for V and I can be derived from (1) in a simple way:

$$\left.\begin{aligned}\frac{d^2V}{dy^2} &= \gamma^2 V\\ \frac{d^2I}{dy^2} &= \gamma^2 I\end{aligned}\right\} \quad \ldots\ldots \quad (2)$$

γ being here the so-called propagation constant, which is given by the formula:

$$\gamma = \sqrt{(r^I + j\omega L^I)(g^I + j\omega C^I)} \quad \ldots \quad (3)$$

The solution of the first of equations (2) is as follows:

$$V = Ae^{+\gamma y} + Be^{-\gamma y} \quad \ldots\ldots \quad (4)$$

where e is the base of natural logarithms and A and B are integration constants, which therefore do not depend upon the position but are determined by current and voltage at one spot on the Lecher system. From (4) and the first of equations (1) it now follows for I that

$$I = \frac{A}{\zeta} e^{+\gamma y} - \frac{B}{\zeta} e^{-\gamma y} \quad \ldots\ldots \quad (5)$$

In this equation ζ is the so-called surge impedance or characteristic impedance, for which the following formula holds:

$$\zeta = \sqrt{\frac{r^I + j\omega L^I}{g^I + j\omega C^I}} \quad \ldots\ldots \quad (6)$$

In general γ and ζ are complex quantities.
We now set

$$\gamma = \alpha + j\beta \quad \ldots\ldots \quad (7)$$

If we first disregard the losses of the Lecher system ($r^I = 0$, $g^I = 0$), then according to (3) $\gamma = j\omega\sqrt{L^I C^I}$ and is thus purely imaginary, so that according to (7) $\alpha = 0$, $\beta = \omega\sqrt{L^I C^I}$.

The wave impedance is now real and ζ is equal to $\sqrt{L^I/C^I}$. One now speaks of the surge resistance. For equations (4) and (5) may now be written:

$$\left.\begin{aligned}V &= A\,e^{+j\beta y} + B\,e^{-j\beta y}\\ I &= \frac{A}{\zeta} e^{+j\beta y} - \frac{B}{\zeta} e^{-j\beta y},\end{aligned}\right\} \quad \ldots\ldots \quad (8)$$

If, as is represented in fig. 1, the right-hand end is terminated by an impedance Z_x, for $y = 0$ the ratio of V to I is equal to Z_x, and thus

$$Z_x = \frac{A + B}{A - B}\zeta, \quad \ldots\ldots \quad (9)$$

from which it follows that

$$B = A\frac{Z_x - \zeta}{Z_x + \zeta} \quad \ldots\ldots \quad (10)$$

so that for the first of equations (8) we may write

$$V = A\left\{e^{+j\beta y} + \frac{Z_x - \zeta}{Z_x + \zeta} e^{-j\beta y}\right\} \quad \ldots \quad (11)$$

The quantity

$$f = \frac{Z_x - \zeta}{Z_x + \zeta}, \quad (11a)$$

which is in general complex, is usually termed the reflection factor. We shall explain its significance in the following.

For the complex representation of the voltage we may now write:

$$Ve^{j\omega t} = A\left\{e^{j(\omega t + \beta y)} + fe^{j(\omega t - \beta y)}\right\} \quad \ldots \quad (12)$$

We now represent the argument of the complex quantity f by φ and the modulus by $|f|$, so that $f = |f|e^{j\varphi}$.

As is known, the momentary value of the voltage is equal to the real part of its complex form. From (12) it now follows for this momentary value that

$$V_{mom} = A\left\{\cos(\omega t + \beta y) + |f|\cos(\omega t - \beta y + \varphi)\right\} \quad (13)$$

The first term of the binomial between the brackets represents a wave travelling towards the right, while the second term represents one travelling towards the left. These two waves are usually called the incident (or direct) and the reflected (voltage) waves. The wave length of the two waves is determined by the equation

$$\beta\lambda = 2\pi \quad \ldots\ldots \quad (14)$$

The complex representation of incident and reflected waves, is, according to (12), given by $Ae^{j(\omega t + \beta y)}$ and $Afe^{j(\omega t - \beta y)}$, respectively. For $y = 0$, *i.e.* for the reflecting end where Z_x is connected, these expressions become

$$Ae^{j\omega t} \text{ and } Afe^{j\omega t}, \text{ respectively.}$$

This shows us the significance of f: it is the ratio of the complex representations of reflected and incident waves at the reflecting end.

We shall now consider the incident and reflected waves in more detail for three simple cases (*cf.* expression (11*a*) for f).

a) When the impedance Z_x is equal to ζ, f becomes equal to zero, *i.e.* there is no reflected wave. The incident wave which is then present alone is called a pure travelling wave.

b) If the reflecting end is open $f = 1$, *i.e.* at that

end the incident and the reflected voltage waves are equal in intensity and phase.

c) If the reflected end is short-circuited $f = -1$, *i.e.* at that end the incident and the reflected voltage waves are equal in intensity and opposite in phase.

In cases b) and c), due to the superposition of the two waves, at certain spots lying at intervals of a quarter wave length alternate voltage minima and maxima (nodes and antinodes) occur. We then speak of standing waves. The equation for this is found by calculating the effective value of the A.C. voltage given by (13). After some reduction we find

$$V_{eff} = \frac{1}{\sqrt{2}} A \sqrt{1+|f|^2 + 2|f| \cos (2\beta y - \varphi)} \quad . \quad (15)$$

A voltage minimum occurs for those values of γ for which $\cos (2\beta y - \varphi) = -1$ and at that minimum according to (15) the voltage is

$$V_{eff\ min} = \frac{1}{\sqrt{2}} A \ (1-|f|) \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

In the above-mentioned cases b) and c) where $|f| = 1$, therefore the voltage at the minima is always zero. We might then speak of pure stationary waves. Such waves also occur when a loss-free Lecher system is terminated by a loss-free reactance. In that case Z_x is of course purely imaginary. From (11a) it then follows, since ζ is real, that $|f| = 1$. If, however, the load impedance Z_x contains a resistance component, $|f| \neq 1$ and thus the voltage is not zero at the minima.

The theory given in the foregoing holds for every form of transmission connection and thus not only for a Lecher system, which is actually nothing but a short section of transmission line used at a very high frequency.

Determination of the reflection factor

According to (15) the positions of the voltage maxima and minima, as well as the further curve of the voltage as a function of y, depend upon $|f|$ and φ, thus on the impedance connected at the end. By measuring this voltage along the Lecher system we can determine $|f|$ and φ, and from them Z_x. The measurement of the variation in voltage can be carried out by means of a measuring instrument (diode voltmeter) which can be slid along the two conductors (*fig. 2a*). We then find a variation in the voltage like that reproduced for instance in fig. 2*b*.

The argument φ of f can be determined from the distance from the end of the Lecher system to the first voltage minimum. The formula for that is

$$\varphi = \frac{4\pi y_0}{\lambda} - \pi \quad . \quad . \quad . \quad . \quad . \quad (17)$$

where y_0 is the distance mentioned (see fig. 2*b*).

The modulus of f can be determined in two different ways. In the first method the variation of the high-frequency voltage is measured in the neighbourhood of a minimum. We then measure to

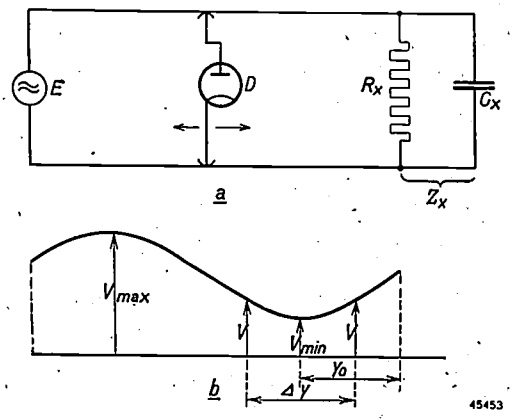


Fig. 2*a*) Diagram showing the principle of the method by which impedances can be measured with the help of a non-tuned Lecher system. Source of voltage E, impedance to be measured Z_x. The shape of the high-frequency voltage curve along the Lecher system is measured with the sliding diode D.
b) Curve of the high-frequency voltage V along the Lecher system.

either side of the minimum the distance at which the voltage is a certain number of times larger than the voltage at the minimum. When V is the voltage at two points which are symmetrical with respect to a minimum and lie at a distance Δy from each other (see fig. 2*b*), the relation between V and V_{min} is determined by:

$$\frac{V}{V_{min}} = \frac{\sqrt{1 + |f|^2 - 2|f| \cos (2\pi \frac{\Delta y}{\lambda})}}{1 - |f|} \quad . \quad . \quad (18)$$

Thus if Δy is measured for a certain value of V/V_{min}, for instance for $V/V_{min} = \sqrt{2}$, $|f|$ can be calculated according to (18).

If Δy is small compared with λ, formula (18) can be further simplified by expanding $\cos (2\pi \Delta y/\lambda)$ into a series. For $V/V_{min} = \sqrt{2}$ we then have

$$2\pi \frac{\Delta y}{\lambda} = \frac{1 - |f|}{\sqrt{|f|}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (19)$$

In the second method of determining the modulus of f the ratio between the voltage at a maximum, V_{max}, and the voltage at a minimum V_{min}, is measured. Since the distance between two voltage maxima is equal to a half wave length, in formula

(18) Δy is then equal to $1/2\lambda$, so that the formula becomes

$$\frac{V_{max}}{V_{min}} = \frac{1+|f|}{1-|f|} \quad \ldots\ldots \quad (20)$$

In the discussion of the causes of errors of measurement we shall also discuss the question of the advantages and disadvantages of these two methods of determining the modulus of f.

Losses of the Lecher system

In the foregoing it was assumed that the Lecher system was free of loss ($\alpha = 0$). If we had not made that assumption the equation for the variation of the voltage along the Lecher system would have been the following:

$$V_{eff} = \frac{1}{\sqrt{2}} A \sqrt{e^{2\alpha y} + |f|^2 e^{-2\alpha y} + 2|f| \cos(2\beta y - \varphi)}. \quad (21)$$

Now it is easy to understand that for a normal Lecher system $e^{2\alpha y}$ will deviate very little from unity. Due to the loss-free dielectric (air) $g^I = 0$, while owing to the very high frequency ωL^I is very large compared with r^I. From (3) and (7) it now follows that $\alpha \ll \beta$, or, in connection with (14): $\alpha\lambda \ll 2\pi$. Since a Lecher system constructed for the measurements in question will never be much longer than one wave length (on this length there are already two maxima and two minima), it will also be true for αy that $\alpha y \ll 2\pi$. Closer investigation of the value of α shows that this quantity is even so small that there is no objection to setting $2\alpha y \ll 1$, and thus $e^{2\alpha y}$ will be only slightly larger than 1.

Upon working out equation (21) further it is now found that the position of the voltage minima, as well as the variation of the voltage in the vicinity of these minima, is in the first approximation not affected by α, so that sufficiently accurate results for practical purposes are obtained when the losses of the Lecher system are disregarded.

Practical arrangement

The voltage variation along the Lecher system is measured with a diode voltmeter. The diode is placed in a holder which can be slid along the two conductors. It may often be of advantage to construct each of the two conductors of the system in two parts separated from each other by a thin insulating layer (see *fig. 3*). The capacity between these two parts of conductor is so large that with respect to high-frequency voltage they behave as one conductor. In this way the heating current can be supplied to the diode without wires being necessary to the diode. The circuit used for that purpose is represented in fig. 3. Between part b of conductor I and part a of conductor II a D.C. voltage occurs which is practically equal to the amplitude of the high-frequency voltage at the

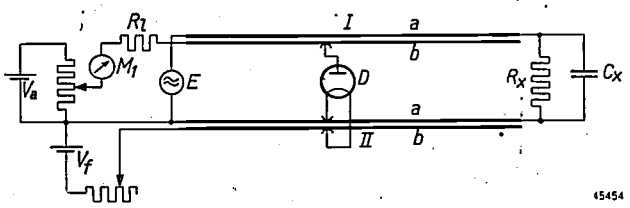


Fig. 3. The two conductors I and II of the Lecher system each consist of two mutually insulated parts a and b. E is the source of high-frequency voltage, V_f is the battery which furnishes the heating current of the diode, while V_a represents a battery by means of which the anode of the diode can be given a small negative bias. M_1 is a micro-ammeter in series with a leak resistance R_1.

point at which the diode is situated. The D.C. voltage can be measured at the beginning of the Lecher system by means of the galvanometer M_1, in series with which the resistance R_1 is connected. Thus no wires to the diode are needed for this either.

When the impedance to be measured is conductive for direct current [4]) it is preferably connected with part a of conductor I; this prevents interference of the measurement of the above-mentioned D.C. voltage by the impedance to be determined.

Errors of measurement

A disturbing element which often causes errors in measurement is the capacity of the measuring diode and of the diode holder. We desire, of course, to know the variation of the voltage along the Lecher system when it is completely "free" between the source of voltage and the impedance to be measured. It is only in that case that the above formulae are valid. Although in the development of the diodes used for these measurements and in the construction of the diode holder every effort was made to keep the capacity as small as possible, it cannot be reduced to such an extent that the Lecher system may be considered as being without load at the position of the diode. The result is that the high frequency voltage on the system is also affected by the location of the diode. It can be shown by calculation that the voltage which is measured in this case upon sliding the diode no longer varies symme-

[4]) The fact that the impedance to be measured can be represented by the connection in parallel of R_x and C_x does not mean that the object to be measured is conductive for direct current. R_x and C_x form only a diagram which has the same impedance as the required object at the frequency in question. Thus for example: the input impedance of a radio valve can be represented by a connection in parallel of a resistance and a capacity, while, nevertheless, practically no direct current conductivity exists between grid and cathode.

trically with respect to the maxima and minima as indicated in fig. 2*b*, but that it exhibits an unsymmetrical character, as shown for instance in *fig. 4*. The voltage read off at one side of a voltage minimum is too high and the other side too low. When the

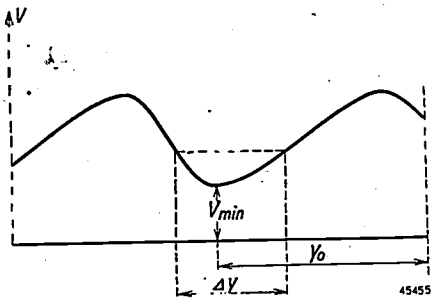


Fig. 4. Under the influence of the capacity of diode and diode holder the curve of the voltage becomes unsymmetrical with respect to the maxima and minima.

dissymmetry in the voltage curve measured is not too great and, moreover, Δy is small compared with the wave length, the deviations at either side of a voltage minimum approximately cancel each other and good measuring results are still obtained. When, however, Z_x differs only slightly from the wave resistance, the condition $\Delta y \ll \lambda$ is not satisfied for $V/V_{\min} = \sqrt{2}$, and incorrect results may be obtained. It is then necessary to eliminate the disturbing influence of the capacity of the diode and its holder. This can be done by introducing in parallel with the diode D a self-induction L_d (see *fig. 5*) which is made so large as to be in resonance

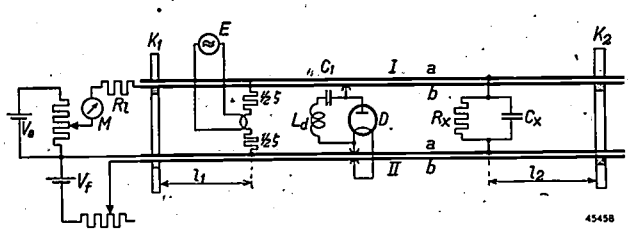


Fig. 5. The capacity of diode D and diode holder is tuned with the help of the self-induction L_d. C_1 is a blocking condenser for DC voltage. The source of high-frequency voltage E is coupled inductively with a loop connected with the Lecher system by means of two resistances which together are equal to the characteristic resistance ζ. K_1 and K_2 are short-circuiting bridges which short-circuit parts a of conductors I and II. The distance l_1 is so adjusted that the capacity of the resistances $^1/_2\zeta$ and of any holder for these resistances is compensated. By adjusting K_2 at the correct distance l_2 it is possible to compensate for the capacity of a holder in which the impedance to be measured is contained.

with the capacity of diode and holder, which can be ascertained from the fact that the voltage curve measured again becomes symmetrical with respect to the maxima and minima. The parallel circuit thus formed then represents a high impedance for the measuring frequency and has practically no effect on the shape of the voltage curve along the Lecher system.

Another possibility of reducing the effect of the diode capacity is to couple the diode with the system *via* very small capacities. One objection to this, however, is that a much higher high-frequency voltage is then necessary on the Lecher system. For many measurements this is of no importance, but for others there are very valid objections. Thus in measuring the input impedance of receiving valves the aim is to carry out the measurements as far as possible under the conditions under which the valves are used in practice, thus, among other conditions, with a low signal voltage. It may therefore be undesirable in this case to couple the measuring diode very loosely with the Lecher system.

In many cases the disturbing influence of the capacity of diode and holder can be reduced in yet another way, namely by constructing the source of voltage connected with the Lecher system in such a manner that it exhibits an internal resistance equal to the characteristic resistance. The resulting improvement can be explained from the fact that a voltage wave reflected at the diode cannot be reflected anew at the end where the source of voltage is situated. Whether or not this exerts a favourable effect on the above-mentioned measuring errors depends, however, upon various circumstances, for instance on the total length of the Lecher system. This should be investigated for each case individually.

Fig. 5 shows how in practice a voltage source with the desired internal resistance is obtained. The high-frequency voltage source E is inductively coupled with a loop, in series with which are two resistances ζ. These two resistances are connected with the Lecher system. As may be seen in fig. 5, the Lecher system is extended a certain distance to the left of this point of connection. On this part of the system we have the sliding short-circuiting bridge K_1 which short-circuits the parts a of the two conductors. This is done for two reasons. In the first place by setting K_1 at a certain distance l_1 the capacity of the above-mentioned resistances as well as that of a holder in which they are usually fastened can be compensated. Since the wave resistance of a Lecher system is practically a pure resistance, the internal impedance of transmitter, coupling loop and the above-mentioned resistances of $1/2\zeta$ should also be a pure resistance. The second purpose of the short-circuiting bridge K_1 is that it is possible to manipulate with the D.C. voltage connections to the left of it without affecting the distribution of high-frequency voltage on the Lecher

system. The system is also extended a certain distance to the right of the impedance to be measured. There is also a sliding short-circuiting bridge K_2 situated on this part of the system, which likewise short-circuits parts *a*. By setting K_2 at a certain l_2 it is possible to compensate for any capacity that may be in parallel with the impedance to be measured, for instance the capacity of a holder in which the object to be investigated is fastened and the capacity of an insulator support at this point.

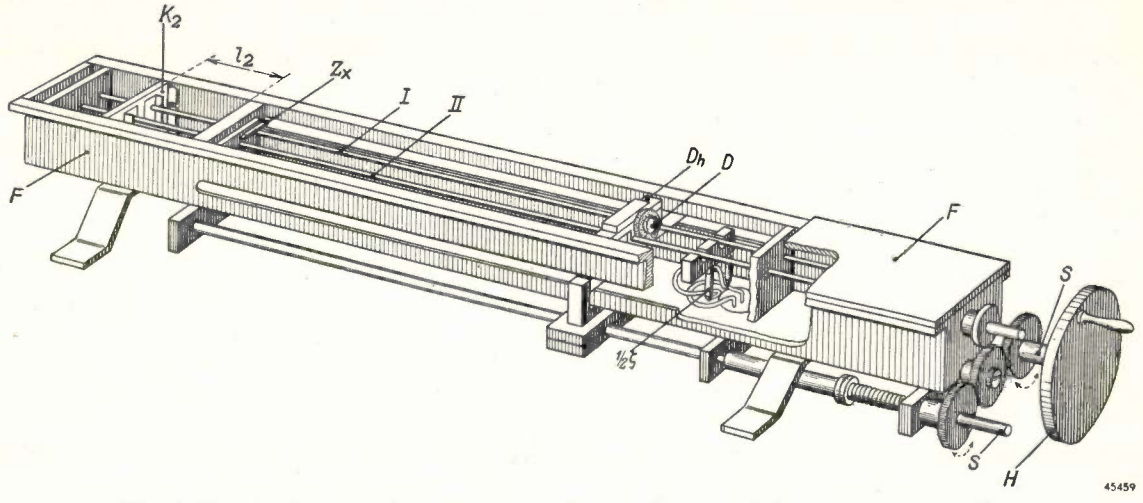


Fig. 6. Sketch of a measuring arrangement for impedance on decimetre waves according to the principle described in this article. In the drawing the cover and part of the wall of the housing *F* have been removed. The diode holder *Dh* is moved by means of the screw rods *S* which are driven by the hand wheel *H*. For the meaning of the other symbols refer to figs. 3 and 5.

The disturbing effect experienced from the capacity of the diode and its holder is greater in the neighbourhood of a voltage maximum than in that of a voltage minimum. This is due to the fact that the ratio between the voltage V and the current I is larger at a voltage maximum than at a voltage minimum. For that reason the first of the above-described methods of determining the modulus of f is carried out by measuring the voltage in the neighbourhood of a minimum, although in principle it would also be possible to measure the voltage in the neighbourhood of a maximum. For the same reason the second method of determining $|f|$ is only to be recommended when the shape of the voltage curve is such that there is little difference between a maximum and a minimum, *i.e.* when the impedance to be measured deviates only little from the characteristic resistance [5]). Of course the objection mentioned against measuring the voltage at other places than at a minimum is less important the more precisely the above-mentioned precautions have been taken for reducing the disturbing effects of the capacity of diode and holder.

Another source of errors in measurement which we should like to mention is a possible mutual induction between the conductors of the Lecher system and the connections of those conductors with the measuring diode. It has been found that large errors may also result from this cause, especially in determining the position of a voltage minimum, thus in the determination of the argument of the reflection factor f. These connections should therefore be kept as short as possible and should as far as possible run perpendicular to the two conductors of the Lecher system.

Fig. 6 is a sketch of a Lecher system used for measurements on the above-described principle. The Lecher system is contained in a housing *F*, the cover of which is partly removed in the drawing. The diode holder *Dh*, which is made of polystyrene, is moved along the conductors by means of screw rods *S* mounted outside the housing.

Diagram for calculation

When the modulus and the argument of the reflection factor f have been determined by the method described above, the impedance Z_x has to be calculated from those values by means of the expression (11*a*). This calculation proves to be rather laborious. It can be considerably shortened by the use of a diagram, an example of which is given in *fig. 7*.

The construction of this figure is based upon the

[5]) If the impedance to be measured is equal to the characteristic resistance, no standing waves occur along the Lecher system, but only travelling waves. The same voltage is then measured at all points.

fact that Z_x is again determined by the connection in parallel of a resistance R_x and a loss-free capacity C_x or self-induction L_x. The real part of f is now plotted as abscissa and the imaginary part as ordinate, while the geometrical positions of f are drawn for certain fixed values of R_x/ζ and of $1/\omega C_x\zeta$ or $\omega L_x/\zeta$, respectively. These geometrical positions are found to be circles. Since the modulus of f is always smaller than 1, the whole diagram falls inside a circle of radius 1. Thus by plotting the measured modulus and the measured argument of f in this diagram the corresponding ratio of the resistance part and the reactive part of Z_x to the wave resistance ζ can be read off. By the use of these quantities without dimensions the diagram can be employed for a Lecher system with any characteristic resistance.

It is clear from fig. 7 that the method described is especially suitable for the measurement of impedances having a resistance component R_x lying between about 10ζ and $\zeta/10$.

For $R_x/\zeta > 10$ and $R_x/\zeta < 1/10$, the circles for successive values of R_x/ζ lie very close together, *i.e.* a small variation in f corresponds to a large change in R_x. The determination of R_x from f outside the limits mentioned above is therefore inaccurate. The method of measurement described in this article is thus a satisfactory complement to the article cited [1]), since the latter possesses a higher precision just outside the limits mentioned.

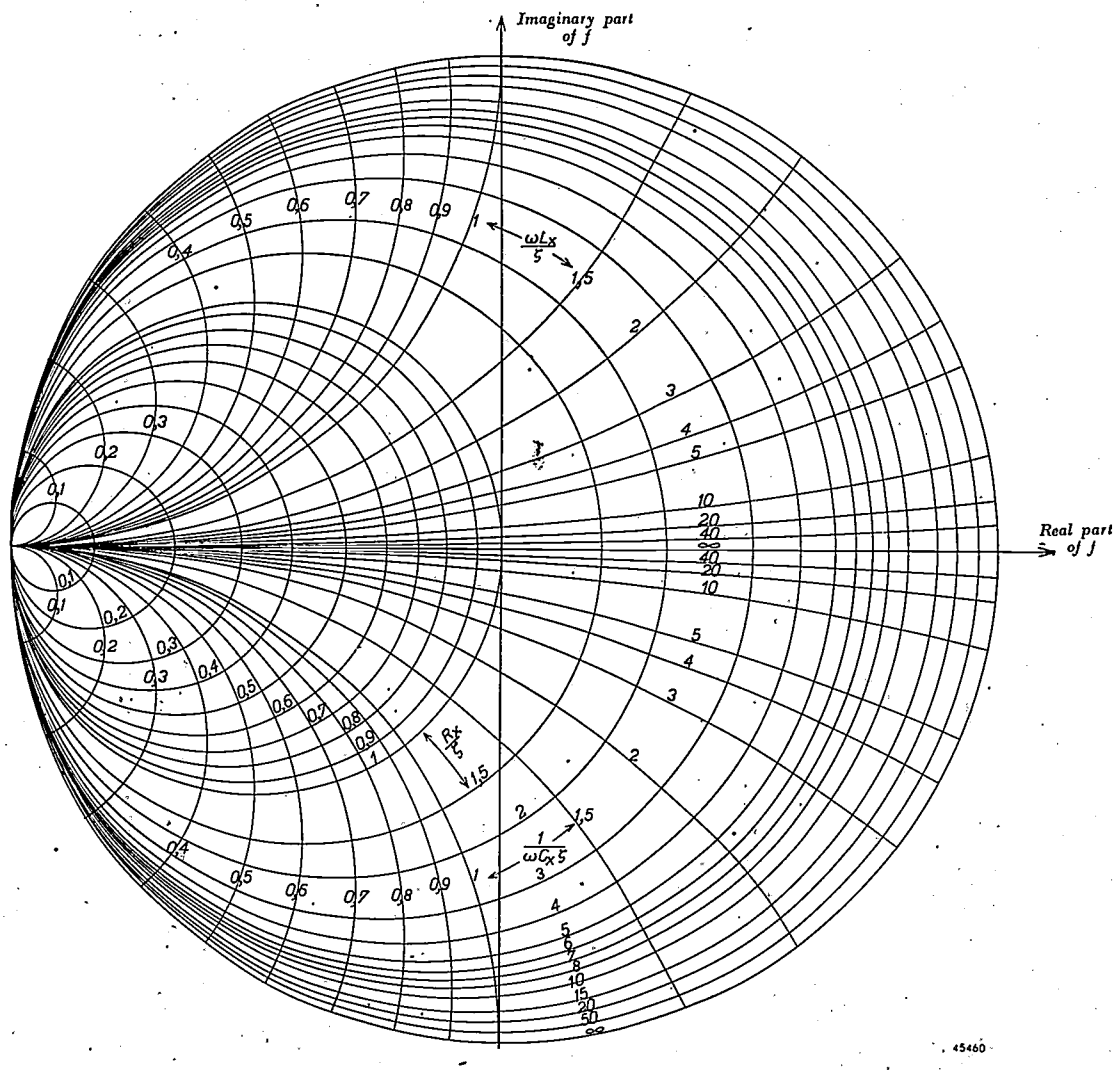


Fig. 7. Diagram for the determination of R_x and C_x (or L_x) from a measured value of the reflection factor f. The whole diagram falls inside a circle of radius 1.

Absolute values of the desired impedances

In the previous method as well as in the one described above the impedances to be measured are not found in absolute values, but only their ratios to the characteristic resistance. In order to determine the absolute value we must know the characteristic resistance. For simple constructions it can easily

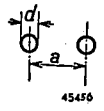
be calculated. The necessary formulae are given below for two commonly used constructions.

For two parallel round wires of diameter d and at a distance a apart the characteristic resistance is

$$\zeta = 120 \ln (p + \sqrt{p^2-1}) \text{ ohm}, \quad . \quad (22)$$

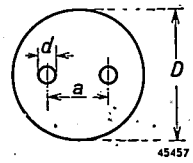
in which

$$p = \frac{a}{d}.$$



If a cylindrical shield of diameter D is placed around these wires the characteristic resistance can also be calculated according to formula (23), but for p the following must be introduced:

$$p = \frac{a}{d} \frac{D^2-(a^2-d^2)}{D^2+(a^2-d^2)}.$$



A completely closed cylindrical shield cannot, however, usually be employed. For structural reasons a square cross-section is often used, while at the same time slits are often necessary in the shielding, for instance for the purpose of moving the diode holder (see fig. 6). Since a precise calculation of the characteristic resistance in such cases is complicated and often impossible, it is usually simpler to measure it. For that purpose it is sufficient to measure the capacity between the two conductors, which may be done at a low frequency [6]). If this capacity per cm is C^I farad and the self-induction of the Lecher system per cm is L^I henry the following relation exists between L^I and C^I:

$$L^I \cdot C^I = 1/c^2, \quad . \quad . \quad . \quad . \quad . \quad . \quad (23)$$

where c is the velocity of light in cm/sec, *i.e.* $c = 3 \times 10^{10}$. This formula holds for shielded as well as for non-shielded Lecher systems.

If C^I is now measured, the characteristic resistance ζ can be calculated from it, since it is given by:

$$\zeta = \sqrt{\frac{L^I}{C^I}} = \frac{1}{cC^I} \text{ ohm} \quad . \quad . \quad . \quad (24)$$

Measurement of impedance without a calibrated instrument

As will have appeared from the above description, in this method of measurement also a diode voltmeter is needed which is only relatively calibrated. Such a calibration is usually performed at a low frequency, namely by comparison with a thermocouple which in turn has been calibrated with direct current. As already stated in the article referred to [1]), at very high frequencies a calibration performed at a low frequency is, relatively, no longer correct either, due to the fact that the electrons in a diode possess a finite transit time. We shall therefore in conclusion give a brief description of a method by which it is possible to measure impedances on decimetre waves without a calibrated measuring instrument being necessary.

This method shows much similarity to the method described in this article. Here also the impedance to be measured is connected to the end of a Lecher system I, while the high frequency voltage source is connected with the other end. In addition to the measuring diode, however, a second Lecher system II can now also be moved along system I (see *fig. 8*).

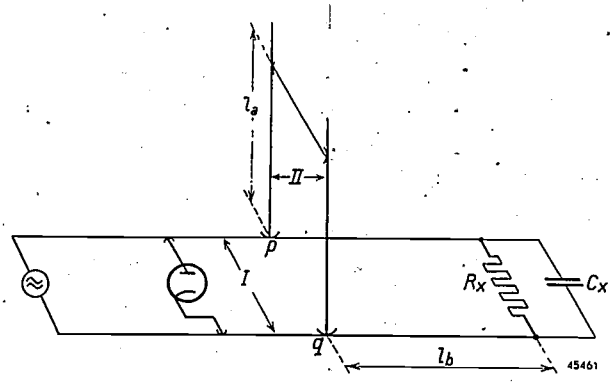


Fig. 8. Diagram showing the principle of an arrangement for measuring impedances on decimetre waves for which a calibrated voltage meter is not necessary. Lecher system *II*, which is provided with a moveable short-circuiting bridge, can be moved along Lecher system *I*. From the lengths l_a and l_b it is possible to calculate R_x and C_x.

System II is provided with a moveable short-circuiting bridge. It can be shown by calculation that at every value of R_x and C_x it is possible to adjust the length l_b so that the part of I lying to the right of II shows an admittance between p and q, the resistance component of which is equal to $1/\zeta$. The reactive component of this admittance can now be compensated by a suitable value of the length l_a. In this case, therefore, the part of I lying to the left of II is terminated by its own characteristic resistances and no standing waves occur along it. Upon moving the diode along the left-hand part of I no variation in voltage will then be observed. Z_x can now be calculated from the lengths l_a and l_b. We shall only give here the formulae to be employed for the case where the characteristic

[6]) When each of the conductors consists of a part a and b (figs. 3 and 5), those parts should of course be short-circuited.

resistance of systems I and II are the same. For modulus and argument of f the following are then valid:

$$|f| = \frac{1}{\sqrt{1 + 4\,\mathrm{tg}^2 2\pi \frac{l_a}{\lambda}}} \quad \ldots\ldots \quad (25)$$

and

$$\varphi = \mathrm{bg}\cos|f| - 4\pi\, l_b/\lambda \quad \ldots\ldots \quad (26)$$

From $|f|$ and φ we can now again calculate R_x and C_x (or L_x), using by preference a diagram like that of fig. 7.

Since in this method the diode voltmeter is only used to ascertain whether or not stationary waves are present on the Lecher system, it need not be calibrated at all. On the other hand it should be possible for the mechanism for adjusting the lengths l_a and l_b to be set with great precision.

The last issue of *Philips Research Reports* (No. 4 of volume 1, August 1946) contains the following papers:

R18: H. A. Klasens: Measurement and calculation of unsharpness combinations in X-ray photography.

R19: C. J. Bouwkamp: A contribution to the theory of acoustic radiation.

R20: M. Gevers: The relation between the power factor and the temperature coefficient of the dielectric constant of solid dielectrics II, III.

R21: J. Th. G. Overbeek: On Smoluchowski's equation for the electrophoresis of colloidal particles.

Readers interested in one of the above mentioned articles may apply to the Administration of the Philips Physical Laboratory, Kastanjelaan Eindhoven, where a limited numbers of copies are available for distribution. For a subscription to Philips Research Reports please write to the publishers of Philips Technical Review.

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

Reprints of the majority of these papers can be obtained on application to the Administration of the Research Laboratory, Kastanjelaan, Eindhoven, Netherlands. Those papers of which no reprints are available in sufficient number, are marked with an asterisk.

1683: H. Rinia and P. M. van Alphen: A new method of producing aspherical optical surfaces (Proc. Kon. Ned. Akad. Wetenschappen, Amsterdam 49, 146-149, 1946).

Descartes and Newton already advocated the use of aspherical surfaces in optical instruments. Recently interest in these has increased in connection with Schmidt's mirror system. Apart from field curvature this system has only one aberration, *viz.* spherical aberration, which can be eliminated by using a correction plate with a 4th degree surface. The manufacturing of these surfaces in mass production by grinding or pressing encounters many difficulties. In this article a method is described, by which the plates are cast from a gelatin solution. Some experiments with lenses provided with this aspherical gelatin surface likewise yielded good results.

1684: H. Rinia: New possibilities for the air engine (Proc. Kon. Akad. Wetenschappen, Amsterdam 49, 150-155, 1946).

After surveying the different types of caloric engines the author describes the principle of the air engine. A favourable temperature ratio, a satisfactory heat transfer and adequate regeneration are the requirements to be fulfilled. Thanks to improvements it has become possible to construct engines working at 3000 rpm at a mean pressure of 10 atm. with an overall output exceeding that of petrol engines.

1685: P. J. Bouma: Zur Einteilung des Ostwald'schen Farbtonkreises (Experientia 2, 99-103, 1946).

The author demonstrates that the „principle of inner symmetry", used by Ostwald as a basis for his colour circle, contains a number of contradictory requirements. After restriction of the principle it proves possible to compute the dominant wavelengths of the 24 colour points of the Ostwald circle situated on the curve of the characteristic colours in the colour triangle. The results of these calculations are compared to the data of Richter.

1686: J. van Slooten: Meetkundige beschouwingen in verband met de theorie der electrische vierpolen (thesis Delft 1946) (Geometrical considerations in connection with the theory of electric four terminal networks).

The theory treated in this thesis consists of two parts: the first two chapters deal with the properties of fourpoles as impedance transformers. In the last three chapters the question is raised how, when connecting in cascade (or in parallel) two fourpoles without losses, whose transformer properties are known, the properties of the resulting fourpoles may be found. It is proved that connection in cascade is equivalent to the addition of motions in non-euclidian space. This may be done in a so-called Cayley diagram worked out for this special purpose. The constructions are useful in the technique of ultra short radio waves.

1687: F. L. H. M. Stumpers: Eenige onderzoekingen over trillingen met frequentie-modulatie (thesis Delft 1946) (Some investigations an oscillations with frequency modulation).

In this thesis attention is paid to the definitions of instantaneous amplitude, phase and frequency as a function of time; the frequency spectrum occurring with different kinds of modulation is calculated and the possibility of interference with frequency modulated signals (*e.g.* by two emitters on one carrier wave, synchronised emitters or two-way reception) is studied.

The influence of noise and disturbances is thoroughly investigated. Whereas only approximative calculations are possible in the case where the noise energy is not small in comparison with the signal energy, in the case of impulse disturbances the calculations can be worked out completely.

The distortion suffered by frequency modulated signals in passing through the electrical network is calculated using Fourier analysis and the series of Carson and Fry. The methods used by these last named authors are critically studied and an alternative development is given, which is more adapted to the case of F.M., and which also has an asymptotic charecter. The theory is applied to simple networks: simple circuit, coupled circuit, frequency detector. Apart from the formation of harmonics intermodulation is considered.

Finally the results are checked by experiments. A new method for determining the distortion of the measuring emitter directly from the spectrum deserves attention.

1688: N. Warmoltz: Over het mechanisme van den capacitieven ontsteker en van den weerstand bij kwikdampgelijkrichters (thesis Delft 1946) (On the mechanism of dielectric-ignition and resistance ignition in mercury arc rectifiers).

A short survey is given of ignition methods in mercury pool rectifiers, based on the field theory of the low pressure mercury arc. The time lag of the dielectric ignitor is measured by oscillographic methods. It is in accordance with the spaces of time required for the rupture by an electric field of a small distortion on the mercury surface as computed by Tonks. The field strength at the sharp point, which is formed during this process, is sufficient for cold emission of electrons, from which an arc develops in the mercury vapour formed simultaneously.

As regards the behaviour of the time lag of resistance ignitors in experiments with frozen cathodes, preheating the ignitor, arguments are found in favour of Mierdel's theory on this subject.

1689: J. van der Vliet: De provitaminen-D van de mossel (mytilus edulis), thesis Groningen 1946 (The provitamines D of the mussel).

The mixture of provitamines D from the sterol fraction of the mussel is investigated. From the product obtained from this mixture by irradiation with ultraviolet rays a crystalline substance with antirachitic properties has been isolated (called D_x), which on closer examination proved to consist of a mixture of calciferol and an unknown antirachitic, almost inactive compound related to vitamine D, having the probable composition $C_{28}H_{44}0$ or $C_{29}H_{46}0$ and containing an unsaturated side branch. Further vitamine D_3 was isolated as a crystalline ester. Oxidation of the irradiation-product with ozone gave different new products, who where also identified. From chemical as well as biological data it is concluded that the composition of the provitamine mixture is:

7-dehydro-chloresterol	50%
ergosterol	17%
cholestatriene -5, 7, 22-ol-3	33%.

The presence of the two first mentioned compounds is proved, that of the last one made probable.

1690: N. H. W. Addink: Complete and incomplete crystals. (Nature **157**, 764, 1946).

The author presents a table containing which to his opinion are the most reliable values of N (Avogrado's number, chemical scale), calculated from the density and from the dimension, as measured by X-ray methods, of the elementary cell in the case of diamond, KCl, quartz, calcite, PbO, Sn and various other metals (partly from measurements by the author and coworkers).

It is seen that the value of N found increases in the order mentioned, corresponding to an increasing deviation from the ideal density. The term incomplete is proposed for crystals showing a density inferior to the ideal value. The value for 10^{-23} N is $6.0227^5 + .0003$ for diamond, which is very near to Birge's value of $6.02283 \pm .0011$.

1691: Balth. van der Pol: The fundamental principles of frequency modulation (J. Inst. El. Engineers **93**, 153-158, 1946).

In a lecture for the radio section of the I.E.E. the author treats:

1) the problem of finding the possible current and voltage in a circuit with L, C, r and condensor leak R being arbitrary functions of the time, and
2) the problem of the response of a given network to a frequency modulated signal. In addition the definitions of amplitude phase and frequency are discussed.

1692: A. Claassen and J. Visser. The determination of uranium with oxy-quinoline (Rec. Trav. Chim. Pays Bas **65**, 211-215, 1946).

The method of Hecht and Reich-Rohwig for the determination of U by means of O-oxy-quinoline has been subjected to a critical investigation. It was found that washing with a hot 0.04% oxine solution is to be preferred to washing with hot water. Results are accurate within 0.1%. Separation from Mg, alkaline earths and alkali metals is complete.

VOL. 8 No. 10, p. 289-320 OCTOBER 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

A NEW SERIES OF SMALL RADIO VALVES

by G. ALMA and F. PRAKKE.

621.396.694-181.4

The development of radio valves having flat bases and of either all-glass construction or with metal envelopes made it possible to reduce the dimensions appreciably, and particularly the overall length, while at the same time maintaining good performance at very high frequencies. The downward trend of dimensions was halted, in the case of metal valves, at the point where further reductions in size resulted in excessive mutual capacitances and dielectric losses between internal leads, and, in the case of all-glass valves, at the point where the proximity of the electrode assembly to the seal between base and envelope exposed the cathode to excessive temperature during the sealing operation. In a new series of valves now developed by Philips, and known as the "A" series or "Rimlock" valves, over-heating of the cathode is avoided by joining bulb to base with a glaze or cement which becomes plastic at a comparatively low temperature. Further substantial reductions in bulb diameter have been achieved. A diameter of 22 mm has been adopted to permit the use of eight contact pins in the base, thus providing the maximum number of connections required in normal receiving valves, and to ensure that there shall be no over-heating of the bulb in the case of valves of the highest dissipation (*i.e.* 14 W for an output pentode). This article describes in some detail the construction and manufacture of these valves.

Over a considerable period the general internal structure and form of envelope (bulb and lead-in wires) had become stabilised, the so-called "pinch" construction being considered normal practice. But certain new requirements in radio engineering called for new developments — developments which would satisfy, for example, the needs of the manufacturers of small, inexpensive sets for which there is a demand in most countries, and the requirements for short-wave and very-short wave reception including television reception.

Small, compact receiving sets call for the production of valves and other components of the smallest possible dimensions, and in this connection the past seven or eight years have seen remarkable developments, the first stages of which have already been the subject of articles in this periodical [1][2]). In the new series of valves about to be described, further reduction in size is once more a prominent feature but, as will be explained, this reduction has been made possible by a new manufacturing technique which has at the same time resulted in improved performance in short-wave operation.

Glass and metal valves with flat bases

In valves which are to be used for short-wave (and, when required, for ultra-short-wave) reception it is essential, amongst other requirements, that capacitances between the leads to the various electrodes, and the dielectric losses in the insulating material between these leads should be very small.

An important advance in this direction was made when the original "pinch" construction was abandoned in favour of a bulb with a flat base [1]). In these valves the lead-in wires were much shorter; they were enclosed in the glass for a short distance only; and they were spaced well apart. These leads could also be made to serve as the contact pins, thus avoiding the necessity of fitting a base cap of plastic material, which has been the source of large and variable losses. Moreover, the short overall length of the leads resulted in proportionally smaller self-induction and opened up new possibilities for the use of valves of this type for very high frequencies (metre waves) [3]).

At this stage, the wished-for reduction in valve dimensions and the equally desirable improvement in performance on short-waves were being attained

[1]) Philips Techn. Rev. 4, 162, 1939.
[2]) Philips Techn. Rev. 6, 318, 1941.

[3]) See for example Philips Techn. Rev. 3, 103, 1938.

by one and the same means. But when further development was attempted on the same lines the two requirements came into conflict. Any considerable decrease in size of receiving sets demanded a reduction in the diameter of the valves, since this dimension largely determines the area of the receiver chassis. But reduction of valve diameter involves a closer spacing of contact pins in the flat valve base, and this results in increased mutual capacitances and dielectric losses.

This difficulty was most pronounced in those valves which employed both bulb and flat base made entirely of metal, with the contact pins fused into the base with glass beads, and this construction gave little prospect of further reductions in valve diameter. It had, however, one important advantage over the all-glass construction, in that the sealing of the metal bulb to the base is achieved by only very slight increase of temperature in the electrode system, since the heavy welding machine employed for this operation generates, in a single current impulse of short duration, an accurately calculated amount of heat which is quickly dissipated to the welding electrodes and surrounding material due to the high thermal conductivity of the metal parts. In the all-glass construction, however, bulb and glass base had to be raised to a temperature of some 800 °C to 900 °C, and in proportion as the diameter of the valve and the length of the leads are reduced, the distance between the parts of the electrode system and the seal also becomes smaller, and the temperature to which they are exposed during the sealing operation becomes greater. The smaller the valve, therefore, the greater the risk that, during sealing, parts of the electrode system may be oxidised and that the cathode may have its emission impaired by the chemical action known as "poisoning". Excessive manufacturing rejects can, in these circumstances, be avoided only by passing an inert gas such as nitrogen through the valve during the sealing operation — a manufacturing complication which cannot be contemplated with equanimity.

The "glazing" technique

The problem outlined above has now been solved in the case of all-glass valves by adopting an entirely new method for joining the bulb to the flat base. Instead of direct fusion, a "glaze" or cement is used, the material selected having a melting point much below that of the glass.

When the glass base with its moulded-in contact pins has been made, a moulded ring of the powdered cement is placed round the upper edge of the base. The whole is then raised to the temperature at which the "glaze" melts and becomes firmly bonded to the glass base (see *Fig. 1*), and is then slowly cooled to relieve the mechanical stresses in the glass.

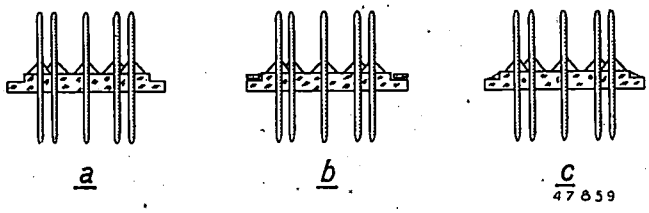


Fig. 1. *a*) Base plate of a radio valve of A series with moulded-in contact pins. Actual size.
b) Same base plate with ring of glaze laid upon it.
c) The ring of glaze is fused on the glass.

The electrodes are then assembled to the supports, the base-plate, complete with electrode assembly, is placed over the inverted bulb in a special sealing machine, and heat is applied until the cement softens, when the base sinks by its own weight far enough to allow the rim of the glass bulb to penetrate the layer of cement. On cooling, the cement sets and adheres to the edge of the bulb to form a vacuum-tight seal (see *Fig. 2*). During the whole of this process it is only the cement and not the glass which is softened, and for the material used in the valves now in production the temperature employed is only some 450 °C. The electrodes, and particularly the cathode, attain a temperature not exceeding 230 °C as compared with 500 °C to 600 °C in the previous method of sealing.

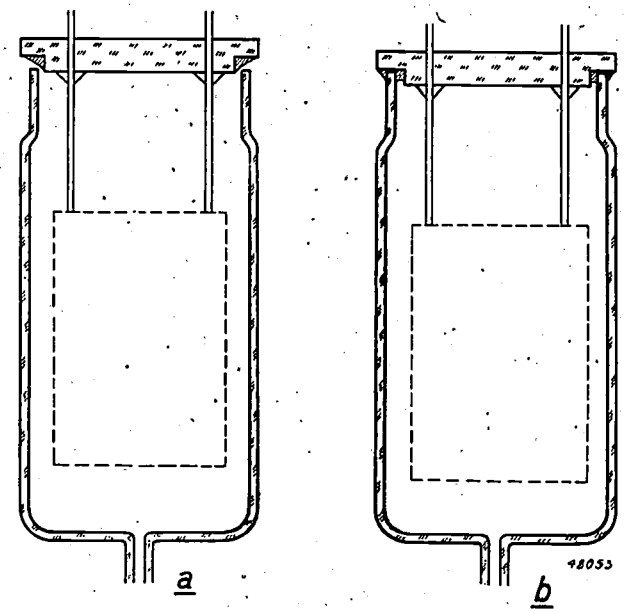


Fig. 2. The joining of bulb and base plate by the glazing technique. 1.5 times actual size.
a) The base plate with the fused ring of glaze is laid loosely on the edge of the bulb.
b) As the glaze melts the edge of the bulb penetrates into the layer of glaze and after the latter has set the bulb and base are securely joined and vacuum-tight.

In order to avoid risk of dangerous mechanical stresses being set up in the valve while cooling after sealing, it has been necessary to find a glaze or cement having a coefficient of thermal expansion practically equal to that of the glass used for the bulb and base. Moreover, the material had to have a suitable surface tension, so that the meniscus of the layer of glaze is slightly concave (see Fig. 1*c*) to prevent the base from becoming displaced to one side when placed on the bulb as shown in Fig.2*a*.

Fig. 3. Detail of the sealing machine on which bulb and base plate are joined. A metal cap is placed on the base plate (left). The whole passes between two rows of gas flames, the layer of glaze being heated to about 450° C, necessary for fusing.

Fig. 3 shows in detail the machine used in sealing the bulb to the base. A metal cap is suspended over the upward-projecting contact pins of the base. The whole passes between two rows of flat flames directed towards the cap. The edge of the bulb and the base are thus heated uniformly, so that no mechanical stresses are set up in the glass. The weight of the cap assists the rim of the bulb to penetrate well into the layer of cement.

This "glazing" technique has made possible the new series of valves known as the "A" series or "Rimlock" valves, having a diameter of only 22 mm. For the sake of comparision it may be mentioned that the "B" series described previously have a diameter of 32 mm., and the "C" series, which included the short-wave push-pull pentode type EFF 50, a diameter of 36 mm [4]).

Some of the types in the "B" and "C" series are not suitable for inclusion in the "A" series for, reasons which are explained later.

Fig. 4a shows a number of valves in the "A" series; Fig. 4*b* a valve in the "A" series compared with valves in the "B" and "C" series; while the reduction in dimensions is best illustrated in *Fig. 5*, where four successive models of two different valve types, namely an I.F. pentode and an output pentode of similar performance, are shown side by side.

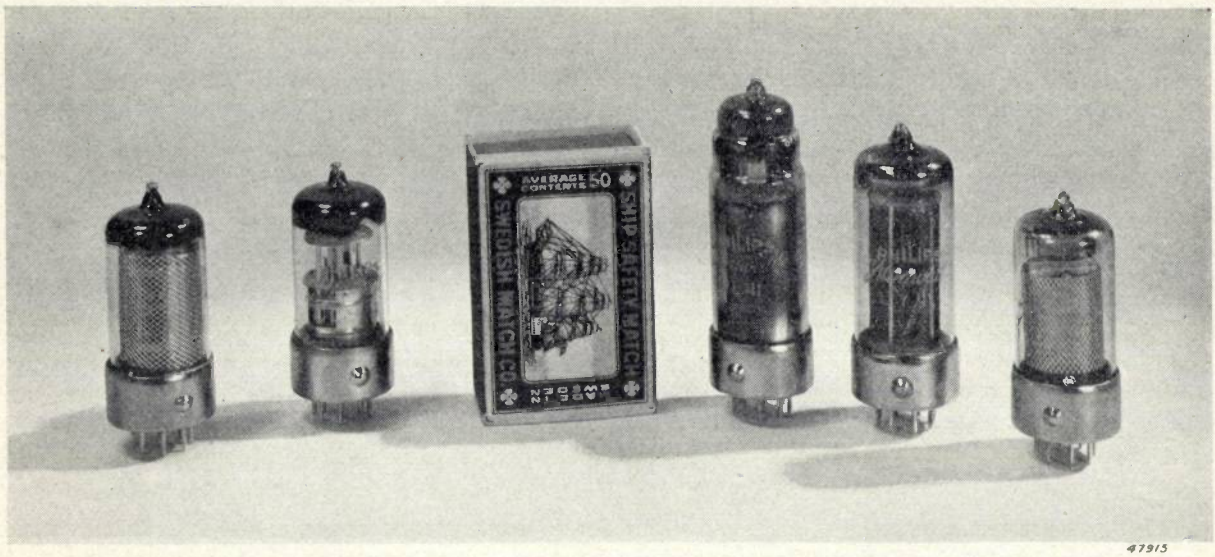
In addition to making possible valves of more compact dimensions, the "glazing" technique confers the general advantage that the bulb can be made to any desired shape, accurate to within 0.1 mm., and that this shape will be maintained during the sealing operation. As examples, *Fig. 6* shows two bulbs without electrode systems. That on the left has a constriction of the envelope close to the base — required for a definite purpose to be explained later. Only by the "glazing" technique can such small details be maintained in the finished product — they would be entirely lost owing to the softening of the glass if exposed to the high temperatures employed in the earlier method of sealing.

A particularly troublesome result of glass softening occurs in small diameter valves, the plastic condition of the glass allowing the contact pins to depart from their vertical alignment, or even to shift their location slightly, with the result that the pins have to be straightened by force. To avoid risk of cracking the glass during this adjustment, the pins must be made of soft metal. This, in turn, introduces the risk that the pins may become bent in service, so that the valve will no longer fit the holder. Using the "glazing" technique, however, the base retains its shape throughout the sealing process, the need for bending the pins disappears, and it is possible to employ pins made of hard metal, not only for the new small diameter "A" series, but also for the valves of larger diameter in the "B" and "C" series.

Apart from the practical advantages already indicated, the "glazing" technique results in a simplification of manufacture and a speeding up of production, the latter being due not only to the

[4]) The development of the B and C series, the so-called "key valves", is described in full in the article referred to in footnote [2]).

a)



b)

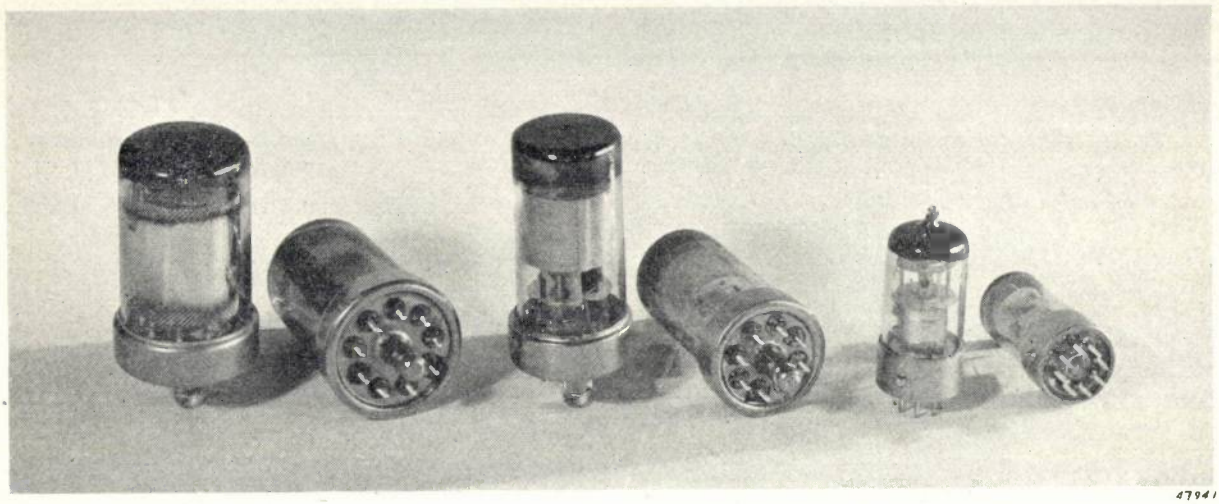


Fig. 4. *a*) Five valves of the new A series ("Rimlock" valves) diameter 22 mm designed for a normal receiving set. The types shown from left to right are: UCH 41, a triode-hexode; UF 41 an intermediate-frequency pentode; UAF 41, a diode-pentode; UY 41 a rectifier valve and UL 41, a 9 W output pentode. These valves are suitable for A.C. and D.C. sets.

b) For the sake of comparison the following are shown side by side: UCH 41; ECH 21, a triode-heptode which because it requires 9 contacts is made in the previously described B-technique (32 mm diameter); and the short-wave push-pull amplifier valve EFF 50 in the C technique (36 mm diameter).

general simplification, but also to the fact that the lower sealing temperature can be attained more quickly than the higher one.

The new valves

Choice of diameter

Had the cathode temperature during the sealing operation been the sole consideration, a diameter considerably smaller than 22 mm could have been selected. But it is the capacitances and other losses between leads which are now the limiting factors, and the extent to which the diameter can be reduced therefore largely depends upon the number of contact pins it is necessary to set in the base. A further factor is the amount of heat generated in the valve which, in conjunction with the valve diameter, determines the working temperature of the envelope and thus affects to a large degree the dielectric losses and electrolysis of the glass between the pins.

The number of pins required and the amount of heat generated vary widely between valves of different types. Thus, a rectifier valve for a radio receiver needs only four pins, while an indirectly-

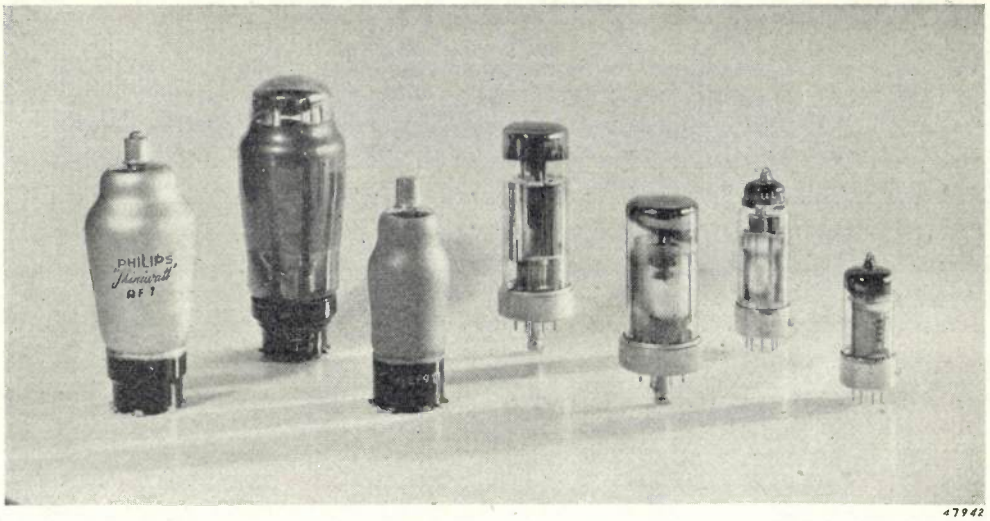


Fig. 5. Successive models, with about the same performance, of an intermediate-frequency pentode (front row) and a 9 W output pentode (back row). Front row, left to right: AF 7 (1937), EF 9 (1937), EF 22 (1941), EF 41 (1946); back row: EBL 21 (1941), EL 41 (1947).

heated frequency changer of the triode-hexode type (which is to be preferred owing to its inherent constancy of tuning on short waves) requires at least eight pins — six for the electrodes (the cathode being common to the triode and hexode sections and two pairs of grids can be interconnected) and two for the cathode heating.

Thus, in principal, different minimum diameters could be selected for each valve type. But this would be very impracticable from the standpoints of both the valve-maker and the set-designer, both of whom desire the greatest possible degree of standardisation of components.

One diameter only has therefore been chosen for the whole of the "A" series — a diameter large enough to accommodate the maximum number of eight pins and one which does not introduce the risk of excessive bulb temperatures being developed at the maximum total dissipation in any valve designed for a normal receiving set — namely the 14 watts dissipated in a 9 W output pentode which, in order to obtain a high mutual conductance, is provided with a long cathode consuming $4^1/_2$ W. The few special valves which require nine contacts are made in the "B" or "C" series.

The diameter of 22 mm. selected for the "A" series provides ample security so far as insulation at high voltage is concerned. It has even been found that a special television valve designed for a peak voltage of 4 000 volts could be operated with that voltage applied between two diametrically opposite pins without risk of electrolysis of the glass or of breakdown. This valve is illustrated in *Fig.* 7.

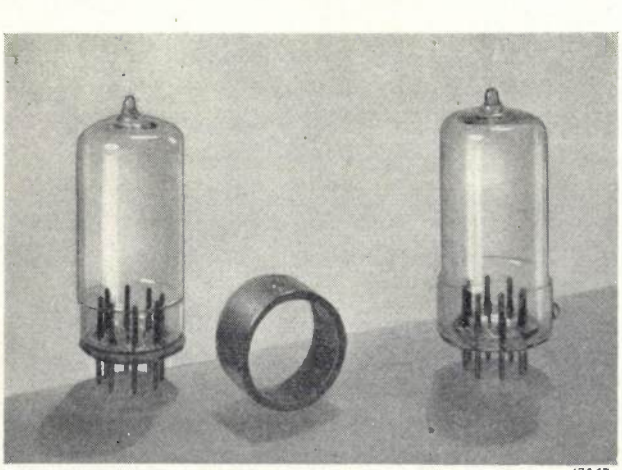


Fig. 6. Photograph of the envelope of the A valves without electrode system. Evacuation and sealing off take place at the top of the valve. The sealed off point is reinforced by moulding it into a compact conical form immediately after sealing off. The two envelopes shown have different forms of guiding mechanism for mounting the valve in the holder (see the final paragraph of this article).

In the U.S.A. very small all-glass valves known as miniature valves with a diameter of 17 mm. have been developed in recent years. These valves, however, have a maximum of seven contact pins. As a consequence, an indirectly-heated frequency changer of the triode-hexode type cannot be made in that series, and only output pentodes of low mutual conductance or small output, unless a high working temperature of the bulb is accepted.

Construction of the electrode system

It is generally desirable to construct the electrode system of a radio valve in such a way that it can be mounted on two or three support rods and that it stands free from the walls of the envelope. This construction is not only simpler than that in which the electrode system is supported from the walls

by rings or discs, but it also reduces the risk of the assembly being distorted while being inserted into the bulb. The free-standing construction is not, however, easy to achieve in bulbs of small diameter owing to the risk of the electrode system striking the walls of the envelope under conditions of mechanical shock or vibration, thus giving rise to noises when the set is in operation.

The selected diameter of 22 mm is, however, sufficiently great to permit free-standing construction to be adopted for most types, including the triode-hexode, pentode and diode-pentode, each of which has a length of 43 mm.

Only for the 9 W output pentode and the rectifier in the new series has it been necessary to support the electrode system from the walls, these valves having comparatively long electrode assemblies and bulb lengths of 61 mm and 52 mm respectively (see Fig. 4*a*).

The multiple valves in the "A" series, such as the triode-hexode and diode-pentode, differ from earlier valves of similar type in that the more complex of the two electrode systems is mounted below the simpler. This simplifies the construction of the support rods and the mica discs employed to give the whole assembly the necessary rigidity. At the same time it confers advantages in connection with the effect of the conduction of heat through the current leads at the bottom of the cathode. A short axial extension of the cathode serves for the simpler system, and the remaining, and longer portion of the cathode for the more complex system. The average temperature of the longer section will obviously be less affected by heat conduction at one end than that of the shorter section.

The "glazing" technique has special advantages in the case of battery valves, in which it is important to keep to a minimum the power required to heat the cathode. A directly heated cathode consisting of the thinnest practicable wire with a very thin coating of a highly emissive oxide is thus indicated [5]). Using the normal nickel filament, wire diameters below 20 microns are not practicable because of the low tensile strength of the material. It was for this reason that for many years Philips have used tungsten filaments which have much higher tensile strength and can therefore be made much thinner. Tungsten wire is, however, more susceptible to oxidation than nickel, and it is in this connection that the low sealing temperature employed in the "glazing" technique confers a great advantage, permitting the use of a filament wire only 8 microns in diameter, for which a heating current of only 12.5 mA is needed.

Fig. 7

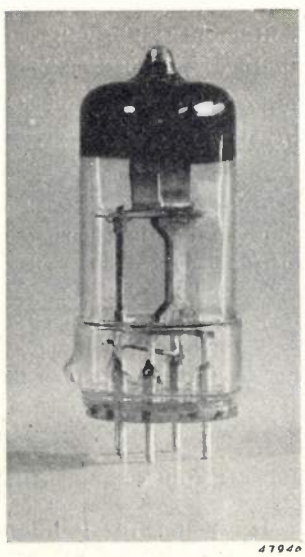


Fig. 8

Fig. 7. Diode for television purposes. The highest voltage that may come to lie between 4 diametrically opposite contact pins amounts to 8000 volts.
Fig. 8. Double diode (mixing diode) for ultra short waves, with glass guiding stud on the bulb.

Valve holders

A radio valve and its holder should be so constructed that when fitting the valve in the holder there is no risk that the pins can enter the wrong sockets.

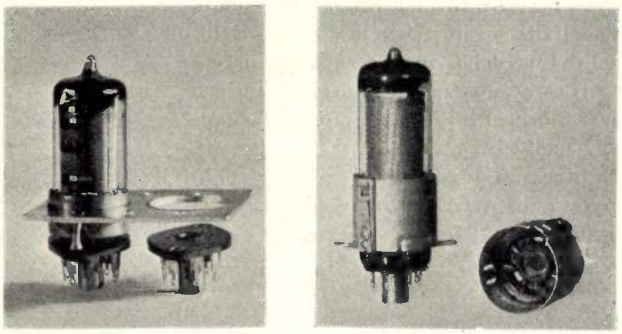
The earliest method of ensuring this was to space the contact pins non-uniformly in the valve base, but it was then difficult to insert the valve correctly into a holder, especially when the latter was located in a more or less inaccessible position in the receiver. Furthermore, in the small all-glass valves where the old sealing technique necessitated the use of soft metal pins there was considerable risk that valves would be forced into the holder, and this introduced the danger of the glass base being cracked.

The method adopted for the "A" series and also for the "B" and "C" series is to space the contact pins uniformly around a pitch circle and to provide some form of locating device to ensure that the valve is inserted in the correct position.

In the "B" and "C" series the locating device consists of a stud at the centre of the base. In the "A" series a metal ring is cemented to the lower portion of the bulb, where the diameter is slightly reduced. This ring carries a small rounded projection or boss which fits into a corresponding groove in the edge of the valve-holder — hence the name

[5]) In these valves the proportionately smaller cathode surface of the thinner wire has practically no effect on the valve characteristic.

"Rimlock". As already explained, the possibility of fitting this ring closely round the bulb within very small dimensional tolerances is entirely due to the low sealing temperatures required by the "glazing" technique.



a) 47943 *b)* 47944

Fig. 9. *a*) Simple valve holder for the A series. The stud on the rim of the bulb falls in a groove in the edge of the holder. The valves are therefore called "Rimlock" valves.

b) Valve holder with locking device. The valve is held fast by means of a spring which locks over the stud when the valve is inserted. The small can in the middle of the holder provides a screen between the contact pins right up to close under the base plate.

Further, owing to the complete absence of bulb distortion, the metal ring can be dispensed with and the locating boss can be formed on the glass wall itself, as is shown in the empty bulb illustrated in *Fig. 6*. *Fig. 8* shows a double diode (mixing diode) for ultra short-wave operation. This solution is particularly advantageous in ultra short-wave valves as the metal ring would substantially increase the capacitances of the leads.

The locating device described above makes possible great simplification in the design of valve-holders. In *Fig. 9a*, for example, a holder is shown consisting of a flat plate carrying the contacts and suspended by two bolts about 8 mm below the top deck of the chassis. A circular hole in the chassis permits the valve to be inserted, and a notch at the correct point on the circumference of this opening serves to accommodate the locating boss of the valve.

It is often desirable, particularly when a receiver has to be transported with the valves in position, to combine a locking device with the locating arrangement. A holder incorporating a locking device is illustrated in Fig. 9*b*.

In Figs. 9*a* and 9*b* can be seen a small metal can at the centre of the holder. When the valve is in position the can reaches just below the underside of the valve base and serves as an electrostatic screen between the contact pins, especially between the anode and control grid pins. In the "B" and "C" series valves this function was, of course, performed by the central spigot.

THE USE OF ISOTOPES AS TRACERS

by A. H. W. ATEN Jr. and F. A. HEYN.

539.16.08:539.155.2

When a small amount of a suitably chosen radioactive or stable isotope is added to one of the substances taking part in a physical or chemical process, it is possible to follow the atoms of that isotope right through the process, to detect their final resting place and to study their distribution by means of their radioactivity or difference in atomic weight. The isotope thus used as a tracer can be employed to furnish information about the nature and the progress of the process in question. This is especially true for processes in which there is an exchange of identical particles. Since such processes are very important in physiology and chemistry, and the tracer method is at present the only known method of demonstrating and investigating such processes, it has found extensive application in these fields. But also for numerous investigations which in principle could be carried out in other ways, the tracer method is found to be of great value, thanks partly to its extremely high sensitivity and partly to the great ease with which the questions presented can be answered. In this article a series of examples is given illustrating the use of the method.

Introduction

When the periodic system of chemical elements was set up in the course of the previous century, it was thought that each element consisted of only one definite kind of atom. Later this was found to be incorrect: an element may consist of different kinds of atoms which have practically identical chemical properties — the criterion for denoting the atoms by the name of the respective element — but differ in atomic weight by one or more units. Such isotopic atoms, so called because they have to be given the same position in the periodic system, may be stable or unstable. In the latter case they undergo a gradual change, accompanied by a radiation, into another kind of atom; they are then radioactive.

Almost all the kinds of atoms occurring in nature are stable. There are only a few unstable ones, namely the well-known substances with natural radioactivity, such as radium, thorium, uranium etc. In addition to these, however, it is nowadays possible to turn each element into one or more new isotopes which do not occur in nature, all of which are unstable (artificial radioactive substances).

A single example will serve to illustrate the above. The element calcium occurring in nature consists of at least six stable isotopic kinds of atoms, namely for 96.96% Ca^{44}, *i.e.* calcium atoms with an atomic weight of 40 (in round numbers), and further 0.64% Ca^{42}, 0.15% Ca^{43}, 2.06% Ca^{44}, 0.0033% Ca^{46} and 0.19% Ca^{48}. Furthermore, up to the beginning of the year 1944 it had been found possible to make artificially six more radioactive calcium isotopes with atomic weights 39, 39, 41, 45, 49, 49. The various unstable isotopes can further be distinguished from each other by the character of their radioactivity. In the case of a radioactive atom a certain percentage of the atomic nuclei present is transformed per second into another sort by spontaneous disintegration, each disintegrating nucleus emitting, according to its sort, a negative or positive electron or an α-particle (helium-nucleus). This is often accompanied by an electromagnetic radiation (γ-radiation). The intensity of the total radiation of a radioactive preparation at any moment can easily be measured. As the number of non-disintegrated atomic nuclei is continually decreasing, while the chance of disintegration remains constant for each atom and thus also the percentage of nuclei disintegrating per second, the radioactivity observed decreases with time. The velocity of this decrease usually expressed by the "half-value time", *i.e.* the time in which the intensity of the radiation falls to one half, is characteristic for each isotope. The above-mentioned six radioactive calcium isotopes have half-value times of 4.05 minutes, 1.06 seconds, 8.5 days, 150 days, 2.5 hours and 30 minutes respectively.

The existence of the isotopes is not merely of theoretical interest. In the last 10 to 20 years isotopes have become an extremely useful practical aid for all kinds of scientific and technical investigations. This use of isotopes is based for a large part on the fact that an isolated isotope of an element takes part in chemical and physical processes in exactly the same way as the familiar mixture of isotopes of that element which occurs in nature, while the isotope in question can always be recognized by the investigator and can be traced even in a chemically identical environment, thanks to its radioactivity or difference in atomic weight. The isotope thus functions as a tracer, capable of furnishing information about the process taking place,

which would be much more difficult or even quite impossible to obtain in any other way.

We shall explain this in more detail, but for better orientation of the reader we shall first discuss four examples out of the large number of investigations which have already been carried out with isotopes as tracers.

What can be done with (radioactive) isotopic tracers

First example

In the production of steel, among other substances the phosphorus, which is present in quite considerable quantities in the crude iron, has to be rendered harmless by adding a slag-forming substance or by lining the crucible with a material that reacts with phosphorus. A continuous check has then to be kept of the amount of phosphorus still present in the molten metal. This can, of course, be done by chemical analysis of samples, but results are obtained much more quickly and easily when a small amount of radioactive phosphorus is added to the melt at the beginning of the process. This is rapidly distributed uniformly throughout the melt, so that the ratio between the natural phosphorus present and the radioactive phosphorus added is the same everywhere. If phosphorus disappears from the melt into the slag floating on the surface or into the lining of the crucible this takes place to an equal degree with the natural and with the radioactive element. The decrease in the percentage of phosphorus in the melt can thus be determined merely by ascertaining the decrease in the radioactive phosphorus. This is extremely simple, since it is only necessary to measure the radioactivity of a sample of the melt, which can be done very easily with an electrometer or an electron-counter [1]).

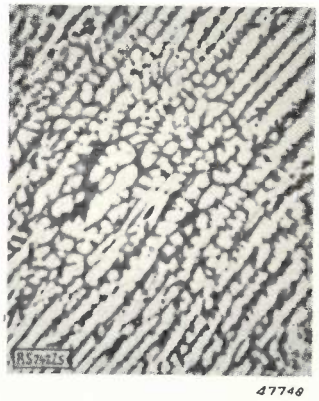
Second example

In many factories the workers come into contact with mercury, and it is known how harmful the regular inhalation of mercury vapour can be; in course of time a concentration of more than 10^{-4} gram of mercury per m^3 of air already becomes injurious to health. It is a difficult problem to detect the presence of mercury in such minute proportions, because chemical analyses are unavailing in such cases. In a certain case which occurred in the manufacture of tubular luminescent lamps in an American factory, where the lamps were filled with mercury vapour at a low pressure, a glass side-tube containing a small drop of mercury had to be "blown" onto the lamp, and inevitably the glass-blower inhaled a very small quantity of mercury vapour into his lungs. In order to determine how much was inhaled a number of tests were carried out with a volume of 2 liters of air that had been in contact with the drops of mercury under exactly the same conditions as in the manufacturing process, this being drawn off by suction and passed over a metal plate kept at the temperature of liquid air. Practically all the mercury in the air condensed on the plate.

The mercury used for the experiments contained a known, small percentage of a radioactive mercury isotope. The radioactivity of the plate, which was quite simple to measure after the experiment, gave an indication of the amount of mercury contained in the air which had passed over the plate. In this way a concentration of 5×10^{-6} g/m³ could be detected. Average mercury concentrations were found of about 10^{-5} and in one case about 4×10^{-5} g/m³, from which it was concluded that in the manufacturing process in question the glassblower ran no danger of poisoning [2]).

Third example

When a piece of metal is fused with a radioactive lead isotope in an atmosphere of hydrogen and then allowed to crystallize again, there are two possibilities. In some metals, such as thallium and magnesium, lead is soluble to a considerable percentage,



47748

Fig. 1. Radiogram of tin with a radioactive lead isotope (thorium B) deposited at the grain boundaries of the polycrystalline metal. Magnified about six times (from G. Tammann and G. Bandel, Z. Metallk. 25, 154, 1933).

the lead atoms being uniformly distributed in the grains and the polychristalline metal obtained. In other metals, on the other hand, such as bismuth, tin, antimony, silver, gold, copper and nickel, in

[1]) An electron counter was described *e.g.* by A. Bouwers and F. A. Heyn, in Philips Techn. Rev. 6, 75, 1941. In the measurements allowance must of course be made for the natural decline in radioactivity with time. We shall return to such practical details on a later occasion.

[2]) J. W. Irvine and C. Goodman, J. appl. Phys 14, 496, 1943.

which lead is practically insoluble, the radioactive lead is situated on the boundaries of the grains. When a microscopic preparation of the metal is made and laid on a photographic plate for several hours after development, the plate will be found to be blackened at those places where it lay against radioactive particles, thus where lead has been deposited. By means of such an "auto-radiogram", of which *fig. 1* is an illustration, it is possible in the first place to ascertain whether and to what degree the added lead is soluble in the metal; if the lead is entirely dissolved the entire surface of the photographic plate is uniformly blackened; if the lead does not dissolve, or only partially, the radiogram shows up very nicely also the boundaries of the grains (fig. 1). Thus the isotopic method can also in this case furnish valuable information about the changes taking place in the structure of the metal upon recrystallization and in rolling [3]).

Fourth example

Friction between two metal surfaces is due partly to adhesion, the result being that when the surfaces slide over each other extremely small particles of metal are torn out of one and taken up in the other. This exchange of metal may take place to such an extent that two surfaces become, as it were, welded together (the familiar seizing). The quantity of material thus transferred from one metal to the other is a measure of the contribution of this effect to the total force of friction. In general it is a question of very small amounts which chemically can hardly be detected at all. An investigation has now been carried out with the help of a radioactive tracer [4]). One metal surface was "activated", *i.e.* it consisted for a small part of atoms of a radioactive isotope of the metal. After this surface had been made to slide over the second, non-activated metal surface, the latter also showed a certain amount of radioactivity. Amounts of 10^{-10} gram of transferred metal were detected in this way, and, what is more, by means of a radiogram as described in the preceding example also the distribution of the transferred material on the surface could be studied. From the radiogram shown in *fig. 2a* it may be concluded for example, that in this experiment the sliding of the two metal surfaces over each other was not continuous but took place in small jerks. By this method the influence of all kinds of factors, such as the pressure, the hardness of the surface, etc. on the transfer of material can be studied, as also the effect of a lubricant; see fig. 2*b*.

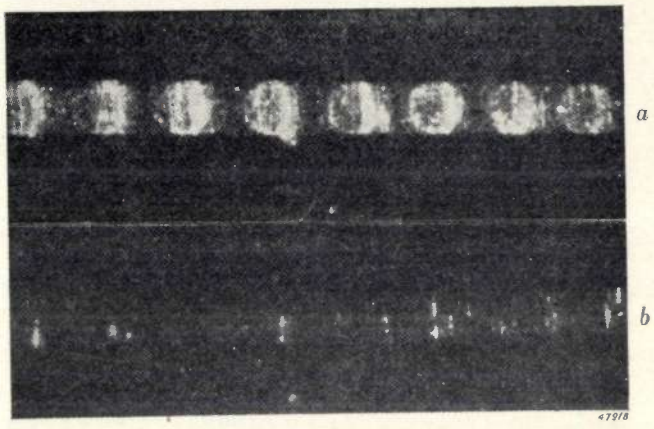


Fig. 2. Radiograms of a metal surface after it has been rubbed with a piece of metal containing a radioactive component. The radioactivity of the surface shows there has been a transfer of material. This explains to a large extent the friction set up when one surface slides over another. The illustration relates to the friction of lead on steel *a*) with no lubricant, *b*) with lubricant. Magnification about six times. (From J. N. Gregory, Nature (London) **157**, 444, 1946).

Why the tracer method is so important

From these few examples we can already deduce the most important aspects which have lent such great significance to the tracer method.

We first call attention to the last example discussed. The transfer of material can also be measured when the two surfaces sliding over each other contain the same metals or even when they are exactly identical. It must be realized that this would not be possible by any other known method, since the transfer takes place in both directions: there is an exchange of identical particles. It would not be possible by any chemical or physical method to ascertain the origin of the metal present on one of the surfaces after the experiment, whereas the radioactive isotope immediately gives the answer.

Processes in which there is an exchange of identical particles are very common in nature, not only in chemistry and metallurgy, but especially in the physiology of plants and animals. The tracer method, which offers the only method of approach, has been used on a large scale for the investigation of such processes, and the publications in that field are innumerable. In a survey of such investigations for the year 1940 [5]) in physiology alone more than 300 publications are cited. But also in the field of technology, for routine tests, and like-

[3]) G. Tammann and G. Bandel, Z. Metallk. **25**, 153 and 207, 1933.

[4]) B. W. Sakmann, J. T. Burwell and J. W. Irvine, J. Appl. Phys **15**, 495, 1944; J. N. Gregory, Nature (London), **157**, 444, 1946 (6 April).

[5]) J. R. Loofbourow, Rev. Mod. Phys **12**, 267, 1940.

wise in agriculture and chemistry, the method is being more and more widely applied [6]).

Although the last example discussed illustrates the possibility of studying the exchange of identical particles with the help of an isotope, the employment of the tracer method in that case is not motivated only by the possibility mentioned. When two different metals slide over each other the transfer of material could in principle be studied also by other methods. The fact that a radioactive tracer is, nevertheless, used, is due to the fact that a much greater sensitivity can be attained with the radioactivity measurements, *i.e.* much smaller amounts of a substance can be detected than with other methods so far available. The same applies to the case of the determination of mercury. In the determination of phosphorus which was discussed as the first example, the indicator method is not necessary in principle either, but it was applied there because of the greater ease with which the quantity of phosphorus could be determined, compared with a chemical or other analysis. Finally, an important point in the employment of a radioactive isotope is that it can be localized so easily when mixed with a different or a chemically identical substance, while also its distribution can be determined; see the radiograms of figs. 1 and 2.

Where radioactive isotopes are applied for the sake of the advantages of greater sensitivity and easier working, while in principle other methods would also provide an answer to the questions raised, one might speak of "untrue" applications of the indicator method. In such cases the nature of the isotope is sometimes a matter of indifference; in order to obtain a radiogram of the grain boundaries in a polycrystalline material it would also be possible to use a radioactive isotope of some metal other than lead, provided it does not dissolve in the base metal. In the "true" applications of the tracer method it is quite different, for there it is essential that an isotope can be detected in identical surroundings.

In this connection attention should be called to the fact that the tracer method can also be applied with non-radioactive isotopes. The atoms of such isotopes are recognizable (labelled) by their different atomic weight and the properties connected therewith, such as specific weight, velocity of diffusion, heat conductivity, etc. The most important atoms to be considered are "heavy hydrogen" (deuterium) of atomic weight 2 (approx.), the oxygen isotope of atomic weight 18 and the nitrogen isotope of atomic weight 15. With such stable isotopes the measurement of the radioactivity of mixtures of isotopes is replaced by measurements of density or the like.

These measurements are generally much less easy than the measurement of radioactivity and also not so sensitive by far. With stable isotopes there is, therefore, no question of "untrue" applications of the tracer method. The reason for using these is solely the possibility of studying processes of exchange, where no suitable radioactive isotopes can be found.

Owing to the very large number of applications of the indicator method (true and untrue) it has become impossible, as well as purposeless, to give a survey of these applications, even if one confined oneself to a definite field. We shall not, therefore, attempt to do so, but in the following we shall say something about the origin of the method and follow this up with a number of suitably chosen examples, with the intention of showing the possibilities of the method from different angles. In a subsequent article we shall go more deeply into the practical performance of investigations with radioactive and also with stable isotopes. As to this practical performance we can only point out here that it is not necessary to prepare the radioactive (or stable) isotopes oneself, for they can be obtained from certain suitably equipped laboratories. In Europe the Philips Laboratory in Eindhoven, among others, has already supplied suitable radioactive substances for a number of applications.

Origin of the method

The tracer method was initiated by Hevesy, who first discovered the possibility of studying processes of exchange by that means and immediately put his ideas into practice (in 1915). He used the method, for instance, to test the theory of Arrhenius about the dissociation of electrolytes. In essence his experiment was as follows. From a certain amount of normal lead a lead salt is prepared, for instance lead chloride, and from a corresponding amount of the radioactive lead isotope, which is formed as a disintegration product of radium, another salt, for instance lead nitrate is prepared. When the two salts are dissolved in water, the solutions mixed, and then the two salts extracted separately from the mixture, the two lead compounds are found to have become equally radioactive.

The lead atoms from the two salts must, therefore,

[6]) See, for example, the survey of chemical applications by G. Seaborg, Chem. Rev. 27, 199, 1940, where more than 500 publications are mentioned, most of them different from those in the article by Loofbourow.

have been completely mixed in the solution. This result agrees entirely with the hypothesis that the lead compounds are dissociated in the solution, *i.e.* that lead occurs therein in the form of free ions.

We have just said that one salt was prepared from normal lead and the other from the radioactive lead isotope. Consequently, in order to carry out the experiment in this way a sample of the pure radioactive lead isotope would have to be available. Actually, however, this is not necessary. It is sufficient if one sample of lead contains only a small amount of the radioactive isotope. Hevesy recognized this from the very beginning, as may be seen from the curious story of the way in which he came to use the isotopes in this way. He had tried in vain to separate radium *D* from a quantity of lead containing a small amount of that substance. Since radium *D* is an isotope of lead (the radioactive isotope mentioned in the radium series; its name dates from the time when there was no clear idea of the situation), it cannot, as we now know, be separated by ordinary chemical means. It was just this failure that gave Hevesy the idea that he could always distinguish the lead of this sample "contaminated" with radium *D*, from a sample of ordinary lead: in all mixtures with ordinary lead every fraction of the "contaminated" (radioactive) lead sample takes an equal fraction of the original radioactivity with it and can thus be determined quantitatively by measurement of the radioactivity.

The "contaminated" lead is thus, as it were, indicated or labelled by the radioactive isotope itself, and provided it is a homogeneous mixture the whole sample can serve as a quantity of labelled atoms. This is in fact obvious when it is borne in mind that the radioactivity of an element only means that per unit of time a certain percentage of the atoms present in a sample disintegrates spontaneously. If the sample also contains a number of isotopic atoms which are stable and thus will never disintegrate, the only result, in the first instance, is that the percentage of disintegrating atoms per unit of time, is smaller, thus the radioactivity is "diluted". In fact, also in the first examples discussed there were certain dilutions of radioactive phosphorus and mercury, and this is usually the case with artificial radioactive substances where the degree of "concentration" of the radioactivity depends upon the preparation of the substance. We shall return to this in a subsequent article.

Hevesy's applications of radioactive substances were not confined to exchange experiments. He realised also the significance of radioactivity measurements as a substitute for chemical analyses, owing to the great ease and sensitivity of the method as illustrated by our first examples, and he thus also made use of "untrue" applications of indicators. Once, when he had reason to suspect the cleanliness of his landlady, he smeared a bit of "dirt" with a radioactive substance on his dinner plate and checked daily whether the plate had been properly washed simply by measuring the radioactivity which (literally) still clung to it. He was indeed able to detect radioactivity of the plate for many days. Whether or not this was to be ascribed to the carelessness of the landlady or to the extreme sensitivity of the method, history fails to relate.

Denomination of the method

Hevesy called a radioactive isotope used for the experiments described, an indicator, and thus following his example one often speaks of the indicator method. In recent years in English-speaking countries the terms "tracer method" and "tracer atoms" have become more usual: the radioactive isotopes are used, as it were, for discovering and following a trace. "Labelled" and "tagged" atoms are also often spoken of. The terms speak for themselves. Finally, to complete the list, we may mention the denoting of these atoms as "spies", as proposed by Evans [7]). This name is meant to indicate that each atom of a radioactive isotope can move about unrecognized in a "crowd" of even similar atoms until at a certain moment it "betrays" its presence and whereabouts by its disintegration. The concentration of "spies" in the experiments usually lies between 1 to 10^{10} and 1 to 10^{15} normal individuals. Translated into terms of human society, this would be equivalent to one spy among a population at least five times as large as that of the whole earth.

Further examples of the application of tracers

The examples which will be discussed in the following in unrelated order, will give the reader an idea of the multifarious nature of the applications of indicators. In order to reduce them to some kind of systematic order we have sorted out the examples into three groups according to the character of the problem. In the first group the problem is only where something is situated (localisation), in the second group how much of a substance there is, remains over or takes part in a process (quantitative problems); in the third group it is particularly a question of exchange processes. It must be said, however, that often the boundaries between the groups cannot be sharply drawn. In localisation problems one is obviously always concerned with "untrue" indicator applications, as is also usually

[7]) R. D. Evans, Applied Nuclear Physics, J. Appl. Phys 12, 260, 1941.

the case in the second group (quantitative problems), while the last group contains only "true" cases.

Localisation

A very old example is the tracing of samples of radium that have been lost in hospitals through carelessness or theft; the places where the sample might possibly be, for instance the refuse heap, are gone over with an electron counter. A more modern case is the tracing of stoppages in an oil pipeline. In the periodical cleaning of the walls of the pipe a screw-shaped scraper is placed in the line and pushed along by the pressure of the oil itself. If the scraper gets jammed somewhere, it has to be located as quickly as possible, in order to open the line at that point and remove the accumulated deposit. With the help of a radioactive indicator this localisation is astonishingly simple. A scraper is used that contains a little radioactive material emitting γ-radiation, which easily penetrates through the walls of the pipe and can be detected with a somewhat modified "electron counter". One rides along the line with this "counting" apparatus until the radioactivity betrays the position where the scraper has stopped.

There is another similar application in the petroleum industry, for determining the setting depth of the cement that is pumped in behind the casing of an oil well. A radioactive mineral, carnotite, is mixed with the cement and when a counting instrument is lowered into the drill hole it indicates a strong radiation at the level of the cement.

Extremely fine cracks in metal surfaces can be detected and localised by applying a greasy paste containing a radioactive substance to the surface of the metal under high pressure. Upon the surface being cleaned, the radioactive paste is left in the cracks, and by making an auto-radiogram of the surface the cracks can then easily be seen.

Quantitative problems

In order to determine the efficiency of a fog, smoke or dust filter, it is necessary to measure the very small quantities of fog-forming or other substances retained by the filter. As fog-forming substance tricresyl phosphate containing radioactive phosphorus is used. When this is passed successively through several filters, their efficiency can be judged by comparing the intensity of their radioactivity.

Radioactive isotopes are sometimes an excellent means of measuring very small solubilities or very low vapour pressures. The vapour pressure of thorium acetyl acetonate, for example, has been determined by preparing a sample of the compound with a strongly radioactive thorium isotope, saturating a given volume of nitrogen with the vapour of the compound and passing the gas through acidified alcohol, in which the compound is absorbed. The thorium concentration can then be calculated from the radioactivity of the liquid and from that the amount of vapour in the given volume of nitrogen.

In general it may be said that radioactive isotopes render valuable service in microchemistry, *i.e.* the chemical investigation of extremely small quantities of a substance, as for instance in adsorption phenomena, in very dilute solutions, etc. Another interesting fact is that radioactivity has made possible the investigation of the chemical properties of the elements "43" and "85" and of several "trans-uraniums", elements which could not be found in nature but from which radioactive isotopes could be prepared artificially in imponderably small quantities.

Important perspectives are opened by the application of radioactive indicators in chemical analysis, so often constituting daily routine work in technology. One example out of many is the following: it is desired to determine the bromine content in a mixture of a bromide and a chloride. A complete separation of the two compounds is very difficult and takes up a great deal of time. If, however, a little radioactive bromine (in the form of the compound in question) is added to the mixture, only a partial separation of the bromine is sufficient. Due to the homogeneous mixing it is known that the ratio between the bromine separated and the total amount of bromine is equal to the ratio between the radioactivity separated out and the original radioactivity. Since the latter ratio is easily determined, it is possible to calculate the desired total content of bromine directly from the amount of bromine separated out.

An application in biology somewhat resembling the above is the determination of the total amount of blood in an animal. After a certain amount of blood has been taken death inevitably sets in and it is then impossible to draw off the rest of the blood. If, however, a solution of some substance or other containing a known quantity of radioactive atoms, is injected into the test animal intravenously and given time to distribute itself homogeneously throughout the whole circulatory system, a small sample of the blood suffices, for the total amount of blood can then be calculated from the percentage of injected radioactive atoms

recovered in the sample. (Of course no appreciable part of the injected substance must have been transferred from the blood to other parts of the body.)

The fact that it is practically unnecessary to interfere with the normal life of the test animal is of importance for many investigations, especially those of a pharmacodynamic nature. We may mention here an investigation into the rate of absorption of insuline which is injected periodically under the skin of sufferers from diabetes. It is often desirable to restrict the number of injections and it is therefore favourable if the insuline is retained for a relatively long time near the point of injection, or stored there as it were, and only slowly taken up in the circulation. By building a radioactive atom (radioactive iodine) into the molecule of three kinds of insuline, *viz.* "ordinary" insuline, globine insuline and protamine-zinc-insuline, and measuring from time to time the decrease of radioactivity at the point of injection, it has been possible to determine that the rate of absorption of the three kinds of insuline in the body decreases in the order given above.

Exchange processes

Although also in chemistry and technology numerous processes play a part where an exchange of identical particles occurs — we mention only auto diffusion, *e.g.* the diffusion of lead atoms in lead — physiology is the most prominent field for such exchanges. One of the most striking examples is the continuous exchange of the building materials of the body. This has been studied in detail with phosphorus in the form of various compounds, with the help of radioactive phosphorus, which lends itself so well for such experiments. Hevesy has particularly done a great deal in these investigations. It has been established that phosphorus does not remain permanently bound in any constituent of the body. The exchange takes place most rapidly between the blood and various organs: of the phosphate ions present in the blood at a given moment after two hours, only 2% are still present, the rest having been exchanged. In the liver and kidneys, too, there is a rapid renewal, but also in the bones and even in the brain the phosphorus is in course of time renewed, though at a much slower rate. The parts of the body that take least part in the continual exchange are the teeth: after 250 days only 1% of the phosphorus in the dental enamel is renewed.

In a certain case it is not so much a matter of exchange as one of selective assimilation of substances by certain constituents of the body, namely where the exchange leads, as it were, to a credit balance for that part of the body. A striking example is the assimilation of iodine by the thyroid gland. With the help of a radioactive iodine isotope it has been determined that, out of an extra amount of iodine administered in the food, after 1 or 2 days a healthy person has stored up in the thyroid gland about 3%, whereas a sufferer from goitre stores up 30% or more; see *fig. 3*. In certain cases of cancer

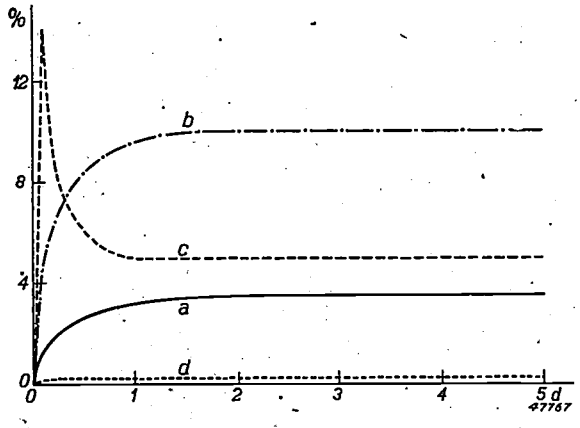


Fig. 3. The assimilation of iodine in the thyroid gland. Along the abscissa is the number of days which has elapsed after the administration of a known extra amount of iodine in food, along the ordinate the percentage of this amount found in the thyroid gland as measured by means of radioactive iodine;
a) for healthy persons,
b) with benign goitre,
c) with Basedow's disease,
d) upon defective functioning of the thyroid gland (myxoedema). (From J. G. Hamilton, Application of radioactive tracers to biology and medicine, J. appl. Phys, 12, 440-460, 1941).

of the thyroid gland the radioactive iodine was found to be accumulated not in the cancer tissue but in the healthy tissue. (Thus we again arrive at the problems referred to under "localisation".) Somewhat similar phenomena are found in the case of the assimilation of radioactive strontium in the blood and in the bones. It has been possible not only to determine how this process is retarded, for instance, by rockets and then promoted by the administration of vitamin *D*, and how it is even led in the opposite direction by Basedow's disease, but it has also been possible to study the finer distribution of the strontium assimilated: the strontium was found to accumulate in the hard bone tissue, and in the case of bone cancer it showed a preference for the cancer tissue and possible metastases thereof. Though straying from our subject, it should be pointed out that this last case may be of value to the doctor not only diagnostically but also therapeutically. Given a sufficiently high concen-

tration the radiation of a radioactive kind of atom has a destructive effect on the cancer tissue. If the radiating substance is selectively attracted by the cancer cells, as strontium by bone cancer, this may eventually serve as the basis of a very effective therapeutic treatment. (On the other hand, for physiological applications of radioactive isotopes as indicators it is a general rule that the destructive effect of the radiation must be avoided by keeping the concentrations of the radioactive isotopes sufficiently small. We shall return to this in the following article.)

We shall leave it at these examples. They are sufficient to give the reader an impression of what can be achieved with the tracer method in research work and routine investigation, a method for which uses are to be found in ever-increasing numbers and in even wider fields.

PENETRATION AND WELDING SPEED IN CONTACT ARC-WELDING

by P. C. van der WILLIGEN.

621.791.753.41

In a previous article a description was given of what contact arc-welding is, how the particular electrodes are made and what can be done by that method of welding. Here we will deal with several important advantages, both in welding technique and economy, which are obtained with the new type of electrodes. The shape of the penetration differs from the usual shape in that it is deeper in the middle and shallower at the sides, thus giving a favourable root penetration and avoiding undercut. Furthermore the welding speed, *i.e.* the amount of metal deposited per second, is greater.

Recently a new method of welding was described [1][2], so-called "contact arc welding", in which the electrodes can be kept resting on the workpiece from beginning to end, while the coating of the electrode is in continuous electrical contact with the workpiece. This new method was the outcome of an investigation carried out with the object, *i.a.*, of making touch-welding easier and of more universal application. With "contact electrodes" touch-welding can indeed be carried out easily without fear of the troublesome "freezing". This is due to the fact that the coating of the electrodes is made very thick by transferring a considerable part of the metal from the core wire to the coating. At the same time this makes the coating slightly conductive and this conductivity can be regulated in such a way that the electrode becomes "self-starting". This does away with the necessity of striking the arc by tapping and prevents extinction when welding with alternating current.

The goal which had been set at the beginning of the investigation was thereby achieved. In research work, however, it often happens that after attaining a given goal the conclusion is also reached that the solution found involves certain drawbacks preventing its application. In the development of contact arc-welding the reverse was true. In this case a number of further advantages appeared, while incidental disadvantages were found to be of little importance and easy to avoid. This will be illustrated by reference to the history of the electrode Contact 15. The facts, however, hold qualitatively for all types of contact-electrodes. Besides the electrode Contact 15, two other types were developed: Contact 18 and Contact 20, which were found in practice to be no less important than the first mentioned type. We shall first give a brief resumé of some of the characteristics of the three types of contact electrodes [3].

The different types of contact electrodes

The contact electrodes Contact 15, 18 and 20 were developed from the ordinary electrodes Ph. 55, 48 and 50, respectively, which, as will be known, differ mainly in the composition of the coating. The properties and field of application of the different types of contact electrodes are thus determined for a large part by those of the corresponding ordinary electrodes. In addition, of course, there are the specific properties connected with the new method of welding.

Thus, for example, Contact 15 and Ph. 55 have in common the high impact value [4]) and the insensibility to sulphur in the material to be welded [5]). Contact arc-welding can be carried out with Contact 15 in all positions except the vertical, where the free arc must be used [6]).

An electrode, the Contact 18, was therefore specially developed for vertical-down contact arc-welding. As basis the electrode Ph. 48 was taken, because this type is particularly well suited for vertical-down welding. Contact 18 can also be used quite well in other positions. Moreover, it gives unusually good results in under-water welding, the advantages of contact arc-welding, namely touch-welding and self-starting, being particularly valuable because under water the welder can see but very little, especially while actually welding, owing to the water around the arc becoming turbid. Contact 20 is an electrode which was developed exclusively for down-hand contact arc-welding. It is known

[1]) Philips Techn. Rev. 8, 161, 1946.

[2]) The Welding Journal, 25, 313-5, 1946.

[3]) Contact 15 and Contact 18 have already been discussed in the article referred to in footnote [1]).

[4]) See Philips Techn. Rev. 6, 97, 1941.

[5]) See Philips Techn. Rev. 7, 91, 1942.

[6]) The reason for this is explained on page 166 of the article referred to in footnote [1]).

that the welding speed, *i.e.* the amount of metal deposited per unit of time, of the type Ph. 50, which as already stated served as basis, is especially high. Since the higher welding speed, as we shall see, is characteristic of all types of contact electrodes, it will be clear that Contact 20 has an exceptionally high speed, since it was developed from the already fast type Ph. 50. This, in fact, constitutes the most important characteristic of the most recent of the different types of contact electrodes.

Shape of the penetration

A question which deserves particular attention in welding is the shape of the penetration. By this is meant that part of the workpiece that is fused during the welding. When a bead is welded on a flat plate the penetration resembles a segment of a circle, as shown diagrammatically in *fig. 1a* for an ordinary coated electrode.

In the first test with the heavy contact electrodes, however, it was already found that the shape of the penetration differed considerably from the normal, the penetration being deeper in the middle and shallower at the sides, *cf.* fig. 1*b*. This different shape of the penetration is characteristic of all contact electrodes.

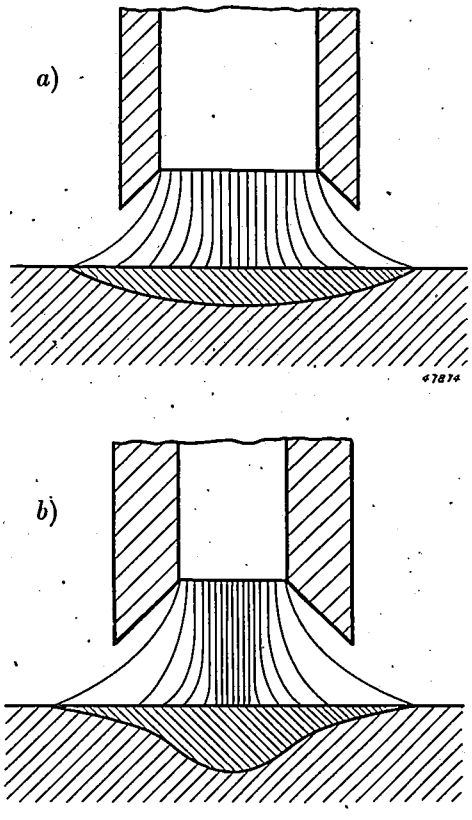


Fig. 1. The shape of the penetration: *a*) in the case of an ordinary electrode and *b*) in that of a contact electrode. The penetration is represented by the heavily shaded parts of the figure (for the sake of simplicity the bead is not drawn). It is also shown diagrammatically that the arc is more concentrated with contact electrodes than with ordinary ones, resulting in the difference in the shape of the penetration.

The explanation of this particular shape of the penetration is not difficult to find, considering that the arc burns only on the core wire; since in contact electrodes the core wire has only a relatively small surface, the arc is more concentrated than in the case of ordinary electrodes where the core wire has a relatively large surface (see figs. 1*a* and *b*). The spraying action of the welding arc, which is the result of the forces acting on the molten droplets [7]) and which, for example, makes overhead welding possible, will therefore, in the case of contact electrodes, also be more concentrated in the centre. Moreover, the longer cup of the contact electrode keeps the droplet better directed, resulting in less spatter. Furthermore, the great depth of penetration in the middle is undoubtedly due to the arc voltages of the thicker contact electrode being appreciably higher than those of the corresponding ordinary coated electrodes, containing the same amount of metal.

It is a known fact that electrodes having a high arc voltage give a deep penetration (such as the "high organic" type in the U.S.A.).

The relatively shallow penetration at the two sides can now be understood, for here the arc burns only more or less laterally.

The peculiar form of the penetration with contact electrodes is in very many cases of the greatest importance, for the following reasons. In the first place the deep penetration in the middle makes it possible with V-welds (and of course with fillet welds made in the flat position) to apply a first layer directly with a heavy contact electrode, say with a core diameter of 5 or 6 mm. It is unnecessary to deposit a first layer with an electrode of small diameter. The result is fewer layers and saving of time, as well as other advantages (*cf.* below on the subject of distortion).

It is to be noted that with a certain diameter of Contact 15 the arc voltage is closely dependent on the current used: with high current the arc voltage is highest. *Fig. 2* shows the relation between arc voltage and current for contact 15-5. (The number 5 indicates the diameter of the core wire in mm.) As was to be expected, the penetration is also found to depend very much on the current [8]).

[7]) As stated in the article referred to in footnote [1]), in special cases it is more advantageous to use Contact 15 with the free arc; this is also true for Contact 18.

[8]) See J. Sack, Philips Techn. Rev. 4, 9, 1939 and the Welding Industry, July 1939.

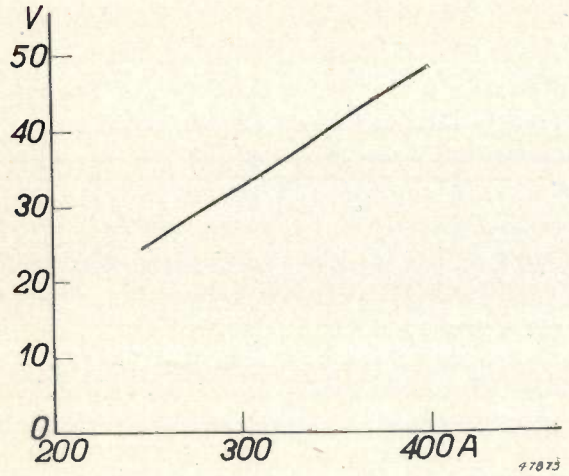


Fig. 2. Relation between arc voltage and current for Contact 15-5.

When a fillet weld is made in the flat position with two strips of 12 mm thickness, successively with different currents, and the depth of the penetration, the so-called root penetration (see text below *fig. 3*), is measured on the cross-sections, the result shown graphically in *fig. 4* is obtained. A similar graph could also be drawn for V welds; the relation between the current and the root penetration would be found to be qualitatively the same. It is therefore clear that when making, for instance, an open V weld with backing strip (see *fig. 5*) and using Contact 15-6 a high current must be used for the first layer in order to secure sufficient penetration to the bottom of the weld.

In the second place, due to the slight penetration at the sides, even when the heaviest contact electrodes and the highest currents are used,

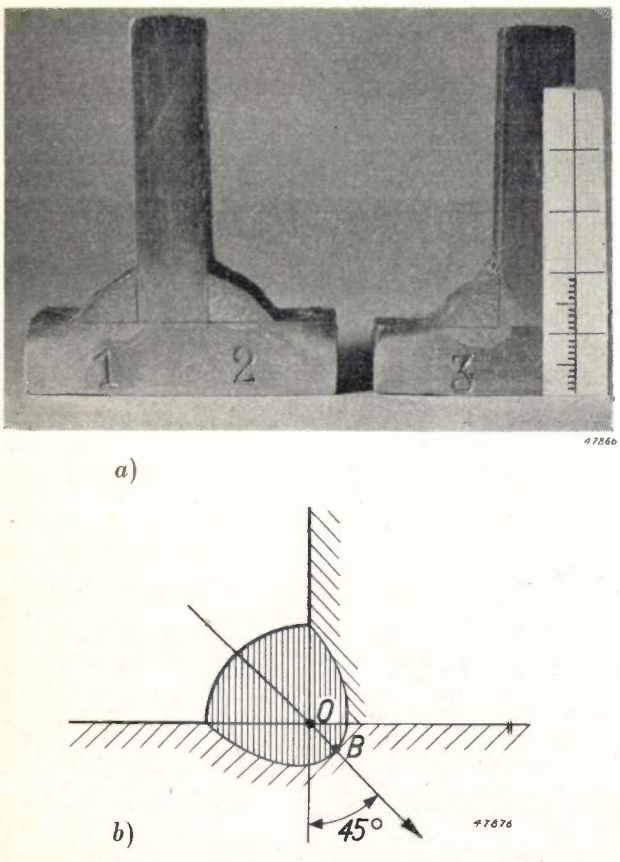


Fig. 3. *a*) In these cross-sections of fillet welds made in the flat position it is clearly seen that the depth of the penetration depends upon the current: the currents in the cases shown as 1, 2 and 3 amounted, respectively, to 375, 300 and 440 amp; the unit of the scale corresponds to 1 cm.

b) The strength of the weld is to a large extent determined by the so-called root penetration, which is defined as the distance *OB*. The root penetration is considered positive in the direction of the arrow. A negative root penetration means that a channel remains open in the corner under the bead.

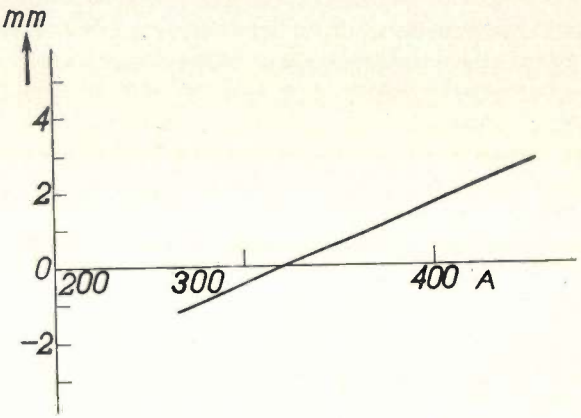


Fig. 4. The relation between the root-penetration and the current used with Contact 15-5, found from fig. 3.

no trouble is experienced from undercut, *i.e.* burning away of the original material at the edges of the bead. Undercut is one of the difficulties that had to be combatted with ordinary electrodes at high currents; undercut reduces the strength of the welded joint.

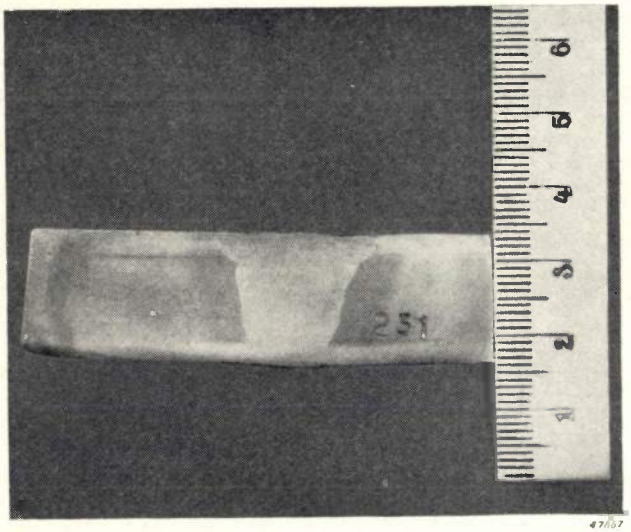
Fig. 6 shows fillet welds made in the flat position with Ph. 55-7 (No. 59) and with Contact 15-5 (No. 76), both with alternating current, 375 amp., 30 volts and 375 amp., 42 volts, respectively. The undercut can clearly be seen in No. 59 (under the 9 for instance). *Fig. 7* shows two cross-sections taken from these welds; with Contact 15-5 (No. 76d) the root penetration is about 1.5 mm deeper than with Ph. 55-7 (No. 59b).

To sum up, it may be stated that the shape of the penetration as found with the heavier contact electrodes offers important technical advantages compared with that obtained with ordinary coated electrodes.

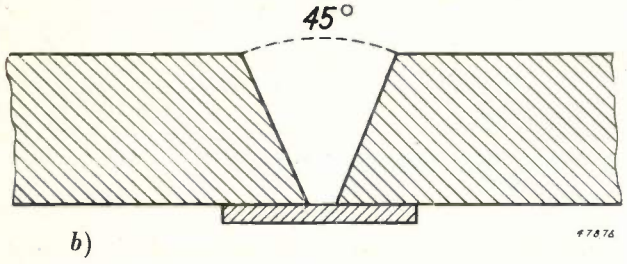
Welding speed

It has already been mentioned that the arc voltage of contact electrodes, especially those of

5 mm and heavier, is considerably higher than that of the corresponding ordinary electrodes. In the course of the investigation it was also found that the heat efficiency of the contact electrodes, *i.e.* the ratio



a)



b)

Fig. 5. *a*) Diagram of an open V weld. In the case represented in the photograph (see *b*), the angle was 45° and the opening, *i.e.* the smallest distance between the two plates to be joined, was 4 mm.

b) Cross-section of an open V weld made in two layers between two plates 17 mm thick. For the first layer Contact 15-6 was used with 520 Amp. A.C.; for the second layer Contact 15-6 with 490 Amp. The backing strip is removed after welding.

of heat expended in the fusing of the metal to the total amount of heat supplied, is higher. This is due to the deep cup enveloping a larger part of the arc than is the case with ordinary electrodes used with the free arc. The heat in the arc is thus used to greater advantage and the radiaton of the arc into space is diminished. As a result of the two facts mentioned, the welding speed, *i.e.* the amount of metal deposited per second, is considerably higher with the contact electrodes.

Averaged over a large number of tests it was found, for example, that Contact 15-5 is about 50% faster than Ph. 55-7. These diameters of the two types are particularly suitable for comparison because they contain equal amounts of iron and slag-forming substances [9]), and because incidentally the maximum current is the same in both cases, namely about 375 amp.

Table I gives the average results:

Table I

Electrode	Contact 15-5	Ph. 55-7
Current in amp.	375	375
Arc voltage	42	30
Milligrams iron per sec.	1700	1150
Milligr. iron per amp. sec.	4.5	3.1

In addition to this greater speed of Contact 15 compared with an ordinary coated Ph. 55 of the same weight, there is also the possibility, already mentioned in the discussion of the shape of the penetration, of using a heavier contact 15 in many cases where a lighter type of an ordinary coated electrode would be required. This also means greater welding speed, since with a heavier electrode more iron is deposited per second. Therefore in such cases this gain in speed must be added to the above 50%.

Several further particulars

Distortion

The distortion or warping of a weld may have unpleasant results in all welded structures and the avoidance of this demands much care and expert knowledge. It is therefore of great importance that in welding with contact electrodes there is little distortion. This fact is very closely connected with the high welding speed and the favourable shape of the penetration of these electrodes. It is known that distortion increases as the welding speed decreases and also increases with the number of layers in the weld. It has already been explained that in welding with contact electrodes, due to the special shape of the penetration, fewer layers are needed than in the case of ordinary electrodes.

Heating of the electrodes

In many cases, and especially with the lighter electrodes [10]), the highest permissible current is

[9]) At low currents when the arc voltage is not higher than with Ph 55, a smaller penetration might even be expected with Contact 15.

[10]) With heavier types the highest current is determined by other factors, for instance by the desired uniformity of the surface of the weld, which becomes less at higher currents (convex, irregular appearance).

determined by the fact that when using up the last piece of the electrode the top end becomes red hot and this is apt to cause the rod to bend. Since it might, therefore, be expected that the highest permissible current for contact electrodes would be much lower than that for ordinary ones, but

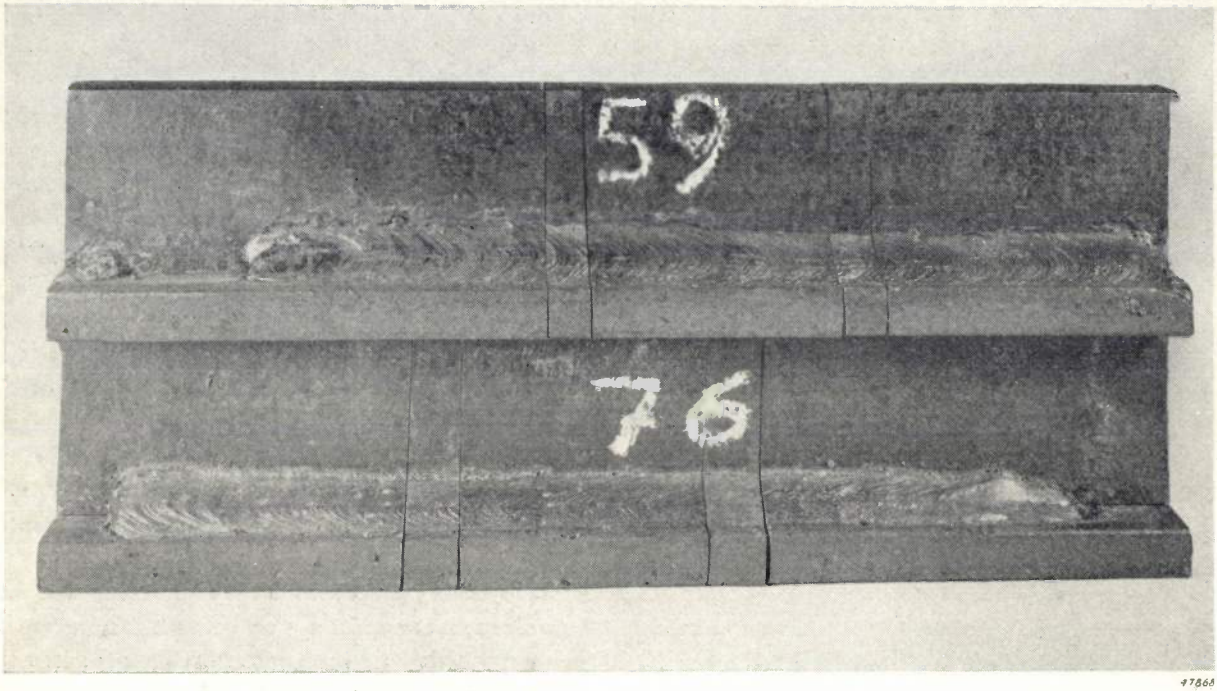


Fig. 6. Photographs of fillet welds made in the flat position with Ph 55-7 (No. 59) and with Contact 15-5 (No. 76). In No. 59 the undercut can be seen under the figure 9, while No.76 shows no faults at all.

the cores of contact electrodes are much thinner than those of the corresponding ordinary electrodes actually there is not such a great difference. Thanks to the fact that the coating of the contact electrodes is heavy and conductive, it dissipates the heat much better than the coating of the ordinary electrodes, thus compensating for a large part the effect of the smaller thickness of the core wire of the contact electrodes. Whereas with ordinary electrodes it is of only secondary importance whether or not the coating makes good contact with the holder, in the case of contact electrodes it is of great importance that the coating should make good contact with the holder, because otherwise the upper end of the core wire would become red hot much sooner than in the case of an ordinary electrode. The holder should have a reasonably high heat capacity and conductivity; the heads of holders specially made for contact electrodes are of copper and constructed in such a way as to fulfil this requirement.

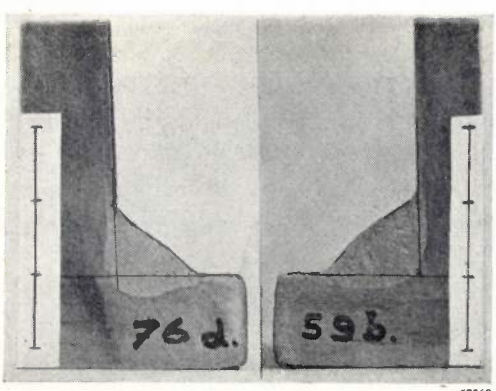
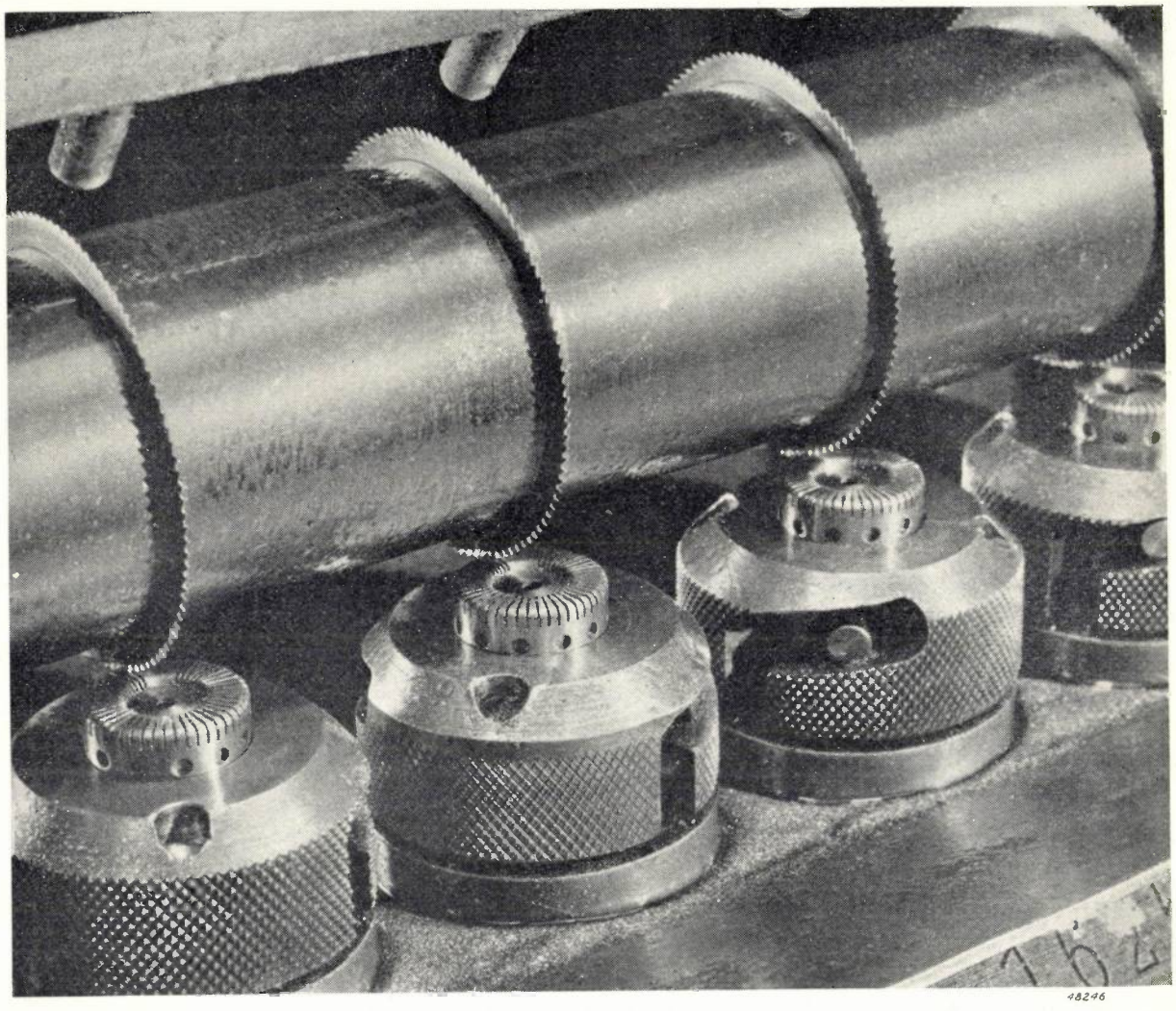


Fig. 7. Cross-sections of the fillet welds of fig. 6. In 76d, which was welded with Contact 15-5, the root penetration is found to be about 1.5 mm deeper. The unit of the scale corresponds to 1 cm.

MASS PRODUCTION OF ELECTRICAL SHAVERS



This photograph shows how the 48 slots are cut in the head of the "Philishave" dry-shaver. A series of circular saws are clamped at equal distances around a rotating shaft and underneath each saw is a special kind of turntable on which a shaving head is placed. These tables are carried on a common slide moved in the direction of the saws. As soon as the slide has passed underneath the saws and a slot has been cut in each head, the slide returns and all the tables automatically turn exactly over the forty eighth part of a circle, ready for the next slot to be cut.

AUTOMATIC CHANGE-OVER TO AN EMERGENCY APPARATUS IN A COMMUNICATION SYSTEM

by G. HEPP.

621.395.365.3

In the case of a breakdown it should be possible for components of a communication system, such as amplifiers and oscillators, to be replaced automatically by analogous emergency components without interruption in the service. In particular the impulse generator, from which the carrier waves of a carrier telephony system are derived, has to satisfy very high requirements as regards safety precautions. Here two methods are described that have been developed by Philips for automatic changeover to an emergency component as soon as the amplitude of the output signal falls below a certain value. One of these methods is employed in a carrier telephony system already described in this periodical. Where low-frequency amplifiers are concerned, which — in contrast to oscillators — produce an output signal varying considerably in amplitude, a constant auxiliary signal is added which has a frequency outside the band of the signal to be amplified. It is this auxiliary signal which brings about the change-over to the emergency amplifier when the intensity becomes too low.

Introduction

In the present stage of electrical communications technique — whether it is a question of an "ordinary" telephone connection or one with carrier system, a cable or a radio connection — the message transmitted passes through numerous electrical apparatus: lines, transformers, amplifiers, etc., and the development of this technique is undoubtedly moving in the direction of greater complexity of the whole system; witness the carrier-wave telephony repeatedly discussed in this periodical. With this greater complexity there is, of course, a greater chance of failures due to defects, which are never entirely avoidable. On the other hand, and rightly, in communications technique very high requirements are set as to reliability requirements which not only call for the utmost care in manufacture, testing and maintenance, but which, moreover, make it necessary to apply measures for limiting failures, which in spite of everything are still apt to occur occasionally, to the shortest possible duration — preferably so short as to pass unnoticed by the persons who are connected. It is usually not sufficient to install an alarm arrangement to warn the operators in case of emergency. Often a certain component will have to be duplicated and a device installed for automatically changing over to the emergency part as soon as a failure occurs. This change-over must be effected with the minimum number of extra parts, because they too may be liable to breakdown. Further requirements which such an arrangement must satisfy will be discussed farther on.

Obviously it will depend upon the importance of a component how far these safety measures will be carried. A typical example of a component which in case of a breakdown — if no emergency reserve is provided — would lead to the simultaneous interruption of a number of connections is the impulse generator, recently discussed in this periodical, which furnishes a carrier system with carrier waves [1]). It is essential that there should be an emergency generator to be automatically switched on without interruption in case of breakdown, while of course the operators have to be warned too. Such an automatic transfer is especially necessary for apparatus located where there is no operating personnel, as is the case, for example, with most repeater stations, which are located at regular intervals along a telephone link.

We shall now discuss in turn some of the requirements which such an arrangement has to satisfy and examine how this can be achieved. First of all, however, let us define the term "failure" as signifying a fall in the output voltage to below a certain limit (for instance 80% of the rated value). For the present we shall consider only those components which, like a generator, give an approximately constant voltage when functioning normally. Later we shall also discuss components where this is not the case (for instance low-frequency amplifiers).

First requirement: metastability

If the output voltage falls below the critical value the emergency generator must be switched in, but if that voltage subsequently rises again to above that level the regular generator must still remain out of service until the defect has been repaired and it has been restored to action by the operating personnel. If this requirement were not fulfilled then every time a bad contact is made or a contact broken there would be a change-over from one generator to the other, which would cause a troublesome clicking in the telephones. The possibility of such an unnecessary change-over taking place is not confined to the case of a loose contact, for it can well be imagined that due to ageing of the valves the output voltage may fall to round about

[1]) The Excitation of the Carrier Waves in an Installation for Carrier Telephony, Philips Techn. Rev. 8, 141, 1946.

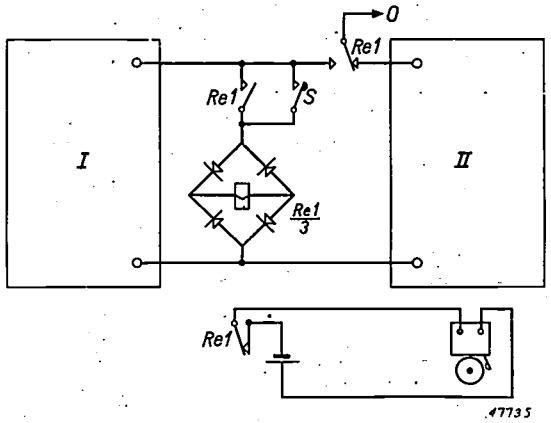


Fig. 1. Automatic change-over and alarm by means of a relay *Re*1/3, see footnote [3]). After pressing the starting key *S* the coil of the relay is excited by the rectified voltage from the generator *I*. This causes the three relay contacts to change from the position indicated to the opposite one, so that
1) the load *O* is connected with generator *I*,
2) the key *S* is shunted, so that after *S* is released the relay coil continues to be excited,
3) the alarm arrangement is switched off.

When generator *I* fails, the relay opens and the contacts return to the positions drawn in the figure, so that
1) the load *O* is connected with the emergency generator *II*,
2) the key is no longer shunted.
3) the alarm comes into action.

If generator *I* recovers, the conditions last mentioned are maintained (metastability) until the key *S* is pressed down again.

the critical value and owing to fluctuations in the mains voltage sometimes be above and sometimes below it.

A circuit which satisfies the requirement first formulated is said to be metastable. *Fig. 1* is an example of how this can be realized with the help of a relay [2]); its functioning is explained in the text below the illustration.

Further requirements

A second requirement is that the transfer apparatus must not use too much energy from the generator. The excitation of a reliable relay requires more power than can be spared for that purpose by most generators, and especially by those used for carrier telephony. An amplifier valve might be used in front of the relay, but this is apt to conflict with the requirement, mentioned at the beginning, that there should be the minimum number of parts subject to breakdown themselves.

To avoid these difficulties two circuits were designed in which one or more of the valves of the generator also fulfils the function of D.C. amplifier for the relay. These circuits differ in the fact that in the one case the combination of the two generators is mestastable, and in the other each separate generator is metastable. By way of example the two circuits will be discussed as applied to simple oscillators.

Circuit in which the combination is metastable

Fig. 2 shows a circuit in which one oscillator furnishes a voltage which, when rectified, blocks the other [3]). Only one of the two oscillators can function at one time, and if that one fails, the other comes automatically into action without the use of a relay. An alarm relay will of course be used (not drawn in fig. 2).

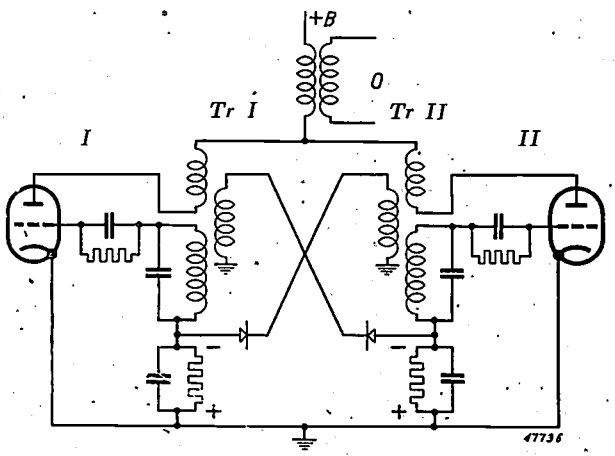


Fig. 2. Circuit in which one oscillator (*I*) excites in an extracoil of the transformer *Tr I* a voltage which, being rectified, blocks the oscillator *II* and inversely. With this arrangement no relays are needed to transfer the load *O*. Relays for alarm signals are omitted for the sake of simplicity. At $+B$ anode voltage is connected.

The circuit of fig. 2 still has the drawback that in the event of the anode being short-circuited with earth or one of the other electrodes neither of the two oscillators can work. This objection can be met by replacing the output transformer by two somewhat more complicated transformers, as indicated in *fig. 3* (*TrA* and *TrB*). If we assume that oscillator *I* is in action, its anode current flows through the two coils *1* of *TrA* and *TrB*, respectively, and induces voltages in all the other coils. These are so connected that the voltages in coils *2* cancel each other, so that in the anode circuit of *II* no voltage is induced (a possible short-circuiting of anode *II* to earth does not then affect the energy furnished by *I*);

the voltages in coils *3* act together so that there is a result ant voltage which blocks *II*;

the voltages in coils *4* cancel each other, so that *I* does not block itself.

Owing to the symmetry of the circuit the same applies, *mutatis mutandis*, when *II* and not *I* is the oscillator in function.

[2]) In the diagrams of this article where relay coils are denoted, for instance, by *Re* 2/4, the first number is the number of the relay in question and the second the number of contacts it has. Each relay contact is indicated by *Re* followed by the number of the relay to which the contact belongs. According to custom the contacts are drawn in the position occupied when the coil carries no current.

[3]) U.S. Patent 2.319.320. This circuit is related to the circuit known as "Kallirotron", in which, however, the D.C. voltage that blocks one valve is taken directly from the other without the intermediate stage of a rectified A.C. voltage; the object in this was to produce a negative resistance.

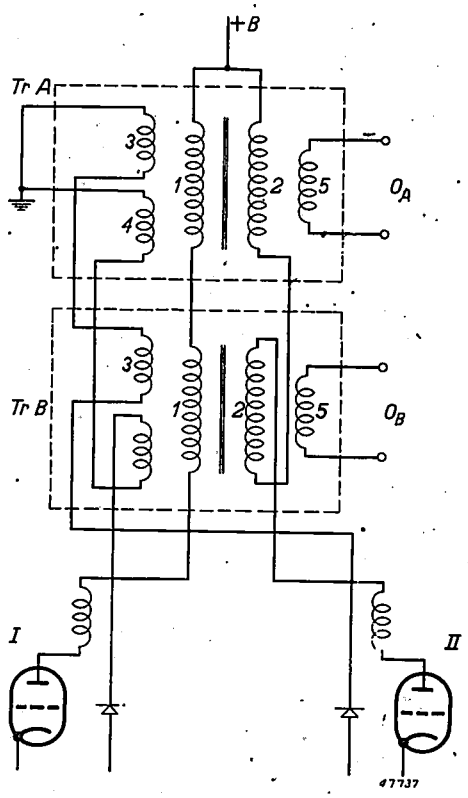


Fig. 3. Extension of the circuit of fig. 2 to include two push-pull output transformers (*Tr A* and *Tr B*), so that when an anode is short-circuited to earth one of the oscillators continues to function. The load must be divided equally between the two coils 5 (O_A, O_B).

The two coils 5 are the output coils, each of which feeds half of the load. The condition for good functioning is that these loads shall be equal, since otherwise the balance is upset.

Circuit in which each of the oscillators is metastable

Fig. 4 shows a circuit [4]) in which initially each of the oscillators is kept blocked by the negative grid voltage from a battery *C*. This blocking voltage can, however, be neutralized by an equal but oppositely directed voltage *E*. The voltage *E* for each of the oscillators is obtained by rectification of an A.C. voltage induced in an auxiliary winding on their output transformer (*TrI* and *TrII*, respectively). Once it is brought into action, therefore, the oscillator continues to generate, but it cannot begin of itself, so that the requirement of metastability is met. In order to start, the key S_I or S_{II} is pressed down, temporarily cutting out the negative grid voltage, so that the valve begins to generate, and when the voltage *E* has been reached the key can be released without any change occurring. Normally both the oscillators function, but only the voltage generated by one of them is used.

In the anode circuits there are the relays *Re1/2* and *Re2/1*, each of which operates an alarm contact, while the former also operates the switching-over of the load *O*. The relay coils are excited by the anode direct current. The energy taken from the output transformer to set up the voltage *E* is naturally much less than the energy necessary to excite the relay.

The relay *Re2/1* can be omitted if necessary, but its retention has the advantage that a warning is also given when there is a failure in the emergency oscillator *II* while it is in a state of oscillation but not in action, thus considerably enhancing the reliability of the emergency unit. In the circuit of fig. 2 such a possibility is entirely absent. In cases where absolute dependability is essential, therefore, the circuit of fig. 4 will be preferred to that of fig. 2 (and the use of relays will have to be accepted). If the requirements are less strict the simpler method of fig. 2 will receive first consideration.

Temporary interruption of metastability; automatic starting

It may happen that both oscillators fail simultaneously due, for instance, to a short interruption of one of the feeding voltages — or that they fail one shortly after the other, without any serious defect in either. The service would then be completely at a standstill. In such an event metastability plays us false and it is therefore necessary to try

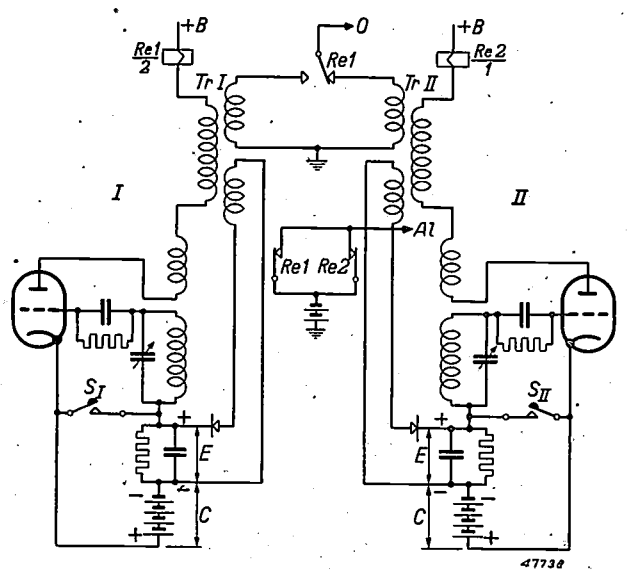


Fig. 4. Circuit in which each of the identical oscillators *I* and *II* excites its own grid voltage *E* equal to the fixed negative grid bias *C* but of opposite polarity: Consequently each of the oscillators is metastable, *i.e.* they cannot be brought into action without pressing the starting keys S_I or S_{II}. The relays *Re* 1/2 and *Re* 2/1 switch over the load *O* or respectively bring an alarm into action. $+B$ = positive pole of the anode battery, *Tr I*, *Tr II* = output transformers, *Al* = alarm arrangement.

[4]) U.S. Patent. 2. 330. 582.

to neutralize it when one of the oscillators fails.

This can be done in the manner indicated in *fig. 5a*, namely by shunting the starting key S_I by a break-contact of *Re 2/1* (and conversely by shunting S_{II} by a break-contact of *Re 1/2*). After a temporary disturbance oscillator *I* can then resume action without outside help.

A slight addition, however, is still desirable. If, while oscillator *I* is in action, "chattering" occurs in *II*, the relay contact *Re 2* in parallel with S_I would continually open and close. This would be harmless if the voltages *C* and *E* were exactly equal, but this is only approximately so (due in part to the fact that *E* decreases as the valves age). The adjustment of the valve *I* therefore changes somewhat as the relay contact opens or closes, resulting in undesired fluctuations in the oscillator voltage.

This can be avoided by connecting in series with the above-mentioned relay contact another break-contact of relay *Re 1/2* (*fig. 5b*). This is open as long as oscillator *I* is in action, so that the opening or closing of the contact *Re 2* in series with it has no effect. Conversely, the other starting key is shunted by two break-contacts in series, one of relay *Re 1/2*

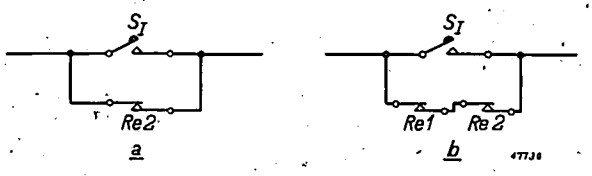


Fig. 5. *a*) The starting key S_I of oscillator *I* being shunted by a break-contact *Re* 2, oscillator *I* (when oscillator *II* has failed), can of itself come into action again after a momentary failure. Conversely S_{II} is also shunted by a break-contact *Re* 1.

b) When oscillator *I* is functioning normally, a faulty contact in oscillator *II* would cause relay contact *Re* 2 to open and close (fig. 5*a*), which might have undesirable consequences. In order to avoid this, a break-contact *Re* 1 is connected in series with *Re* 2. S_{II} is also shunted by a connection in series of two break-contacts *Re* 1 and *Re* 2.

and one of relay *Re 2/1*. The two pairs of contacts are closed when both oscillators are out of action; these are thus given the opportunity of beginning to oscillate anew (as far as they are in a condition to do so), without it being necessary to press the starting key. This is what is called automatic starting.

Practical example

In a carrier-wave system designed by Philips automatic change-over of the impulse generators

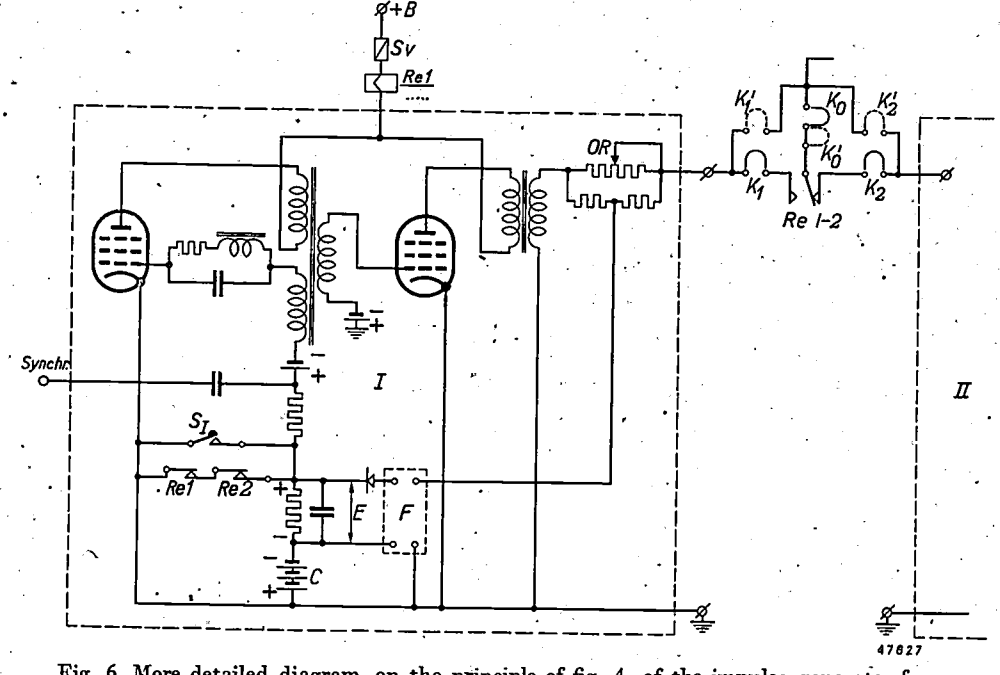


Fig. 6. More detailed diagram, on the principle of fig. 4, of the impulse generator for a telephony system with carrier waves.
F = filtcr tuned to the fundamental frequency (4Kc/Sec), *Synchr.* = connection for the synchronizing voltage, *OR* = signal volume control, *Sv* = Safety fuse, *Re* 1-2 = relay contact connecting the load cither with generator *I* or with generator *II*. It is indicated in this way because the relays *Re*1/2 and *Re*2/1 of fig. 4 can be exchanged at will by a switching arrangement (not shown). In this way one is free in the choice of which generator will be considered the regular one. This change-over is done by means of the control key of fig. 7. K_0, K_1, K_2 = plug connections to make the installation independant of the contact *Re* 1-2, If this checking has to take place during the operating of Osc. *I* the plug is taken from K_2 and placed in the position K_1'. Subsequently the plug *K* in the position K_0 is brought into the position K_0' and finally the conection between *Re* 1-2 and the Osc. *I* is broken by removing the plug from *K*1 and spring it temporarily in the position K_2.

is employed on the principle of fig. 4. The circuit used is given in somewhat more detail [5]) in *fig. 6*. The grid voltages C and E referred to in the foregoing may be distinguished, and also the starting key S_1 shunted by two relay contacts according to fig. 5*b*. The filter (F), however, is new; the voltage E with which the generator unblocks itself should only come through when the desired signal is produced and not, for example, a "hum" or some chance oscillation. To this end a filter or circuit (F), tuned to the fundamental frequency of the A.C. voltage to be provided (here 4 kc/sec), is placed in front of the rectifier cell which furnishes the voltage E.

Fig. 7 shows the key panel. The function of some of the keys is explained in the text below the figure.

Automatic change-over of audio-frequency amplifiers

The methods of automatic change-over described, both those of fig. 2 and those of fig. 4, are based on the rectification of an A.C. voltage proportional to the signal emitted and the use of this rectified voltage either to block the other oscillator (fig. 2) or to deblock its own oscillator (fig. 4). This works satisfactorily as long as it is applied to an apparatus that gives a reasonably constant signal. In the case, however, of an amplifier for speech or music, for instance, the output voltage is most variable and at intervals even zero. If either of the methods described were applied to such an amplifier, then the intervals and even the soft passages would cause the transfer mechanism to act as if there were a defect, thus leading quite unnecessarily to repeated change-over.

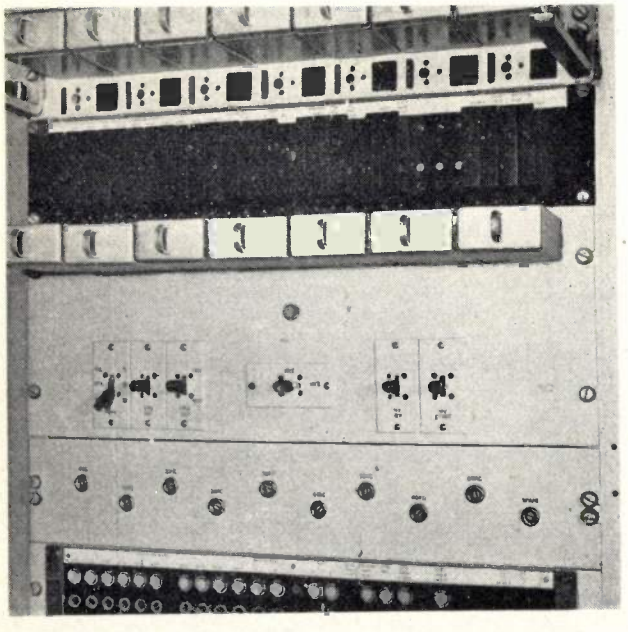


Fig. 7. Key panel for carrier wave feeding with automatic transfer according to fig. 6. The central key is used to determine which generator will be the regular one and which the emergency. The key at the extreme right is the starting key (S_I or S_{II}, according to the position of the central key). The key second from the right, in the position "M.V. Repair" during repair of the emergency generator, also switches off the alarm of the latter.

[5]) This diagram is at the same time an elaboration of figs. 2 and 5 of Philips Techn. Rev. 8, 140 and 143, 1946.

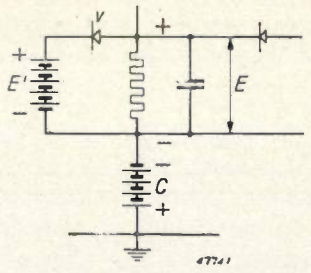


Fig. 8. When the principle of fig. 4 is applied to low-frequency amplifiers the blocking voltage E has to be limited (upward) to maintain the correct operating conditions of the amplifier valve. This limitation can be achieved by employing a battery with voltage E' in series with a rectifier cell v which passes current when E becomes higher than E'.

This can be avoided in the following way. In addition to the low-frequency signal to be amplified, a constant auxiliary signal of a frequency lying outside (usually higher than) the band to be amplified is laid on to the input of the amplifier. This auxiliary signal is always present and gives at the output the necessary blocking or deblocking voltage for the amplifier valve or valves. A filter F as in fig. 6, tuned to the frequency of the auxiliary signal, is of course indispensable.

If this is applied in a circuit like that of fig. 4, it is necessary to introduce a limitation of the voltage E, because in contrast to the case with the impulse generator with low-frequency amplifiers only a small part of the valve characteristic is used, so that the operating conditions must remain constant within narrow limits.

If for some reason or other the voltage E should become too high, the valve will function in a region with a larger mutual conductance and the output signal will also become greater, which causes E to increase again, and so on. In order to prevent this one can connect in parallel with the resistance and condenser on which the voltage E acts (*fig. 8*) a battery of the desired voltage E, in series with a rectifier cell v which passes current when E becomes higher than E'.

By this means E is limited to the value E'. If, due to a defect, E falls below the value E', the cell v keeps the battery switched off as it were, so that E can drop unhindered and finally the amplifier in question is cut out. Under normal conditions, however, E is practically equal to E', so that the operating conditions of the amplifier valve are confined within narrow limits.

NON-FERROUS COPPER WIRE FOR MOVING-COIL METERS

by P. G. MOEREL and A. RADEMAKERS. 621.317.715.004.64:538.22

The material of which the rotating system of moving-coil meters is made may contain particles of iron. The ferromagnetism of the latter, in combination with the express or accidental inhomogeniety of the magnetic field, causes certain defects in the measuring instrument, such as a non-reproducible zero point or, in the case of coulometers too large, a residual couple. Measures are discussed particularly for keeping the copper wire used for the moving coil as free of iron as possible, and some results thereby obtained are mentioned.

Moving-coil meter without directional couple

As far back as the beginning of this century it was known that a magnetic flux or a quantity of electricity ($\int i dt$) can be measured with the help of a moving-coil meter provided it satisfies certain requirements. Such an instrument is the fluxmeter of Grassot [1]).

The main requirement to be satisfied, contrary to what is demanded of an ordinary moving-coil meter, is an extremely small directional couple.

This follows from the equation of motion of the moving system:

$$J\frac{d^2\varphi}{dt^2} + a\frac{d\varphi}{dt} + D\varphi = Ai \quad \ldots\ldots\ldots (1)$$

where J represents the mass moment of inertia with respect to the axis of rotation, φ the angle of rotation, t the time, a the damping constant, D the directional couple at $\varphi = 1$, A a proportionality factor and i the current in the rotating coil. This equation expresses the fact that the external couple (Ai) is at equilibrium with the resultant of the couple of the mass force, the damping couple and the directional couple.

If the instrument possesses a sufficiently small inertia and directional couple, such that the first and third terms of the left-hand member of (1) are negligible compared with the second term, there remains:

$$a\frac{d\varphi}{dt} = Ai \quad \ldots\ldots\ldots (2)$$

Or integrated:

$$\varphi_1 - \varphi_0 = \frac{A}{a}\int_{t_0}^{t_1} i\,dt = \frac{A}{a}Q \quad \ldots\ldots\ldots (3)$$

The difference between the readings at the moments t_1 and t_0 is thus a measure of the charge Q which has flowed in that interval of time.

In the complete absence of directional couple and inertia the coil only moves while current is flowing through, but has no definite zero position. If the moments t_0 and t_1 of (3) are chosen before the beginning and after the ending of the passage of current (which will naturally be the case when dealing with a current impulse), φ_0 and φ_1 are read when the pointer is stationary. This is one of the advantages of this method compared with the ballistic method, where the largest deviation has to be read.

The current i in the right-hand member of (2) may be an impulse generated by the change in flux in a coil to which the instrument is connected. Then

$$i = \frac{1}{R}\cdot n\frac{d\Phi}{dt}$$

(R = total resistance of the circuit, n = number of windings of the coil, each of which envelops the flux Φ), so that the solution of (2) becomes

$$\varphi_1 - \varphi_0 = \frac{A}{a}\cdot\frac{n}{R}(\Phi_1 - \Phi_0) \quad \ldots\ldots\ldots (4)$$

If the change in flux is obtained, for instance, by sliding the coil off a magnet under examination, then Φ_0 is the flux to be determined and Φ_1 is practically zero, so that here again the change in reading is a measure of the quantity desired.

The directional couple is kept small in the first place by entirely omitting the spiral springs which cause this couple in ordinary moving-coil meters, and further by leaving the connection wires to the moving-coil limp and balancing the moving system as truly as possible. It has, however, long been known that often a small directional couple still remains [2]), as a result of which the pointer does not become stationary after the conclusion of the current impulse. This "parasitic" directional couple is to be ascribed to a combination of two causes, *viz.* the presence of ferro-magnetic impurities [3]) in the materials of which the moving coil is made, in combination with the fact that the magnetic field in which the coil moves is not entirely homogeneous (although that is usually desired). If the coil contains particles of iron large enough to exhibit ferromagnetism [4]), those particles are attracted

[1]) E. Grassot, Fluxmètre, J. Phys. 3, 696, 1904.

[2]) H. Busch, Das Kriecbgalvanometer, Zt.f. techn. Physik, 7, 369, 1926.

[3]) See for example F. W. Constant, Ferromagnetic Impurities in Metals, Rev. Mod. Phys. 17, 81, 1945, where further references to literature are also given.

[4]) As is known, ferromagnetism is based upon a reciprocal action between iron atoms (or nickel or cobalt atoms). It does not occur in the case of iron atoms in solution among other atoms, at relatively large distances from each other.

to the spots where the field strength is greatest, so that a couple acts on the rotating system and consequently when it is carrying no current it still shows a certain preference for a definite position.

Parasitic directional couple caused by ferro-magnetic impurities

This is illustrated by *fig. 1*. Even when the iron core between the pole pieces is perfectly centred the field in the air gap is not absolutely homogeneous, because of the spreading of the lines of force at the edges of the pole pieces; it is strongest on the *N-S* axis. A ferromagnetic particle of iron at point *O* becomes a (practically) radially directed magnetic dipole and will experience a force which tends to move it towards the position of greatest field strength, thus in the direction of the *y* axis indicated in fig. 1. This force leads to a couple which is found to be proportional (see below) to the imhomogeniety of the field, which is expressed by $\partial H_r/\delta\varphi$. Added together over all the ferromagnetic particles of iron situated in the magnetic field, whose combined mass we shall call *G*, this leads to a couple

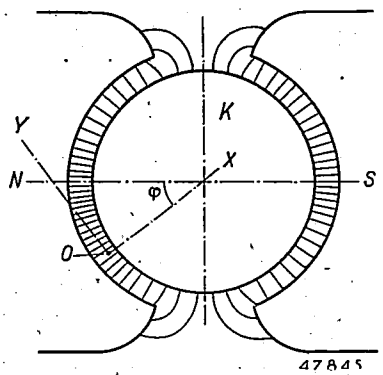


Fig. 1. Not entirely homogeneous field distribution between the iron core *K* and the pole piece *N-S* of a moving-coil meter.

$$K_{Fe} = CG \cdot \frac{\partial H_r}{\partial \varphi} \quad . \quad . \quad . \quad . \quad . \quad (5)$$

acting on the rotating system, where *C* is a constant.

For the component in the *Y* direction of the force *F* which a magnetic field *H* (with components H_x, H_y, H_z) exerts on a dipole with the magnetic moment *M*, the following applies:

$$F_y = M_x \frac{\partial H_y}{\partial x} + M_y \frac{\partial H_y}{\partial y} + M_z \frac{\partial H_y}{\partial z}.$$

In the case in question where the dipole is (practically) radially directed, $M = M_x$ and $M_y = M_z = 0$, thus,

$$F_y = M \frac{\partial H_y}{\partial x}.$$

The components F_x and F_z in the first instance have no effect on the deflection of the measuring instrument and therefore are not considered here.

In the absence of electric current rot $\vec{H} = 0$; in another form this is expressed by three equations, one of which is as follows:

$$\frac{\partial H_y}{\partial z} = \frac{\partial H_z}{\partial y},$$

so that the following may be written for F_y:

$$F_y = M \frac{\partial H_x}{\partial y};$$

thus on the moving coil of diameter *d* a couple *k* acts of the following magnitude:

$$k = \frac{d}{2} M \frac{\partial H_x}{\partial y} = M \frac{\partial H_r}{\partial \varphi} \quad . \quad . \quad . \quad . \quad . \quad (6)$$

when φ is the angle indicated in fig. 1.

In order to arrive at the total couple K_{Fe} exerted on the moving system due to the presence of particles of iron, (6) must be totalled up for all the particles concerned. Assuming that, roughly speaking, all these particles are situated in an equally strong field, with the same inhomogeniety, and that they are similar in shape, the total couple is found to be proportional to their combined mass *G*. If the other factors are included in a single constant *C*, the following is indeed obtained:

$$K_{Fe} = C \cdot G \cdot \frac{\partial H_r}{\partial \varphi} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

The effect in question is thus dependent on two factors which one can try to influence; the amount of iron impurities in the coil and the inhomogeniety of the field (G and $\partial H_r/\partial\varphi$ in (5)).

Fig. 2a shows diagrammatically the various couples which acted in a given case on the moving system of a (currentless) millicoulomb meter: K_{Fe} the couple due to the presence of iron in the not particularly pure coil, K_B the maximum

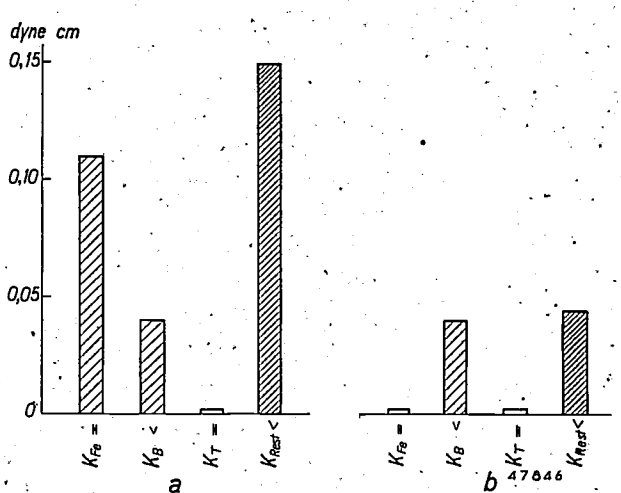


Fig. 2. Couples which act on the moving system of a currentless millicoulomb meter, *a*) for a moving system contaminated with iron. *b*) for one practically free of iron. K_{Fe} = iron couple, K_B maximum balancing couple, K_T = torsion couple, K_{rest} = resultant of K_{Fe}, K_B and K_T. (It is assumed that these three couples act in the same direction, *i.e.* in the most unfavourable case).

couple which remains due to unavoidable imperfections in the balancing, K_T the torsion couple of the current supply wires. Together they form the residual couple K_{min} which, if it is larger than the frictional couple, makes the system move. *Fig.2b* shows these couples for the same instrument with materials which have carefully been kept free of iron. It may be seen that the residual couple of max. 0.15 has been reduced to about max. 0.045 dyne cm.

Moving-coil meters with directional couple

Disturbed linear scale in galvanometers

The parasitic directional couple may be disturbing not only in the above-mentioned flux-meters and coulo-meters, but also in instruments with intended directional couple: it disturbs the linearity of the scale, owing to the total directional couple no longer being truly proportional to the deflection of the instrument. This phenomenon, it is true, is not noticeable in the usual sort of technical moving coil meters where the directional couple of the spiral springs is appreciably larger than the parasitic couple — even when no special attention is paid to the purity of the materials used for the coil; in the case of sensitive moving-coil galvanometers, however, with their so much weaker directional couple (torsion wire), deviations from linearity may occur, and it is just with these instruments that great value is often attached to a linear scale.

Varying zero position in meters with a logarithmic scale

There is still another unpleasant phenomenon which must also be ascribed to iron particles in the coil, namely a changing zero position. By this is meant the failure of the pointer of a measuring instrument to return exactly to zero, although it has previously been set accurately at zero; the difference from the true zero depends upon the magnitude of the deflection that the instrument has just recorded. If upon rotation of the coil not only the magnitude but also the direction of the field with respect to the coil changes, upon return to the neighbourhood of the zero position the ferromagnetic impurities will exhibit a remanent magnetization the direction of which depends upon the magnitude of the deflection. A directional couple thus remains which causes the so-called non-reproducible zero point. It is obvious that this phenomenon will also be stronger according as the field is less homogeneous.

Although, as a rule, a linear scale is desired in moving-coil meters, for special applications instruments are required with a non-linear scale, and in particular a logarithmic one, for instance meters calibrated in decibels as used for acoustic investigations and in communication technology [5]); further, also photographic exposure meters (with a photocell). In these cases the sensitivity of the meter should decrease as the deflection of the meter increases. This can be obtained by giving the magnetic circuit such dimensions that the field in which the coil moves changes according to a logarithmic law. In this case, therefore, the field is very inhomogeneous, resulting in a great variation of the zero point when the moving system contains ferromagnetic particles.

Relative determination of the iron content

From the above it follows that in the construction of the special kinds of moving-coil meters mentioned care must be taken that the materials of which the moving system is made are free of iron. This applies to the aluminium frame on which the coil is wound, and even for the small weights which balance it (*fig. 3*), although the latter are outside

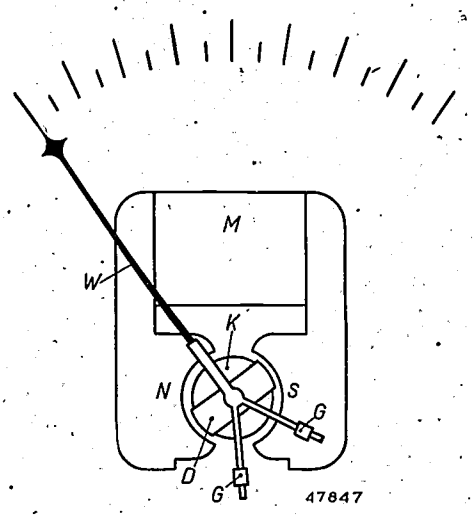


Fig. 3. Moving-coil meter. The permanent magnet M magnetizes the pole pieces N-S between which the moving coil D can rotate about the core K. The pointer W is attached to the coil. The movable counterweights G serve for balancing the moving system.

the air gap and thus situated in only a weak magnetic field. But the most important part, of course, is the copper wire used for the coil, which will be the main subject of our discourse in the following.

In order to be able to manufacture the necessary non-ferrous material, we first sought a simple method of quickly testing different samples for iron content. The choice fell upon a determination of the

[5]) For the reasons why a logarithmic scale is used here by preference, see Philips Techn. Rev. 2, 50, 1937.

remanence after magnetization in a field strong enough to reach saturation. Taken per unit weight of the material examined, the result is a measure of the magnetic behaviour of the substance. The remanent magnetic moment M_r is determined in a manner as illustrated in *fig. 4* and further explained in the text below the illustration.

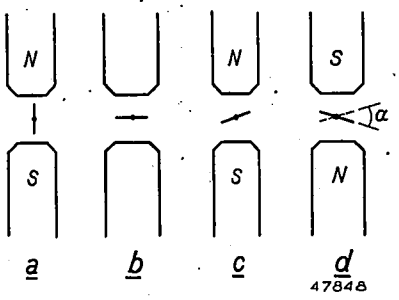


Fig. 4. Determination of the remancnt magnetic moment of a copper specimen containing iron, for instancc a small rod suspended on a torsion wire betwecn the *N-S* poles of an electromagnet.
a) With a strong field the iron particles are magnetized.
b) After switching off the field the rod is turned 90°.
c) The electromagnet is weakly excited, causing a deflection of the rod.
d) After commutation the rod shows the opposite deflection. The angle through which it turns upon transition from *c*) to *d*) is measured with mirror and scale.

From the angle α through which a test rod of the material under examination rotates upon commutation of the field, thus upon transition from the state of *fig. 4c* to that of *fig. 4d*, and from the prevailing field strength H, it follows for the remanent magnetic moment M_r that:

$$M_r = \frac{a\tau}{2H},$$

where a is the torsion constant of the suspension wire.

If G_{Cu} is the mass of the sample of copper, then $m = M_r/G_{Cu}$ gives the above-mentioned measure of the ferromagnetic behaviour.

Measurements were first carried out with the apparatus of fig. 4 on technical electrolytical copper wire. The values found were very divergent, due for a large part to contaminations on the surface, as was very soon found. After removing a thin layer by etching in nitric acid the results lay between 2×10^{-5} and 4×10^{-5} Gauss cm³/gram and showed no change upon further etching (until the diameter of the wire had been reduced by one half). From this it may be concluded that the remaining iron particles are distributed homogeneously in the copper.

In the meantime it had been found that for the construction of meters with no directional couple the iron content of etched technical copper was a factor 10 to 20 too high. Consequently copper with a value of m not higher than about 10^{-6} was necessary.

Manufacture of non-ferrous copper wire

A suitable material for this purpose was found in so-called *OFHC* copper ("Oxygen-Free, High-Conductivity"), which was available as wire 1 mm thick, and after etching off a layer 10 μ thick, this had a homogeneously distributed iron content corresponding to $m = 1.5 \times 10^{-7}$. In this state of purity the material is diamagnetic, in contrast to technical copper. The measurement of the remanence, however, is still of value, because the diamagnetism shows no hysteresis.

This material thus reasonably satisfied the requirement as to purity, but it still had to be drawn out to a thickness of 25 μ needed for the construction of a certain meter. When this was done in the ordinary way, the iron content sometimes rose to $m = 10^{-5}$. It therefore proved necessary to draw the wire in a specially equipped wire drawing machine from which iron in every form had been eliminated as far as possible, which means, among other things, that:

the dies (diamonds) must be kept reserved exclusively for the drawing of non-ferrous copper wire; instead of steel only bronze tongs, forceps, etc. must be used;

no tap water, but only distilled water, must be used in the etching baths and for rinsing;

the atmosphere must be kept free of dust;

the wire must not be touched, not even with clean hands.

The importance of such measures will be evident from the following example.

In the above-mentioned apparatus a piece of copper gave a deflection of 12.5 cm. It was removed with steel forceps, rubbed slightly along the forceps and again hung in the apparatus. The deflection was then 19 cm. This corresponds to an increase in the magnetic moment of 3×10^{-7}, *i.e.* twice as much as 1 gram of the purest material contains! The fact that the iron content is indeed very small also follows from its determination by a chemical method, which leads to results of the order of 10^{-2}% for technical and 10^{-4}% for "non-ferrous" copper.

Practical results

The use of non-ferrous copper wire opens the possibility of considerably improving the quality of various kinds of moving-coil meters. Thus it is

possible in the case of meters with a non-linear scale to reduce the zero-point error, which may otherwise amount to one scale division, to an inappreciable value. Fig. 2 already showed the improvement in the residual couple of a millicoulomb meter. Before the use of non-ferrous material this couple caused the system to continue to move after with a velocity of about 1/20 scale division per second after the passage of current has ceased. When changing over to non-ferrous material, among other measures, this velocity was brought down to a maximum of 1/60 to 1/100 scale division per second in only a small part of the scale, while in the rest of the scale it was reduced to zero.

Issue No. 5 (of volume 1) of *Philips Research Reports* contains the following papers:

R22: A. A. Padmos: Checking the thermal expansion of glasses and metals by stress-optical tests.

R23: J. L. H. Jonker: The control of the current distribution in electron tubes.

R24: W. Elenbaas: Similarity of high-pressure discharges of the convection stabilized type.

R25: M. Gevers: The relation between; the power factor and the temperature coefficient of the dielectric constant of solid dielectrics IV.

R26: A. van der Ziel: Theory of grounded grid amplifier.

Readers interested in one of the above mentioned articles apply to the Administration of the Philips Physical Laboratory, Kastanjelaan, Eindhoven, Holland, where a limited number of copies are available for distribution. For a subscription to Philips Research Reports application should be made to the publishers of „Philips Technical Review".

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

Reprints of the majority of these papers can be obtained on application to the Administration of the Research Laboratory, Kastanjelaan, Eindhoven, Netherlands. Those papers of which no reprints are available in sufficient number, are marked with an asterisk.

1693: J. M. Stevels: Chemische binding en reactiviteit van eenvoudige organische verbindingen (Chem. Weekblad 42, 150-154, 1946) (Chemical bond and reactivity of simple organic compounds).

From experimental data the author shows that, at least for halogenated methane derivatives, the following properties of a C-X bond go together: stronger binding, larger dissociation energy, larger activation energy and collision number in the reaction with Na in the gas phase, smaller binding refraction, greater force constant of the C-X vibration. Extending to unsaturated compounds one may add stronger double bond character, smaller atomic distance.

1694: Balth. van der Pol. Merkwaardige eigenschappen van geheele getallen (Natuurkundige Voordrachten, Diligentia, band 23, den Haag, van Stockum, 1946).

(Remarkable properties of whole numbers).

This lecture treats some remarkable properties of whole numbers: prime numbers, Goldbach's conjecture, complex prime numbers, perfect numbers the number of prime numbers in a given interval, Skewes' number.

1695: P. J. Bouma. Colour equations (Physica 12, 189-194, 1946).

Two types of equation occur in colorimetry: algebraic equations and colour equations. The difference between these two kinds of equations is explained, and a special symbol (- - - - - -) is introduced in order to avoid the danger of confusion between the two types. This danger is illustrated by a worked example where the same problem is solved without and with the use of colour equations.

1696: J. F. H. Custers and J. C. Riemersma. The texture of straight-rolled and of cross-rolled molybdenum (Physica 12, 195-208, 1946).

With the aid of pole figures the textures of straight-rolled and of cross-rolled molybdenum are determined. The pole figures thus obtained show, that earlier authors described these textures in too simple a way; they are at least twofold.

For example, after cross-rolling, there is found besides the so-called, (100) (110) texture (where the (100) plane is parallel to the rolling plane, and the (110) direction is parallel to the rolling direction) a second texture, which is rotary symmetrical around the normal to the rolling plane, and which has a (111) plane parallel to this plane.

The texture of straight-rolled molybdenum turns out to be in good agreement with the texture of straight-rolled iron, as determined by Krudjumow and Sachs; the texture of cross-rolled iron is not known.

VOL. 8 No. 11, p. 321-352 NOVEMBER 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

X-RAY FLUOROSCOPY WITH ENLARGED IMAGE

by G. C. E. BURGER, B. COMBEE and J. H. van der TUUK.

621.386.8:
616-073.755:
620.179.152

The intrinsic blurring of a fluoroscope screen (for instance about 1 mm) interferes with the observation of small details showing little contrast. By placing the object at a distance from the screen and close to the focus an enlarged image is obtained in which the (constant) screen blurring has relatively less effect. In order to prevent the geometrical blurring, which increases rapidly upon enlargement, from spoiling the desired effect, X-ray tubes with a very small focus must be used for enlarged fluoroscopy. Upon enlargement all the contrasts in the image also increase considerably, due to the fact that with a large distance between object and screen practically no secondary, scattered radiation falls on the screen. By the improvement in definition and contrast certain groups of otherwise invisible details become perceptible, while at the same time the enlarged images can be studied more quickly and with less fatigue. In certain cases the enlargement is accompanied by a significant loss in (primary) screen brightness. The detrimental effect of this on the perceptibility of details is analysed. In medical diagnostics the enlargement makes it necessary to limit the time a patient is kept under observation. Phantom tests and a small series of examinations of lung patients have led to the conclusion that enlargement may well constitute an appreciable advance in diagnostics. The technique to be followed in this case has to be further worked out. In a separate section a description is given of the Philips X-ray tubes which were specially developed for the experiments in question. They work with electrostatic focusing and provide a choice of two foci, one normal focus of 2.0 or 1.2 mm width and a fine focus of 0.3 mm width.

Blurring of the image in fluoroscopy

The shadow picture obtained in X-ray screening of any object is always affected by a certain blurring, which may be ascribed two to causes. In the first place the focus of the X-ray from which the projecting rays are emitted is not a point. The finite width of the focus causes a "geometrical blurring" O_g (half shadow width), which according to *fig. 1* is given by

$$O_g = \frac{b}{a} f, \quad \ldots\ldots\ldots \quad (1)$$

where a and b are respectively the distance between object and focus and that between object and screen. In the second place the mechanism of the excitation of the light on the fluorescent screen causes a blurring of the image. Every fluoroscope screen has for this reason an intrinsic blurring, the "screen blurring", O_s, which for a normally good screen may amount, for example, to 0.6 to 1.0 mm. The geometrical blurring is usually smaller. For instance in the case of normal lung screening $a \approx 50$ cm, $b \approx 20$ cm, and foci of for instance 1.2 mm width are used; for such a case $O_g \approx 0.5$ mm.

In the testing of materials as well as in medical diagnostics the object in fluoroscopy is to be able to observe very small objects, often with little contrast (minute holes or cracks, errors in assembly of the piece of work; disease nuclei in the lungs, small ulcers in the stomach, etc.) The blurring of the image will be a handicap in such cases. We shall explain this further, but for the present it may be stated roughly that the perceptibility of

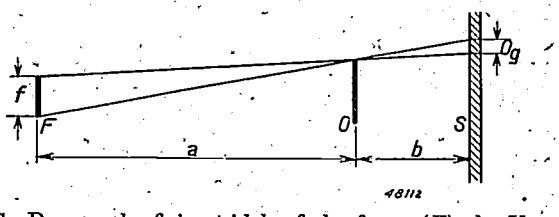


Fig. 1. Due to the finite width of the focus (F) the X-ray shadow picture of the object (O) on the fluoroscope screen (S) has a geometrical blurring O_g.

a detail with little contrast will become doubtful as soon as the detail is no larger than the blurring of the image. With the above mentioned numerical values of the blurring, therefore, all kinds of details (sometimes important ones) smaller than about 1 mm would escape observation.

Principle of enlarged fluoroscopy

It is obvious that only in case of necessity will such a limitation be tolerated. There now exists a method, which in principle is very simple, to improve the perceptibility of small details, namely fluoroscopy with enlarged image. For that purpose the arrangement of fig. 1 is modified in such a way that b becomes several times as large as a; see *fig. 2*. It may be read off from the figure that the object is projected on the screen enlarged by a factor

$$v = \frac{b+a}{a} \quad \ldots\ldots\ldots \quad (2)$$

Since all details are enlarged in the same proportion (or, with a thick object, in approximately the same proportion), while the intrinsic screen blurring remains unaltered, small details then have

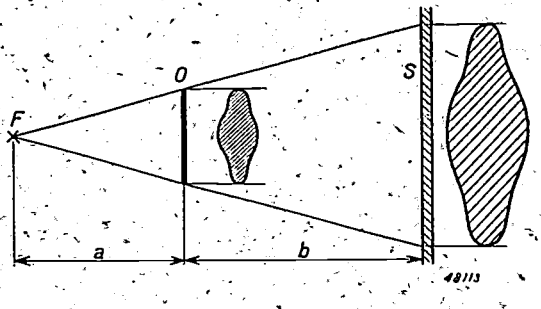


Fig. 2. By making the distance b between object (O) and screen (S) large enough, compared with the distance a between object and focus (F), a more or less enlarged fluoroscopy image of the object can be obtained.

a greater chance to show up above the screen blurring. This would indeed mean that with the change in the ratio of distance b/a mentioned the geometrical blurring would increase considerably (equation (1)) and therefore neutralize the advantage gained. It is possible, however, to compensate the increase of the factor b/a in equation (1) by making the width of the focus f correspondingly smaller. For the practical application of enlarged fluoroscopy therefore special X-ray tubes with a very small focus are desired.

In the following we shall replace this rough picture with a more quantitative formulation and also discuss the influence of various other factors involved in enlarged screening. It will be found that the situation for the testing of materials is somewhat different from that for medical diagnostics. The possibilities of medical applications, which have been studied carefully by us in recent years, will be dealt with in somewhat more detail. At the same time the X-ray tubes especially developed by Philips in connection with the magnification technique will be described.

The improvement attainable by magnification

The critical size of detail

It is unnecessary to prove that a larger and sharper image is more pleasant and offers more certainty of judgement. But it may be asked whether in the enlarged, relatively sharper image details can indeed be observed which were formerly invisible. In order to answer this question it is first necessary to consider the detrimental effect of the blurring.

To begin with, we condiser the case where b is very small; no enlargement is thus employed ($v \approx 1$) and we also assume a point focus, so that the geometrical blurring is zero. Outside the shadow of the detail to be observed let the screen brightness be H_2, in the middle of the shadow H_1 (see *fig. 3a*). At the edge of the shadow, due to the screen blurring, the brightness does not fall abruptly from the value H_2 to the value H_1, but there is a transition region of width O_s in which the brightness gradually varies from H_2 to H_1 (for the sake of simplicity we consider this variation to be linear). For the perceptibility of the detail drawn in fig. 3*a*, where the diameter d is larger then the screen blurring O_s, this gradual transition of brightness has no unfavourable result. If the contrast between H_1 and H_2 is equal to or larger than the minimum contrast which the observer can still just distinguish at the angle of vision corresponding to d, the detail will be observed in spite of the blurring [1]).

The situation becomes different when the detail is smaller than the screen blurring. In fig. 3*b* it may be seen that the regions of the transition in brightness at opposite edges now partially overlap. The brightness behind the detail nowhere falls to the low value H, *i.e.* the screen blurring results in a decrease in contrast. Thus if a detail of the size $d < O_s$ without the screen blurring has just sufficient contrast to be observed, then

[1]) With decreasing angle of vision the value of the minimum perceptible contrast increases. Due to the cutting down of the region of full contrast of a detail following from fig. 3*a* (difference in brightness $H_2 - H_1$), therefore, perceptibility is somewhat unfavourably affected. The impression of contrast upon a gradual transition in brightness $H_2 \rightarrow H_1$ is also probably smaller than upon an abrupt transition.

owing to the decrease in contrast caused by the blurring it will remain below the required size. In this case we may call O_s the critical size of detail.

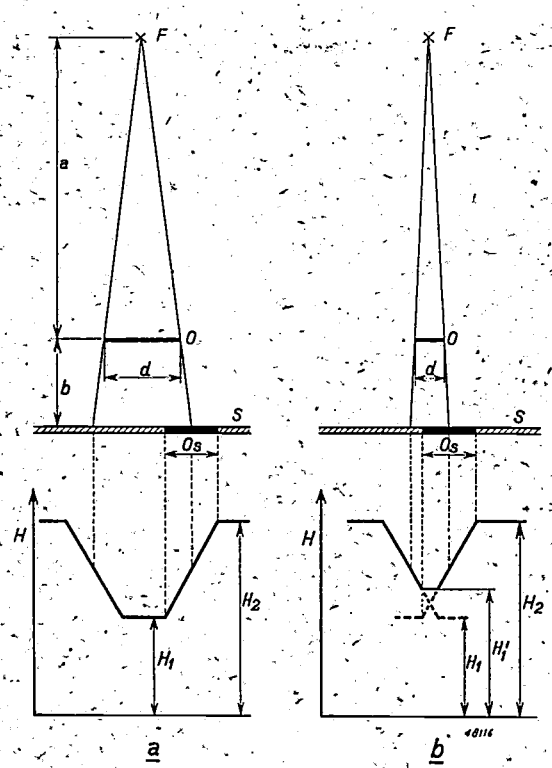


Fig. 3. Influence of the intrinsic screen blurring O_s on the projection of a detail of size d. The focus is assumed to be a point, so that the geometrical blurring $O_g = 0$. The distance b is assumed to be still so small that no appreciable enlargement occurs ($v \approx 1$). Below, the brightness H on the fluoroscope screen is plotted as a function of the position along a cross-section.
a) Large detail, $d > O_s$.
b) Small detail, $d < O_s$. The blurring here causes a decrease in contrast; $d = O_s$ in this case is the "critical detail size".

Let us now consider the same case when screening with magnification ($v > 1$). The shadow picture of the detail becomes v times as large, and the constant screen blurring O_s can now only cause a decrease in contrast when $v \cdot d < O_s$, *i.e.* in the case of details which are v times as small as before. Thus only v times smaller details are apt to escape observation. This is the effect with which we are concerned in the enlargement.

The effect is partially neutralized due to the fact that upon enlargement the total blurring is in any case increased; as already mentioned the geometrical blurring now comes to the fore. If we assume for the moment that only the geometrical blurring O_g is present ($O_s = 0$), then the situation for a large and a small detail is represented by *fig. 4a* and *b* respectively. O_g decreases the contrast for small details in exactly the same way as explained above for the screen blurring O_s. "Decrease of contrast" is now only another expression for the simple fact that the fluoroscope screen is no longer affected by the core shadow of the detail (see fig. 4*b*). The condition for decrease of the contrast is, analogous to the above, $v \cdot d < O_g$ and since

$$O_g = (v-1) \cdot f \quad \ldots\ldots\ldots \quad (3)$$

(see equations (1) and (2)), the perceptibility is unfavourably affected for details with

$$d < f\left(1 - \frac{1}{v}\right) \quad \ldots\ldots\ldots \quad (4)$$

If, therefore, the enlargement were very large ($v \gg 1$), so that the screen blurring becomes negligible compared with the geometrical blurring, then according to equation (4) the critical size of detail simply becomes equal to the width of focus. From this the importance of making the focus as small as possible for magnification technique becomes quite evident.

The simultaneous occurrence of geometrical and screen blurring results in a total blurring O_t, which can be approximately represented [2]) by

$$O_t = \sqrt{O_g^2 + O_s^2} \quad \ldots\ldots \quad (5)$$

Decrease in contrast occurs at $v \cdot d < O_t$, and thus

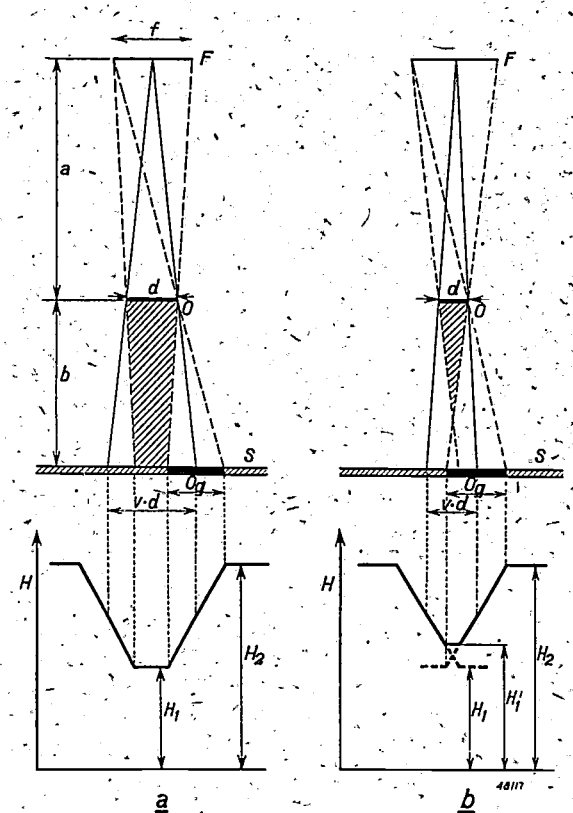


Fig. 4. Influence of the geometrical blurring O_g. It is assumed that $O_s = 0$. The figure is analogous to fig. 3.
a) Large detail, $d > O_g$.
b) Small detail, $d > O_g$. There is a decrease in contrast.

[2]) R. R. Newell, Brit. J. Radiology **30**, 493, 1938.

the critical size of detail in the general case is

$$d = \frac{1}{v}\sqrt{(v-1)^2 f^2 + O_s^2} \quad . \quad . \quad . \quad . \quad . \quad (6)$$

or, if the formula is written in the non-dimensional form

$$\frac{d}{O_s} = \frac{1}{v}\sqrt{(v-1)^2\left(\frac{f}{O_s}\right)^2 + 1} \quad . \quad . \quad . \quad (7)$$

In *fig. 5* this relation is represented graphically. The ratio d/O_s is plotted as a function of the enlargement v for different values of the parameter.

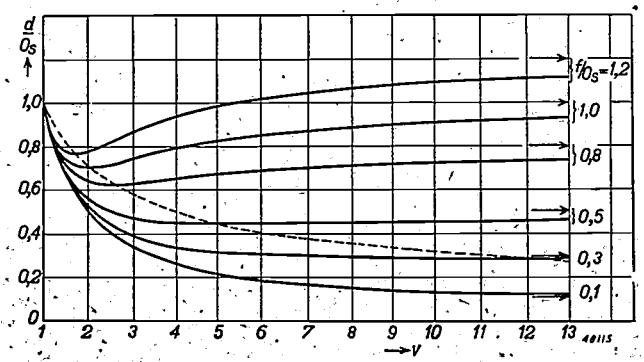


Fig. 5. In the presence of a screen blurring O_s and simultaneously a geometrical blurring (focus width f) the critical size of particle d, below which a decrease in contrast occurs due to the blurring, is given by equation (7). The ratio d/O_s is here plotted according to this equation as a function of the enlargement factor v, with the quotient f/O_s as parameter. Each curve passes through a minimum at an enlargement factor $v_m = 1 + (O_s/f)^2$; the level of these minima is given by $(d/O_s)_{min} = 1/\sqrt{v_m}$ (dotted-line curve). With very small foci, for instance for $f/O_s = 0.3$, however, at an enlargement of $v \approx 4$ practically the whole theoretical gain is already obtained.

f/O_s. It may be seen that with a small enough focus the critical size of detail steadily decreases with increasing enlargement, but that at the same time this decrease is limited by the width of focus f. At the value $f = 0.3$ mm, which has been realized in the X-ray tubes to be described later, and with $O_s = 1.0$ mm, the practical enlargement will be not more than 4 times, since the further decrease of d is then no longer appreciable. The critical size of detail here is $d = 1/3$ mm, while without enlargement $d = O_s = 1$ mm!

The scattered radiation

In addition to diminishing the influence of the screen blurring enlargement also has another favourable effect which lies quite outside the considerations already discussed. The longer the distance (b) between object and fluoroscope screen, the smaller the percentage of secondary X-radiation [3]) scattered by the object which falls on the screen. This scattered radiation gives a uniform fog over the image which reduces all contrasts — an effect, therefore, which in the case of enlargement becomes less potent or even almost disappears. A gain in contrast is thus obtained over the whole image. The degree to which the contrasts may increase is illustrated by the fact for instance in lung fluoroscopy the scattered radiation may have the same intensity as the primary radiation and in screening of the abdomen its intensity may even be three or four times as great [3]).

The effect of the enhanced contrasts is further increased by the fact that upon enlargement the visual angles within which all the details of the object are seen increase proportionally. For details with a larger visual angle an observer has a higher sensitivity to contrast (*cf.* footnote [1])).

In *fig. 6* two photographs (*b* and *c*) are given of an object (*a*) on the fluoroscope screen with and without enlargement. It may be seen that small details originally invisible or almost so are rendered visible and the contrasts in the enlarged image become larger. For the sake of comparison a reproduction (*d*) is also given of the normal fluoroscopic image subsequently enlarged optically. Since the screen blurring is here also enlarged and the fog due to the scattered radiation remains, no improvement is obtained.

Ease of observation

A comparison of figures 6*b* and 6*c* also gives an idea of the more convenient and easier observation of the enlarged image to which reference has already been made. The concrete advantage of this in practice is that perception of details is much quicker and less fatiguing.

The screen brightness

Influence on the improvement obtained

It would be premature to conclude from the above that in every case (with sufficiently small focus) enlargement would be an advantage. We have untill now disregarded the fact that upon enlargement the brightness of the image on the fluoroscope screen may decrease.

We must here make a distinction between the contribution to the brightness of the primary (the image-forming) and that of the secondary (scattered) X-radiation. The contribution to the brightness of the secondary radiation decreases in any case upon enlargement. This means that all kinds of physiological factors which remain to be discussed are affected unfavourably to some extent, but on

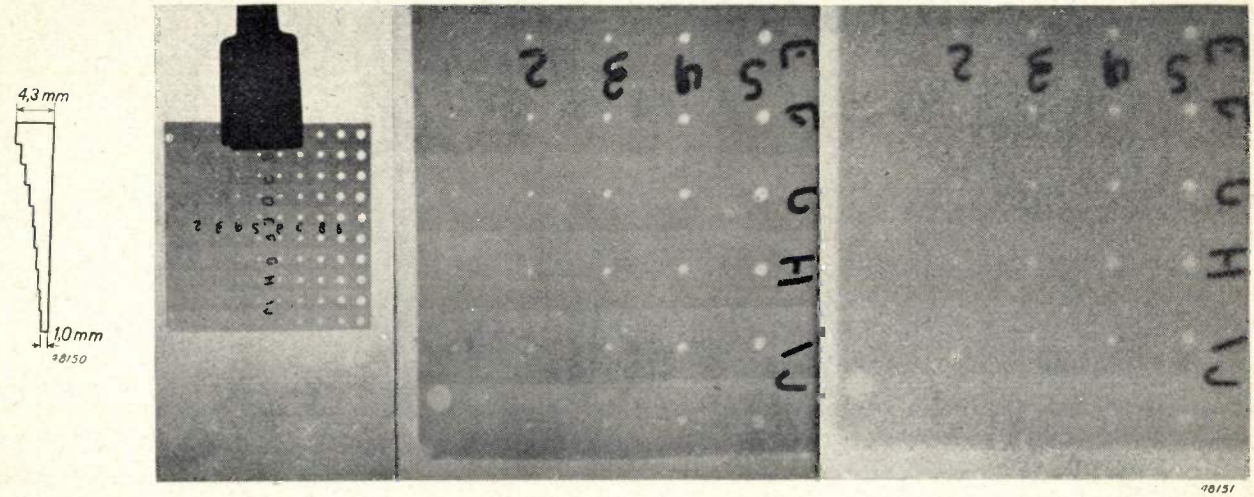
[3]) See for example W. J. Oosterkamp, Combatting the scattered radiation in the medical X-ray image, Philips Techn. Rev. 8, 183, 1946.

the whole the disappearance of the scattered radiation can only be favourable. However, the contribution to the brightness of the primary radiation may also decrease appreciably upon enlargement. We shall see presently that in the testing of materials this is very often the case, while in medical diagnostics it need not occur invariably.

With decreasing brightness of screen the visual acuity and sensitivity to contrast of the observer diminish. The advantage of the increase in all contrasts by enlargement (including the already mentioned higher sensitivity due to the larger angles of vision) is partly neutralized by a diminution of the sensitivity to contrast. As to the blurring the critical size of detail caused thereby does not depend upon the brightness, but the decrease in contrast in the case of details with dimensions smaller than the critical, when the observer's sensitivity to contrast decreases, becomes fatal already at greater contrasts, so that a larger group of details falls into the category of those no longer perceptible. Part of the gain which enlargement brings especially for the observation of small details poor in contrast is thus lost again.

In addition to the (small and large) details with poor contrast to which all the above discussion refers, we must also pay attention to small details with high contrast. In addition to the advantage of the sharper reproduction on the whole, enlargement gives for these details no improvement in perceptibility. A loss in brightness which accompanies the enlargement may even worsen the perceptibility of such details. This may be explained as follows:

The perceptibility of the details rich in contrast is limited by visual acuity, *i.e.* by the minimum angle of vision α_{min}, in which the observer can still just distinguish a detail. When the brightness decreases due to enlargement, α_{min} increases. At the same time, however, the angle of vision α of the detail increases proportionally with the enlargement factor v. The perceptibility is unfavourably affected when α_{min} increases more rapidly than v. The relation between α_{min} and the brightness is known from a number of investigations. The relation between the brightness and v may, as already indicated be very different. If we assume, for example, a variation of the screen brightness proportional to $1/v^2$, it is then found that as long as the brightness remains above a certain limit the increase of α remains ahead, so that upon enlargement the "effective" visual acuity improves. At what enlargement this limit of brightness is met depends, of course, on the initial brightness (for $v = 1$) and this in turn is quite different according to the nature and thickness of the object. Enlargement above a certain factor v, which also depends upon the object, will thus have an unfavourable effect on the observation of the very small details rich in contrast, although the visibility of the details poor in contrast is improved. This fact may be of importance in the testing of materials.



a b c d

Fig. 6. Normal (*b*) and four-fold enlarged (*c*) fluoroscopic picture of a test object (*a*) consisting of an aluminium plate with stepwise varying thickness and in each step a series of holes of different diameter. *d*) The same picture as in (*b*), *i.e.* obtained with normal fluoroscopy, but reproduced on a scale four times as large as (*b*); no appreciable improvement compared with (*b*). The fact that in the reproductions of this review more details can be discerned in *d* then in *b* is due to the new „blurring" which the lattice of the autotype introduces; in this lattice small details disappear rather than large ones.

Screen brightness and method of enlargement

Beginning with normal fluoroscopy where object and screen are set up close to each other, it is possible to enlarge in two ways:

1) the object and the X-ray tube can be left in the same positions and the screen placed farther away (a constant, b and thus also $a + b$ larger);
2) the screen and the focus can be left in position and the object placed closer to the focus (a smaller, b larger, $a + b$ constant). The screen

brightness is inversely proportional to the square of the distance $(a + b)$ between focus and screen; in the first method, therefore, upon enlargement a considerable decrease in brightness is obtained, while in the second method the (primary) brightness remains constant.

In order to avoid the above-mentioned disadvantages of a decrease in brightness, the second method seems to be the better one. For the testing of materials, however, it cannot always be applied; in particular it cannot be applied in those cases where with normal fluoroscopy, in order to obtain the greatest possible screen brightness, the object and the screen are placed as close to the X-ray tube as the dimensions of the object allow. It is therefore only possible to enlarge in such a case by placing the screen farther away, and the decrease in brightness proportional to $(a + b)^2$ or, according to equation (2), proportional to v^2, must be accepted into the bargain.

In medical diagnostics, on the other hand, with normal fluoroscopy the patient is always placed at a fairly great distance from the focus in order to obtain an overall picture of the part of the body concerned [4]). Increasing the screen brightness by placing patient and screen closer to the focus would be of no use here since a greater brightness could also be realized by increasing the current through the X-ray tube. The fact that this is not done and thus that the focus is not loaded to anywhere near the permissible limit is due to the limitation of the X-ray dose which may be administered to a patient in fluoroscopy. The intensity of radiation on the patient, by which the screen brightness is also approximately determined, is chosen so low that the doctor may devote several minutes to the perception of the fluoroscope image and still remain far below the permissible dose. If it is now desired to pass over to enlarged fluoroscopy the second method can in this case be applied, *i.e.* the patient can be placed closer to the focus provided a correspondingly shorter time of observation is allowed in order to keep the dose on the patient within safe limits in spite of the stronger intensity of radiation. Since the distance between focus and screen here remains unchanged the (primary) screen brightness with enlarged fluoroscopy is no less than with normal screening.

Screen brightness and focus loading

We have explained that for enlarged screening a much smaller focus is necessary than is normally used. If the focus is normally loaded to the limit of the power it can receive continuously per unit of surface without becoming too hot, decrease in size of the focus necessarily means a reduction in the current in the X-ray tube and thus in turn a decrease in the screen brightness. In testing materials where powers of 2 kW, for instance 150 kV_{max}, 20 mA and even more are used continuously, with the normal large foci the permissible focus temperature is indeed approached. Nevertheless, the decrease in size of the focus need not have such an unfavourable effect on the screen brightness as would follow from the ratios of the surfaces of the foci. A smaller focus has a higher specific loading capacity (at the same focus temperature). This is due to the fact that the heat developed on the small focus is dissipated not only towards the interior of the anode but to an appreciable extent also laterally (edge effect). Moreover, the specific focus loading capacity can be further increased by the employment of a rotating anode [5]).

In medical diagnostics, due to the above-mentioned limitation of the dose on the patient, with normal fluoroscopy the permissible specific focus loading is far from being attained. A tube voltage of for instance 75 kV_{max} with a current of 2 mA is used, *i.e.* a continuous loading of about 150 watts. It is here unnecessary to have the anode rotate. If the focus is now decreased in size, the focus temperature increases and at the smallest foci realized, of 0.3 mm, it is found necessary to let the anode rotate in order to be able to employ the above-mentioned 150 watts [6]).

Constructions of X-ray tubes with very small focus

The requirement for obtaining a small focus is a strong concentration of the electron beam emitted by the filament of the X-ray tube. In tubes for testing materials, where the focus must sometimes

[4]) If the patient is close to the focus the relative differences in distance for various details in the body are so large that some are projected considerably larger than others.

[5]) See for example J. H. van der Tuuk, X-ray Tubes with Rotating Anode ("Rotalix" tubes), Philips Techn. Rev. 8, 33, 1946; also Philips Techn. Rev. 3, 296, 1938.

[6]) It should be pointed out that the enlarging technique, at least as far as the improvement in definition is concerned, does not as yet offer advantages in making X-ray photographs. It is true that here too there is a kind of "screen blurring", namely that caused by the grains of the usual reinforcing screens. But in enlarged photography in the case of moving objects the blurring due to motion would increase very much, since the small focus for instantaneous photography in any case can be less heavily loaded than a large focus and for that reason alone the exposure would have to be longer. For stationary objects there is no objection to longer exposure, but just for that reason it is then possible, if greater definition is desired, to omit the reinforcing screens, so that only the very slight intrinsic lack of definition of the X-ray film itself remains.

be introduced into all kinds of cavities, the anode with the focus is often placed in a projection of the tube at a distance of 20 to 30 cm from the filament. For the small as well as for normal foci the required concentration can then only be attained by focusing with the help of one or more magnetic coils. This method gives good results but is rather elaborate. With a given magnetic field the focusing is only good for a given velocity of electrons. If it is desired to vary the voltage on the X-ray tube according to the object to be examined, the exciting current of the magnetic coils must also be accurately adjusted anew each time. If an extremely small focus is desired, this adjustment has not only to be very precise but, moreover, for the proper functioning of the tube a carefully smoothed and constant D.C. voltage is necessary, since otherwise the size of focus alternates appreciably with the ripple of the tube voltage.

In X-ray tubes where cathode and anode are placed only about 1 cm apart — in medical diagnostics such tubes are used exclusively — the problem can be solved much more satisfactorily by employing electrostatic focusing. This means that by giving a suitable shape to the cathode the potential between cathode and anode is made to vary in such a way that a strong electrostatic "lens" is formed. This focusing is independant of the tube voltage. The potential of every point in the deflecting field changes proportionally with the tube voltage; since the deflecting action on an electron is proportional to this potential, and on the other hand inversely proportional to the tube voltage which accelerates the electron, the two effects exactly cancel out each other.

The tubes with very small focus developed in the Philips laboratories ("small focus tubes") employ electrostatic focusing. In fig. 4 the cathode of such a tube is shown. It may be seen that two spiral filaments are assembled in grooves of the cathode, one large and one very small. These can be switched on according to choice. The thin spiral furnishes the small focus of 0.3 mm width, the thick one gives a focus of 2 mm or, in other tubes, of 1.2 mm, which is intended for normal fluoroscopy and for making photographs.

The size of the focus obtained is checked in the familiar way by making a photograph of it with a pinhole camera. The aperture of this camera, which is normally made 0.2 mm, must in this case be considerably smaller (about 0.03 mm), in order to obtain a sufficiently sharp projection of the small focus.

Since the tube also serves for taking X-ray photographs when the larger focus is used, the anode is of a rotating type which in any case is desirable in order to be able to give the small focus a high load. In the usual constructions of rotating anodes

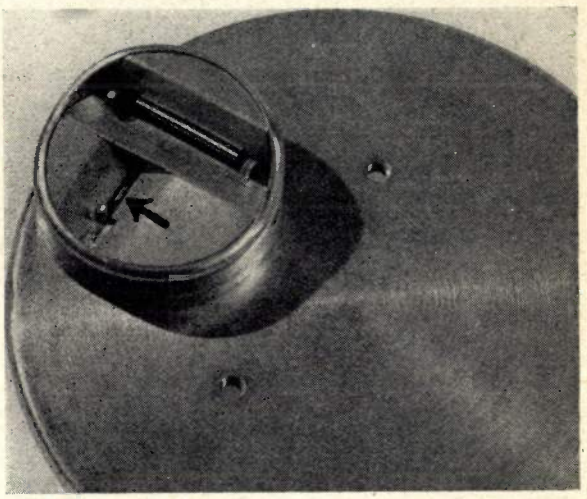


Fig. 7. Cathode of a double-focus X-ray tube, with rotating anode, specially developed by Philips in connection with the enlargement technique. The thick filament gives a focus of 2.0 mm width, with which photographs are made, the thin filament gives a focus of only 0.3 mm width, intended for normal and enlarged fluoroscopy. Thanks to the special construction of the cathode cap the electrons emitted are focused electrostatically on the anode.

the rotor runs in two ball-bearings. The guiding rings of the latter in the rotor as well as on the anode carrier have several tenths of a millimeter axial play in order to take up the difference in thermal expansion of rotor and anode carrier. For working with the small focus, however, the play referred to is undesirable. Axial vibrations of the anode might be caused; since the focus is thereby displaced perpendicular to the direction of the effective beam of X-ray, a relatively considerable increase in the apparent width of focus occurs. In order to avoid this effect a special bearing of the rotor has been made for which the axial tolerance is limited to several hundredths of a millimeter.

Practical application of enlarged fluoroscopy

In the testing of materials

The above statements about the limitation of the dosage in medical diagnostics might give the impression that the general situation for normal fluoroscopy is much more favourable for testing materials. It is possible to work with a very small distance between focus and object, with the maximum focus loading, with arbitrarily long times of observation (apart from economical considerations). These advantages, however, are practically neutralized by the disadvantage that in testing mater-

ials the objects dealt with have a much greater absorption. An iron plate 0.02 mm thick or an aluminium plate of 4.5 mm absorbs (with the same tube voltage of 75 kV) just as much as the chest of a patient. Thus in material testing the screen brightness is by no means so much more than sufficient as might be supposed. Moreover, in a workshop it is impossible to use such low screen brightness as can be used in the X-ray room of a medical establishment. The result is that the employment of fluoroscopy in material testing has always been limited to the light metals with relatively little power of absorption, such as aluminium and magnesium alloys.

Enlarged screening in the testing of materials, as was explained above, involves opposing effect in definition and brightness. Such is also the case, however, with normal screening technique. The screen blurring of 1.0 mm already mentioned is only valid for very sensitive screens having a high brightness. If a less sensitive screen were used, which gave for instance only half the brightness, the blurring would only be half as great. The compromise, which in normal fluoroscopy leads to the choice of a given screen, must be considered anew in enlarged fluoroscopy, and it is thus *a priori* still a question whether enlargement is really an advantage in material testing. Experience has answered this question in the affirmative and in the screening of light metals magnification technique has been employed for a number of years. We shall not go more deeply into the results obtained but refer only to the literature [7]).

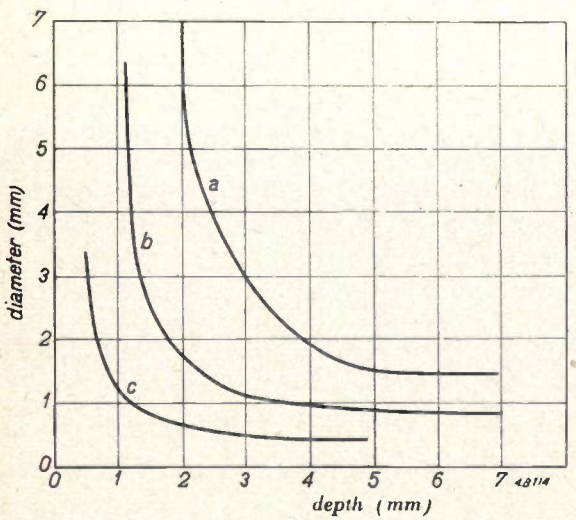


Fig. 8. Visibility tests on a phantom of "Philite" which is about equivalent to the human lung for X-ray fluoroscopy. Holes of different depth and diameter were bored in the phantom. For every depth (in mm) the diameter of the smallest perceptible hole (also in mm) is plotted.
a) For fluoroscopy without enlargement, focus width 0.4 mm, tube voltage 70 kV_{min}, tube current 1.4 mA.
b) For fluoroscopy with four-fold enlargement; data as in (*a*)
c) On an X-ray photograph under optimum conditions.

In medical diagnostics

We have seen that in medical diagnostics enlarged

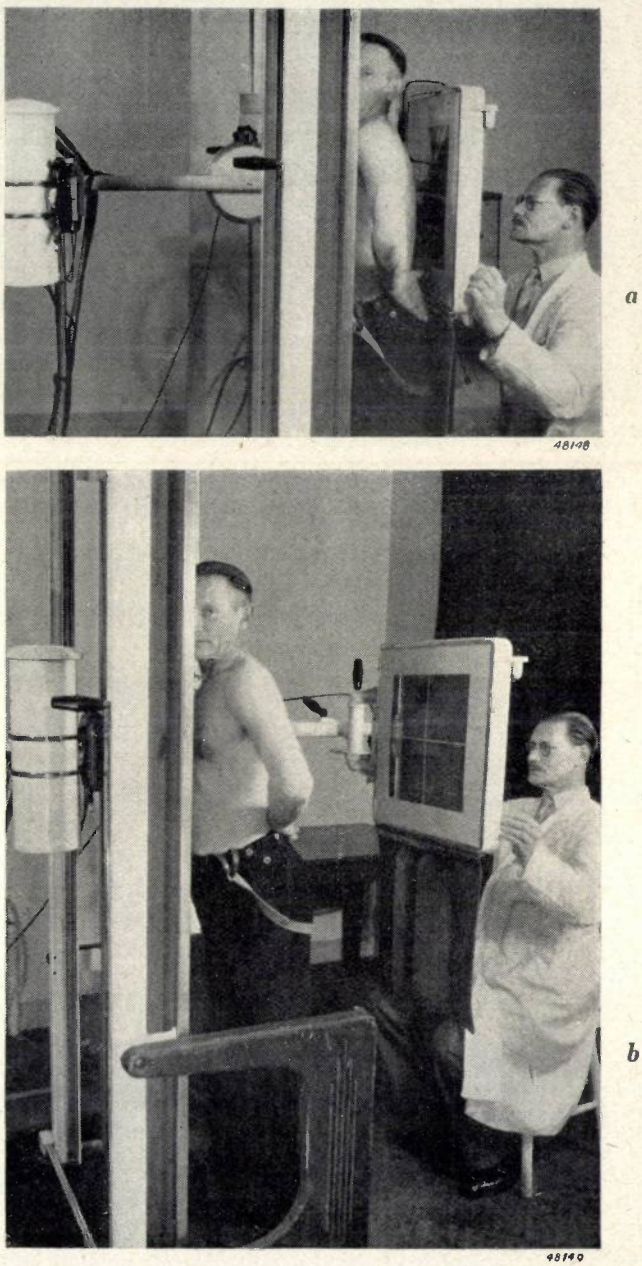


Fig. 9. Improvised arrangement for the enlarged screening of lung patients at the consultation bureau of the Medical Department of the Philips Factories. The patient's position is fixed, the fluoroscope screen and the X-ray tube are rigidly connected with each other and can be moved by the practitioner. Since in the periodically repeated, normal fluoroscopy the patient always stands with his back to the tube and thus the surface of the back always receives the highest radiation intensity (four times as much as a point in the middle of the lungs), for greater safety the patient is here placed with his chest facing the tube.
a) Situation for normal fluoroscopy.
b) Situation for enlarged fluoroscopy.

[7]) See for example R. Berthold, Atlas der zerstörungsfreien Prüfverfahren T/Rö 10/7-10, J. A. Barth, Leipzig 1938.

fluoroscopy need not lead to a loss in (primary) screen brightness, but only to a limitation of the time of observation. The favourable results which could be expected from this led to a systematic investigation of the possibility of using enlarged fluoroscopy in the Medical Department of the Philips factories.

This investigation, of which several preliminary results have been published elsewhere [8]), was mainly in two directions. In the first place observation tests were performed with a phantom of "Philite" which, as far as the absorption and scattering of X-ray is concerned, corresponds approximately to the chest of a patient. A number of holes of different diameter and depth were bored in the phantom. The graphic representation in *fig. 8* shows the visibility of these holes with normal fluoroscopy (*a*), when screening with four-fold enlargement (*b*) and on a good X-ray photograph (*c*). It may be seen that there is a significant improvement between normal and enlarged fluoroscopy. Moreover, these experiments already showed how important it was that the perception could take place more easily and more quickly with the large image than with the small one. Because of this the necessary shortening of the time of observation of a patient (to for instance 10 to 20 sec.) already mentioned is less of a disadvantage.

In the second place a group of lung patients in the routine examinations of the consultation bureau were examined not only with ordinary fluoroscopy but also with enlarged image. The improvised arrangement with which these experiments were carried out is shown in *fig. 9*. A standard on rollers supports the X-ray tube and the fluoroscopy screen. The screen first lies close against the patient so that a normal orientation image is obtained. After having examined this briefly and determined what part of the image requires special attention, the observer takes one or two steps to the rear drawing with him the standard with screen and tube, while the patient remains in his place. The observer can now see the part in question more clearly on the enlarged image.

One of the disadvantages for the doctor is that upon enlargement he literally loses contact with the patient. In normal fluoroscopy it is very useful to be able to turn the patient slightly and to move him about so as to make use also of the sense of touch.

When using enlargement this is not possible, because the time is too short and, moreover, the doctors' arms are not long enough to reach the patient from his position behind the more distant screen.

In spite of this the investigation gave encouraging results. From statistics of the small number of 51 cases examined [9]) it was found that by enlarged fluoroscopy in the case of 10 patients better insight was obtained into the structure of a lung process, 8 times the presence or absence of a cavity was ascertained with greater security (the diagnosis was later checked with the still better photographic image) and in general in 35 cases an improvement was noted, compared with 16 cases where enlarged fluoroscopy gave no concrete advantage.

One of the practical problems still to be solved before enlarged fluoroscopy can become part of the daily practice of the röntgenologist is the adequate protection of the patient against too high doses of X-radiation. This will be more necessary, because just the difficult or interesting cases which form the most fruitful objects for enlargement technique are usually observed for a longer time even in normal fluoroscopy. And this holds even more for the screening of the abdominal organs than for that of the lungs, since in the former case, due to the greater absorption, higher radiation intensities and in some cases (stomach examinatons) longer times of observation must be employed.

[8]) G. C. E. Burger, Hand. 29 Ned. Nat. Gen. Congres, Amsterdam 1943, p. 47. Fig. 8 is also borrowed from this.

[9]) G. C. E. Burger, not yet published.

THE TECHNIQUE OF INVESTIGATIONS WITH RADIOACTIVE AND STABLE ISOTOPES

by A. H. W. ATEN Jr. and F. A. HEYN. 539.155.2:539.167.3:539.16.08

When tracers (tagged atoms) are used for scientific or technical investigation, one is concerned with radioactive and, in some cases, with stable isotopes of certain elements. In as far as the choice is not prescribed by the problem itself, it is often the half-value time of the radioactive isotopes which determines it; the half-value times must be neither too long nor too short. Only in the cases of oxygen and nitrogen is one now limited to the use of stable isotopes, while heavy hydrogen is also still much used. The preparation of the radioactive isotopes is by means of nuclear reaction brought about, for example, with the help of a cyclotron, a high-voltage apparatus or in an uranium pile. Stable isotopes are obtained by the separation of the mixtures of isotopes occurring in nature, for instance by fractional electrolysis, distilation or diffusion. The synthesis of possibly desired compounds of tagged atoms is in practice still one of the most difficult problems that can occur in the application of the method. A neat solution is sometimes found by a biological method of synthesis, *i.e.* in micro-organisms or in the animal body. The analysis of the mixtures of isotopes obtained when applying this method is carried out, in the case of the radioactive isotopes, by measurement of the radioactivity, for instance with an electron counter; many details of this are discussed in this article. For accurate localization of the radioactive isotopes, auto-radiograms are very well suited. Mixtures of stable isotopes are analysed with the mass spectrograph, or, as in the case of heavy hydrogen for instance, by measurement of density.

The concise survey of the tracer or "tagged" atom method and its possibilities of application [1]) which we gave in the previous number of this periodical will be supplemented here by a discussion of the manner in which the method is applied.

The following is a resumé of the most important points in the article mentioned.

Of every chemical element occupying a certain position in the periodic system there are different kinds of atoms, differing in atomic weight. These mutually "isotopic" chemically indistinguishable kinds of atoms may be either stable or radioactive. When two isotopes of an element take part in a physical or chemical process they behave in practically the same way. Nevertheless one isotope, either because it is radioactive or because of its different atomic weight, can in principle be recognized, localized and quantitatively determined by the investigator at every stage of the process. These atoms are, as it were, tagged and may be used as "tracers".

Use is made of this fact for the investigation of very diverse processes, firstly because of the great eare with which the place, quantity and distribution of a radioactive substance is determined, and because of the extremely great sensitivity which can easily be attained; and secondly because certain types of processes have only been made accessible for investigation by the tracer method, namely those processes where there is an exchange of chemically identical particles. Technical examples are the diffusion of a substance in an identical environment and the transfer of material upon friction between two identical surfaces. Especially in the field of chemistry and in the physiology of plant and animal life such processes play an important part.

If it is desired to carry out an investigation with the help of tagged atoms, the following are the main practical problems encountered.

1) What kind of atom should be chosen?
2) How is it obtained.
3) In what form (chemical compound) should the atoms be used and how can they be brought into that form?
4) How are the mixtures of isotopes occurring in the experiments analyzed?

The answers to questions 2, 3 and 4 will often be simple: the tagged atoms, or in some cases the desired compounds of them, are ordered from the nearest laboratory equipped for making them, and the mixtures to be analyzed are sent to the same address for analysis. The fact that various laboratories possessing the necessary equipment — in Europe, among others, the Philips Laboratory in Eindhoven — have opened the opportunity for the application of this method, has very greatly promoted the rapid spread of its use. However, even for those who are thus relieved of the difficulties involved in questions 2, 3 and 4, it will also be of importance to be informed abut the technique followed in solving these questions.

Choice of the kind of atom

In the first experiments with tracers carried out by Hevesy and others a radioactive isotope of lead was usually employed, and sometimes also thallium or bismuth. For example, the assimilation of lead ions by plants was investigated, or the distribution of lead administered between a tumor and the normal tissue in the animal body. Since the plant and the animal body normally

[1]) Philips Techn. Rev. 8, 296, 1946.

contain no lead, it would in principle also have been possible to carry out these experiments by means of a chemical analysis, but then there would have been the practical difficulty of determining the extremely small quantities of lead involved, for which the measurement of radioactivity is particularly suited.

It was obvious that such experiments, performed with substances which do occur in the plant or animal and which are involved in their natural functions, could furnish much more important results. Moreover, from the point of view of methodics, such experiments were more interesting, since chemical analysis must in principle fail here, and only the use of tagged atoms which are also recognisable in a chemically identical environment was able to make these experiments possible.

The reason why the pioneers of the method chose lead, thallium or bismuth was simply because at that time no other tagged atoms were available. The three mentioned are the only non-radioactive elements of which radioactive isotopes with sufficiently strong radio-activity occur in nature.

The situation is now quite different. Since the discovery of artificial radioactivity in 1934 we have learned how to make one or even several radioactive isotopes of every element. The development of methods for the separation of isotopes has also made it possible to prepare stable isotopes of a series of elements in sufficient quantities for practical use. There is such a wide choice of tagged atoms that one might almost speak of an "embarras de choix", were it not that there are again a number of factors limiting the choice.

To begin with, the choice is often prescribed by the problem itself. If, for example, it is a question of the diffusion of certain alloy components in steel and their influence on the properties of the alloy, because of the usually very specific influence of an alloying element, tagged atoms of that same element will have to be chosen. The same is the case with specific biological effects of certain elements, such as the assimilation of iron by anaemics.

An example of a problem permitting a certain freedom of choice is the previously described investigation of the friction between metal surfaces, at least in so far as it is only a question of the mechanism of the friction and not of possible specific effects with certain metals.

In many cases the choice is limited by the half-value time of the radioactive isotopes which can be used (*i.e.* the time in which the radioactivity decreases to one half of its initial value). A list of the half-value times of a series of often used isotopes is given in the table. The half-value time must be neither too short nor too long. If it is very short it means that the radioactivity of a sample decreases rapidly and, in spite of the great sensitivity of the measuring instruments, falls below the limit of detection so quickly that there is no time to perform the experiment in a proper way. This is especially true for biological problems, where, in contrast to chemical problems, the processes to be investigated cannot be accelerated by raising the temperature or otherwise. If the half-value time is very long it means that only a few atoms disintegrate per unit of time, the radioactivity then being so slight already from the beginning that it may not be possible to measure it.

The curves in *figs. 1a* and *b* illustrate these facts. They also show that both with too short and with too long half-value times conditions for the measurements are improved when the concentration of the radioactive isotope in the substance is sufficiently increased. For instance, it was originally very difficult to carry out experiments with radioactive carbon, while because of the dominating position of carbon in all organic chemical compounds

Table I. Half-value times of several radioactive isotopes *)

The symbol of each chemical element is accompanied by an index (upper right) indicating its so-called mass number (atomic weight in whole numbers) and another (lower left) giving its number in the periodic system (nuclear charge).

Element	Atomic weight of the isotope mixture occurring in nature	Symbol of the radioactive isotope	Half-value time
Hydrogen	1.0081	${}_{1}H^{3}$	31 years
Carbon	12.010	${}_{6}C^{11}$	20.5 minutes
		${}_{6}C^{14}$	>1000 years
Nitrogen	14.008	${}_{7}N^{13}$	9.93 minutes
Oxygen	16.000	${}_{8}O^{15}$	2.1 minutes
Sodium	22.997	${}_{11}Na^{22}$	3.0 years
		${}_{11}Na^{24}$	14.8 hours
Magnesium	24.32	${}_{12}Mg^{27}$	10.2 minutes
Phosphorus	30.98	${}_{15}P^{32}$	14.3 days
Sulphur	32.06	${}_{16}S^{35}$	87.1 days
Chlorine	35.457	${}_{17}C^{38}$	37 minutes
Potassium	39.096	${}_{19}K^{42}$	12.4 hours
Calcium	40.08	${}_{20}C^{45}$	180 days
Manganese	54.93	${}_{25}M^{52}$	6.5 days
Iron	55.84	${}_{26}Fe^{59}$	47 days
Copper	63.57	${}_{29}Cu^{64}$	12.8 hours
Bromine	79.916	${}_{35}Br^{82}$	34 hours
Strontium	87.63	${}_{38}Sr^{89}$	55 days
Iodine	126.92	${}_{53}I^{126}$	13 days
		${}_{53}I^{128}$	25 minutes

*) Of most of the elements listed here several other isotopes exist which are not listed. The data are taken from G. Seaborg, Rev. mod. Phys. **16**, 1, 1944.

it was desirable to be able to carry out just such experiments. The isotope C^{14} with a half-value time of about 10 000 years was not active enough, and in the case of C^{11} with a half-value time of only 21 minutes the experiments often had to be given up prematurely. Later on, however, C^{11} was successfully prepared in such large concentrations that experiments concerned with the photosynthesis of the carbohydrates in plants could be continued with it for more than 5 hours. This corresponds to a decrease in the activity to about 1/30 000 of the initial value; actually the difference in activity between initial and final substance may be very much greater still, since the nature of the experiment may determine the fact that the final substance contains only a fraction of the original amount of carbon.

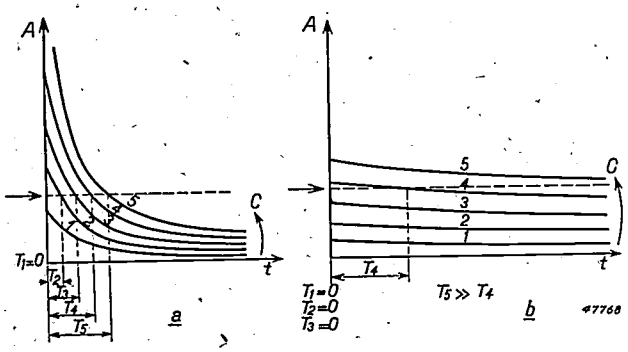


Fig. 1. Disintegration curves of a radioactive mixture of isotopes *a*) with a short half-value time, *b*) with a long half-value time. Each curve holds for a definite initial concentration of the radio-active isotope in the mixture, the concentration increasing in the direction of the arrow. The dotted line indicates the minimum activity that can be detected with the measuring instruments. T_1-T_5 are the respective times during which a measurement of the activity is still possible with the five preparations in question. The limit of the measurement indicated by the dotted line will in general be lower for substances with long half-value times than for those with short half-value times. In the first case the measurement of the activity can be accomplished by counting current impulses for a longer time, so that with fewer impulses per unit of time the statistical fluctuations of activity and zero effect are sufficiently exceeded (see under fig. 2 and at the end of the article).

If it is not possible in practice to realize the required initial concentration of an isotope with an unfavourable half-value time, there is still the possibility of using a stable isotope of the element in question, even though it is usually much more difficult to work with than the radioactive isotopes. It is unfortunate that the available radioactive isotopes of just the most important elements in organic chemistry — carbon, hydrogen, oxygen and nitrogen — have decidedly unfavourable half-value times. In the case of nitrogen and oxygen, for most experiments one must still use the stable isotopes N^{15} and O^{18}. For experiments with tagged hydrogen, the heavy hydrogen isotope with atomic weight 2, deuterium (H^2 or D), was formerly used on a large scale and is still being used.

In the case of a high initial activity, which according to fig. 1*a* is needed in the case of isotopes with a short half-value time, there may be a complication. When metabolic experiments are being performed on animals it must be taken into account that the radioactive radiation not only acts as a "spy" but may also have a destructive effect on tissues of the test animal. There is fortunately a tolerance dose which must be exceeded before any harmful effect of the radiation is manifested. If the initial activity is so great that this dose is exceeded, care must be taken that the progress of the experiment is not altered as a result of the irradiation.

In this connection it should be pointed out that it is one of the great advantages of the tracer method that, in principle, the metabolism can be studied without disturbing the processes thereof in normal living animals.

Incidentally we may mention here the remarkable fact that not only radioactive but also stable isotopes in too high concentrations can damage the life functions of a test animal. This was ascertained in experiments on the assimilation and excretion of water in test animals. The animals were allowed to drink water of which the molecules to a known percentage were built up of heavy instead of ordinary hydrogen, and the density of the water was then studied in the blood, lymph, urine, etc. In the human body, which consists for 80% of water, it was found that half of all the water in the body was renewed in 13 days. In goldfishes, which were simply allowed to swim in "heavy water", it was found that the water molecules in a fish are practically entirely renewed within a few hours. When, however, the concentration of D_2O in the water was very high, so that many H_2O molecules in the fish were replaced by D_2O, this caused the fish to die.

The preparation of tagged atoms

The radioactive isotopes not occurring in nature can be formed by certain transmutations of atomic nuclei (nuclear reactions). The following are examples of such transmutations:

$$_{16}S^{32} + {}_0n^1 \rightarrow {}_{15}P^{32} + {}_1H^1,$$
$$_{15}P^{31} + {}_1H^2 \rightarrow {}_{15}P^{32} + {}_1H^1.$$

In the first case the unstable phosphorus isotope P^{32} is formed from a sulphur atom and a neutron, in the second case from ordinary phosphorus and deuterium. These nuclear reactions do not take place, like chemical reactions, as soon as the electron clouds of the atoms approach each other, for which only little energy is necessary, but they require the direct influence of certain particles (for instance protons or neutrons) on the nuclei, which influence has to be brought about by causing the two to collide with each other with very great force. In our first example the necessary neutrons are obtained by shooting deuterium ions at a preparation containing lithium, the neutrons thus liberated being allowed to enter a vessel containing carbon disul-

phide in which, according to the equation, sulphur nuclei are transmuted into $_{15}P^{32}$. After a certain amount of ordinary phosphorus has been added in order to obtain quantities which can be dealt with practically, this phosphorus, together with the ordinary phosphorus, can be isolated by chemical means.

If the final product of a nuclear reaction is an isotope of the original atom, as in our second example, a chemical separation would seem impossible at first glance. If, however, one begins with a compound of the element to be transmuted, each atom transmuted will be thrown out of the molecule by the shock, so that the reaction product is in the form of free atoms or ions, which can then be isolated from the compound.

The strong forces required for nuclear reactions can be given to the atoms in a cyclotron, in which the (ionized) atoms are accelerated along a spiral trajectory to enormous velocities corresponding to voltages of several million or even tens of millions of volts; or in a high-tension apparatus working with a voltage of several hundred thousand or millions of volts [2]). The efficiency of the nuclear reactions, however, is generally very low, so that very intensive beams of bombarding particles are necessary to obtain the desired concentrations of tagged atoms. The apparatus for the preparation of the tagged atoms therefore becomes so elaborate and expensive that it can only be installed in a few laboratories. As already mentioned, however, various institutes are already making available to every investigator the preparations of tagged atoms which they make. It should be pointed out that a new possibility for obtaining radioactive elements has been opened by the construction of the uranium piles which played such an important part in the development of the atom bomb; numerous radioactive substances can be prepared with their help.

The separation of the radioactive substance formed in the nuclear reaction is not necessary in every case. The experiments already mentioned on the friction between metal surfaces, for example, were carried out with metal plates which had been irradiated in a cyclotron, whereby a small part of the atoms at the surface were transmuted *in situ*. The plates could be used without further treatment.

A similar case is found in a certain method of chemical analysis. If in a mixture of different elements one element occurs in which radioactivity can be excited by irradiation, without a radioactivity of comparable half-value time being excited in any of the other elements, it is possible to conclude from the intensity of the radiation whether, and in what concentration, that element is present. In this way it was possible to demonstrate in a simple way the presence in paper of traces of sulphur which could not possibly be detected by the methods of analytical chemistry, by exposing the paper to the intense neutron radiation of a cyclotron and afterwards demonstrating the presence of a radioactive transmutation product of sulphur. The tracer method here appears in a rudimentary form: with the preparation and detection of the tagged atoms their role is finished, and they need no longer take part in any process.

We can be brief about the preparation of stable isotopes as tracers. Of the elements occurring in nature many consist of different stable isotopes which are mixed in fixed proportions. Nitrogen, for example, consists for 99.6% of N^{14} and for 0.4% of N^{15}. These components can be separated by making use of the difference in velocity of diffusion or reaction, which differences are based upon the difference in atomic weight. Although when the mixture is subjected for instance to a diffusion process the ratio between the isotopes at the end of the process will differ by only a few percent from that at the beginning [3]), if the mixture thus obtained is again subjected to the same process and this is repeated many times, very large deviations from the original ratio of isotopes can finally result. This will occur more quickly the greater the relative difference in mass, thus the lighter the element. Heavy water, which can be concentrated by the electrolysis of water, since the normal hydrogen disappears from the water at the cathode more quickly than the heavy hydrogen (originally present in a ratio of 1 : 5 000 to the normal hydrogen), is obtained in a purity of 96% or higher after only a few steps.

The stable isotopes $_8O^{18}$, $_7N^{15}$, and heavy water are at present obtainable commercially. Heavy water was previously very expensive: a preparation with 96% D_2O cost several times as much as its own weight in gold. The situation has improved very much because of the fact that now, for instance

[2]) See for example F. A. Heyn and A. Bouwers, An Apparatus for the Transmutation of Atomic Nuclei, Philips Techn. Rev. **6**, 46, 1941. On the cyclotron see for instance W. B. Mann, The Cyclotron, Methuen's Monographs, London 1940.

[3]) The approximate constancy of this ratio of isotopes is in many cases an essential condition for the usefulness of the tracer method; all the quantitative investigations according to this method are based upon the fact that the spatial distribution of the one isotope ascertained at the end of a process is a faithful picture of the whole preparation or of the whole amount of that substance which was made to take part in the process. We shall return to this later.

in Norway, D_2O is obtained on a large scale as a by-product in the preparation of the hydrogen needed for the synthesis of ammonia; that process is carried out electrically in that country, where electrical energy is very cheap. The oxygen isotope ${}_8O^{18}$ is prepared by fractional distillation of water, making use of the fact that H_2O^{18} is less votatile than H_2O^{16}. The nitrogen isotope ${}_7N^{15}$ is obtained by a chemical process repeated many times.

The synthesis of tagged compounds

In many cases one works directly with tagged elements. Very often, however, experiments are made with compounds of those elements, which means that one actually needs "tagged compounds". How are such compounds obtained?

It is usually a question of compounds which in their untagged form could be purchased by the investigator from the chemical industry. If this is not the case, as starting materials for the synthesis he could at least choose compounds obtainable commercially which are as close as possible to the desired product and which can be converted into the latter by means of simple reactions. The tagged compounds, however, will usually have to be prepared by the investigator himself, and this will sometimes be made more difficult by the fact that he cannot start with a closely related compound but will be compelled to begin with the tagged element as such, or perhaps from the compound in which the isotope is formed in its preparation. In the case of some radioactive isotopes, such as C^{11}, there is the further difficulty of the limited lifetime, which makes it necessary to carry out the synthesis very rapidly. If, owing to such difficulties, the investigator has no success or the normal methods of synthesis are too laborious, the solution of the problem may perhaps be found in the excellent method of biological synthesis. The more or less complicated organic substances needed for many physiological investigations often occur as products of the metabolism of animals or of certain micro-organisms. If, for example, glycerine is administered to a culture of B pentosaceum (a kind of bacteria), together with carbon dioxide with radioactive carbon, within 30 minutes 80% of that carbon is obtained in the form of proprionic acid and succinic acid. With other micro-organisms other compounds can be made.

A method which can be used in many cases is to allow animals to drink "heavy water" for a time, or to water plants with heavy water. After some time numerous compounds can be isolated from the tissues in which a part of the normal hydrogen occurring in them has been replaced by heavy hydrogen.

A similar case of biological synthesis may be seen in the preparation of lecithin (a group of fats containing phosphorus) which is tagged by means of radioactive phosphorus. Interesting experiments have been performed with this substance, which has been recently much advertised as a "nerve tonic". When radioactive lecithin was fed to a rat it was found that after some time the blood of the rat became radioactive; but the radioactivity could not be localized in the lecithin which is always present in the blood plasma, but in the inorganic phosphate which is also present in the plasma. This means that the lecithin is not taken up as such through the wall of the intestines to reach the blood and organs in that way, but that the lecithin is split up in the intestines into its components. Thus, instead of taking lecithin, the same result can be obtained by consuming a mixture of fats and inorganic phosphates in the correct proportions. The body itself then builds up the lecithin. Use was made of this very fact in order to obtain the radioactive lecithin required for these tests; radioactive phosphate is injected into a hen and several days later radioactive lecithin can be isolated from the yolks of the eggs laid by the hen.

The analysis of the mixtures of isotopes

In the case of radioactive isotopes the analysis is performed by measuring the radioactivity. Since practically all the radioactive isotopes considered emit negative or positive electrons, the measurement can be performed with a so-called electron counter [4]). This instrument contains a gas-discharge tube made conductive for a very short time by an electron entering it. Thereby a current impulse is excited in the electric circuit in which the tube is connected. These impulses are amplified and counted by a counting apparatus. The number of impulses per second is a direct measure of the radioactivity of the substance placed in the vicinity of the counting tube.

It is clear that the value of the activity thus measured will still depend very much on the geometrical configuration, *i.e.* on the shape and size of the prepared object and on the relative distance and position of object and counter. It is always sufficient,

[4]) See for example A. Bouwers and F. A. Heyn, A Simple Instrument for Counting Electrons, Philips Techn. Rev. 6, 74, 1941. In addition to electron counters ionization chambers with electroscopes are often used for measuring radioactivity. These instruments are simpler but less sensitive. The following discussion is equally valid for all measuring arrangements, so that we need only speak about the electron counter.

however, to perform relative measurements: it is a question of comparing the radioactivity of two substances or of one substance at different times. For this it is only necessary to provide for satisfactory reproducibility. In the most common case where one is concerned with preparations in powder form, it is customary to spread out the powders in a certain thickness, in a dish of certain dimensions and then to place the dish in a well defined position with respect to the counter (*fig. 2*).

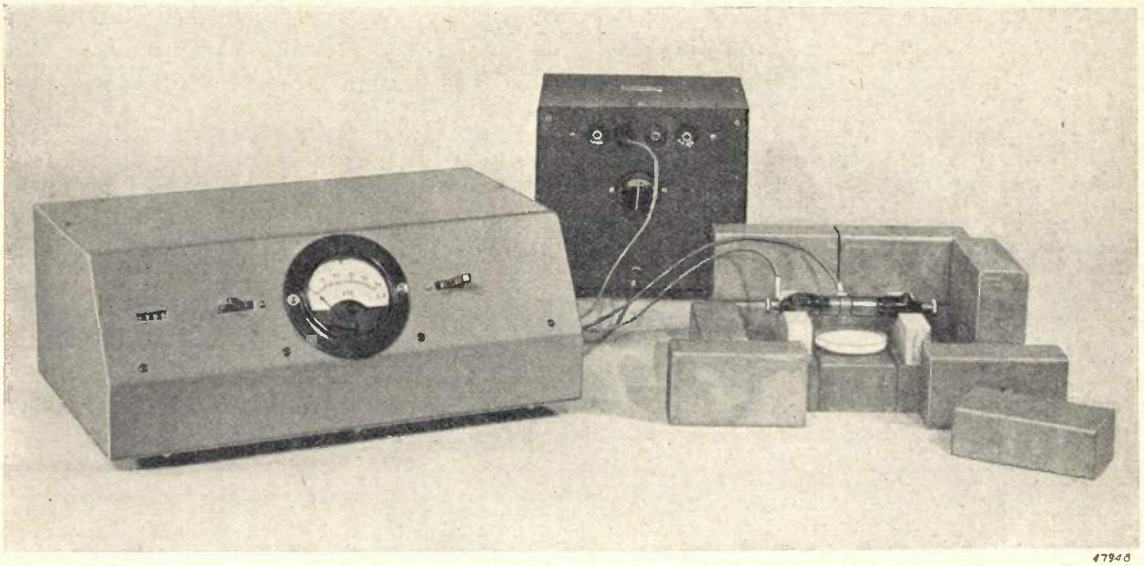


Fig. 2. Arrangement for the measurement of radio-activity with an electron counter. Under the counter tube is a dish in which the substance to be investigated is spread out (right). The two poles of the counter tube are connected with the counting apparatus seen on the left of the photo. In the middle background may be seen the feeding apparatus. Even in the absence of a radioactive substance the counter records a certain number of impulses per minute which are caused by cosmic rays and by the natural radioactivity of the surroundings. In order to diminish this so-called zero effect the counter tube with the substance is surrounded on all sides by lead blocks, some of which can be seen in the photo.

If it is only a question of localizing the radioactive substance, as in the case of the scraper stuck fast in an oil pipe line — mentioned as an example in the preceding article [1]) — or as in experiments for the assimilation of phosphorus or potassium out of the soil in different parts of a plant, the pipe line or the plant can simply be gone over with the counter. More precise localization is possible by the method of the auto-radiogram also previously mentioned. The assimilation of zinc in the fruits of a tomato plant, for example, has been studied by cutting through a fruit and laying the cut surface on a photographic plate. The spots where the fruit contained radioactive zinc are indicated on the developed plate by a blackening. The tomato was found to assimilate the zinc mainly in the seeds (*fig. 3*).

For more quantitative work in experiments like the above the different parts of the plant are reduced by combustion, so that the activity of each part is concentrated in a very small volume. The activity of these different preparations can be compared by placing the same quantity in each case under the counter. For comparison with the original (very concentrated) preparation of the nutritive substance, the latter is made up, if necessary, to the required amount by the addition of inactive ash in order to obtain the same configuration for the measurement.

One complication generally occurring in the analysis of radioactive preparations is the natural decrease in the activity of a substance with time. When it is desired to calculate from the measurement of activity the fraction of the total amount of radioactive substance which has entered into the preparation to be analyzed, a factor depending upon the time should enter into the calculation. This, however, is avoided in a very simple way when the activities measured are always expressed as a percentage of the activity of a portion of the original preparation which has not been used in the experiment. Since the activity of the latter decreases at the same rate, the time factor is automatically eliminated.

The sensitivity that can be obtained with the electron counter is extremely high. With ordinary counters the presence of an amount of 10^5 radioactive atoms can be demonstrated, *i.e.* 10^{-18} to

10^{-19} gram molecules. The limit is set by the statistical fluctuations of the number of atoms disintegrating per unit of time and the fluctuations of the zero effect (see the text under fig. 2).

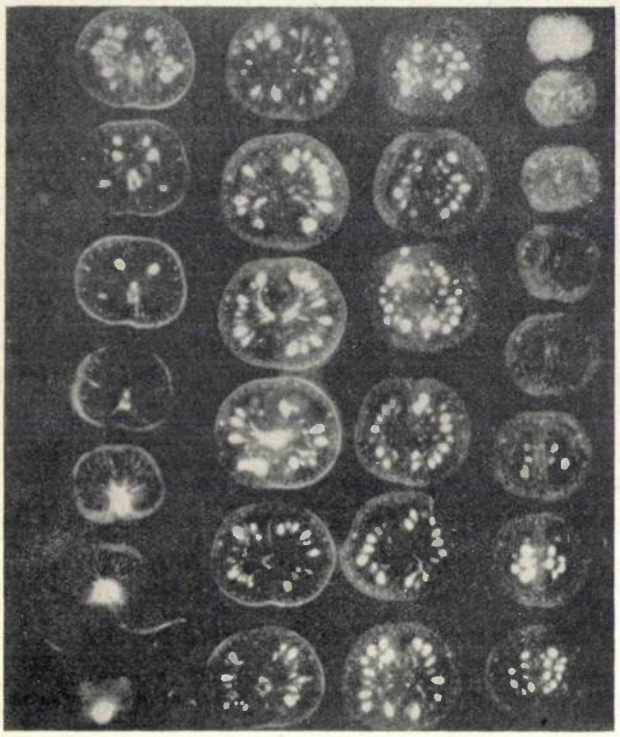


Fig. 3. Auto-radiogram of a cut surface of a tomato. The plant had been nourished for some time with radioactive zinc, which for the greater part was accumulated in the seeds.

With the preparations used in the tracer method the concentration of the radioactive isotope is usually not more than $1 : 10^{10}$ (see the article referred to in footnote [1]): it is then possible to demonstrate the presence of 10^{-9} gram molecules of the preparation. With particularly sensitive counters and very concentrated preparations it has been possible to demonstrate the presence of 2×10^{-12} gram molecules of copper, *i.e.* an amount of 10^{-10} gram.

According to the attainable reproducibility of the geometrical configuration, the accuracy of the radioactivity measurements with the counter varies between about 0.1% and 5%. In the neighbourhood of the limit of sensitivity the accuracy naturally decreases. An accuracy of 10% means that for example in exchange experiments it is only possible to determine the amounts of the isotopes exchanged with a possible error of 10%. Sometimes, however, this is sufficient for the purpose in view. The difference in mass of two isotopes is also accompanied by a slight difference in diffusion and reaction velocity (see above under the preparation of stable isotopes), so that the ratio of isotopes in a physical or chemical exchange process does not always remain entirely constant, but may change slightly, sometimes even by several percent.

The analysis of mixtures of stable isotopes is a problem by itself, which we shall only touch upon lightly here, because these isotopes are of less importance for practical applications. The oldest and in modern form still much used method of analysis is the "mass spectrograph" method of Aston. In a gas-discharge tube ions of the mixture are formed and then deflected to a greater or lesser degree by an electrical and magnetic field according to their mass. This method has been developed to a high degree of accuracy, namely to errors of only 0.5% in the ratio of isotopes to be determined. A much easier method of analysis is that based on the measurement of density, which is used for example for deuterium, the heavy isotope of hydrogen, in the form of "heavy water" (D_2O). Since it is possible without too great difficulties to measure the density of water to an accuracy of 10^{-6} — in this connection it is also important that water can so easily be purified by distillation — it is possible to determine D_2O quantitatively in extremely great dilution in H_2O. If the deuterium forms part of other hydrogen compounds the same method of analysis can be applied by first burning the compound and concentrating attention on the water formed in the combustion. Sometimes only very small quantities are then available for the measurement of density, but this is no objection since there are methods of measurement in which 1/100 or even 1/1000 cm^3 of water is sufficient.

48111

SECRET PRODUCTION OF RADIO RECEIVERS IN OCCUPIED TERRITORY

by "ONE OUT OF MANY." 621.396.181.4"36"

During the German occupation the Dutch were strictly forbidden to listen to foreign broadcasts, but as long as they were left in possession of their wireless sets, few obeyed the orders. When, however, in 1943 the German authorities enforced their commands by confiscating wireless sets on a large scale, there arose a need everywhere for pocket-size receivers, which could either be hidden away in a small corner or camouflaged in an article of daily use. It was, of course, primarily the task of the Philips concern in Eindhoven to satisfy this demand, and though a normal and organized production was out of the question, personal initiative was not lacking. Soon a lively activity sprang up in small workshops and private rooms and the amateur set-making of earlier days underwent a temporary but intensive revival.

Official production figures are lacking and no one knows exactly what was achieved, but it is estimated that several thousands and perhaps tens of thousands of these pocket-size receivers were produced in Eindhoven. Though we can only draw on our own experience and on that of a small circle of friends we hope that the following notes will nevertheless give a fair idea of the difficulties and anxious moments of all who partook in this more or less risky venture.

Preferably "button" or "acorn" valves were used to keep the set as small as possible, and the so-called reflex circuit was very popular owing to the high sensitivity to be obtained from it with only two of these miniature valves. In this circuit the first tube serves both as high frequency and low frequency amplifier, while the second tube is used as grid detector with feedback; the low frequency signal from the anode of this last valve is impressed on the grid of the first valve. In this circuiting system the H.F. and L.F. parts must be carefully separated by filters to avoid self-oscillation.

Fig. 1 illustrates an example of such a circuit, which when using a headphone would give a satisfactory reception on an aerial of two or three meters stretched out in the room.

Since there was no space for a normal tuning condenser, so-called trimmers with a capacity varying from 3 to 30 $\mu\mu$F were used (*fig. 2*). With

well chosen coils the 30, 40 and 50-meter bands could then be covered, which left an ample choice of stations to be tuned into.

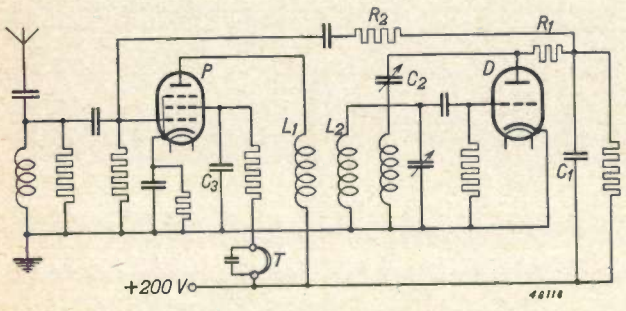


Fig. 1. A reflex circuit frequently used. The pentode P serves as h.f. and as l.f. amplifier; its anode is inductively coupled with the grid of the triode D which is employed as grid detector. From the anode of D the low frequency signal is fed back into the grid of P *via* a low-pass filter comprising the elements R_1, C_1 and R_2. The headphone is taken up in the lead to the screen grid of P, the condensor C_3 being inserted to avoid h.f. signals in the telephone leads.
The triode D has also been fitted with a normal h.f. feed-back which is regulated by the condenser C_2. A weak inductive coupling between the coils L_1 and L_2 precludes the transmission of l.f. signals from P to D.

Another circuit frequently adopted is illustrated in *fig. 3*. In this circuit the triode-hexode UCH 21 is employed, the triode being used as grid detector and the hexode as L.F. amplifier. This combination is made possible by the fact that the grid of the triode is not internally connected with one of the grids of the hexode.

Despite the wide choice of stations referred to above, listening in was often far from a pleasure owing to the many jamming stations set up by the Germans especially in the short wave range. Combined with the unavoidable fading these made reception difficult and frequently impossible, and in the course of time this induced many set-makers to change over to longer wavelengths. The smaller range of wavelengths then to be covered was compensated by a more reliable reception.

On the medium wave the British "Home Service" in particular was free from jamming for a long time, and on the 1500 meters the news could still be distinguished with sufficient clarity from a disturbing background. When, however, the station usually tuned into happened to change its wave length it became necessary to alter the coils and this was generally no easy matter with such minute home-made sets.

The circuits illustrated by figures 1 and 2 were specially designed for sets of very small dimensions. Those, however, who had given up their wireless to the German authorities, according to orders, and consequently did not live in constant fear of their house being searched, could take the risk of somewhat larger dimensions. In a collection recently exhibited to members of the B.B.C. staff visiting the Netherlands we observed a couple of receivers mounted on a sturdy brass chassis and fitted with interchangeable coils and two normal valves, the EF 6 and EBC 3, which reproduced various English programmes on a loudspeaker with an aerial of only one meter. We know that 15 of these

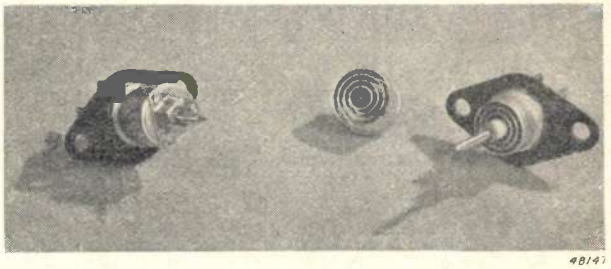


Fig. 2. A trimmer, ordinarily used for the correction of capacities in tuned circuits, is now employed as tuning condenser. The construction shown above is that of a sliding condenser, one electrode being moved with respect to the other by a screw movement; the capacity ranges from 3 to 30 $\mu\mu$F.

sets were made by one person single handed, some of which were supplied to the underground movement. And this is by no means to be considered as a record.

Practically all pocket-size receivers were made for direct use on the 220 volt AC mains, the anode

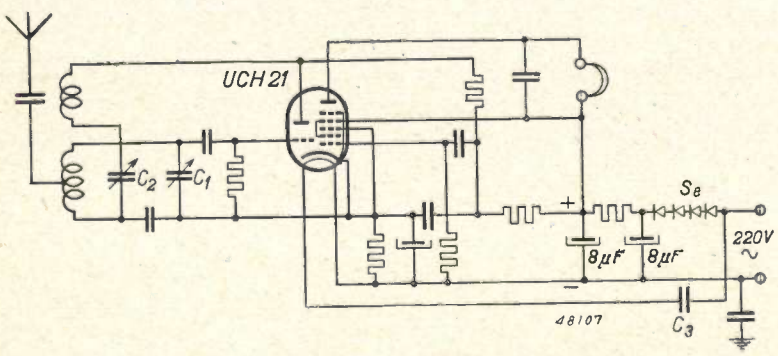


Fig. 3. Another circuit of frequent application, combining grid detection and l.f. amplification in a single valve (UCH 21). A normal feed-back can be adjusted by means of C_2. To avoid the use of a transformer the heater is fed from the mains *via* the condenser C_3. The anode voltage is also taken from the mains after rectification by the selenium rectifier *Se*.

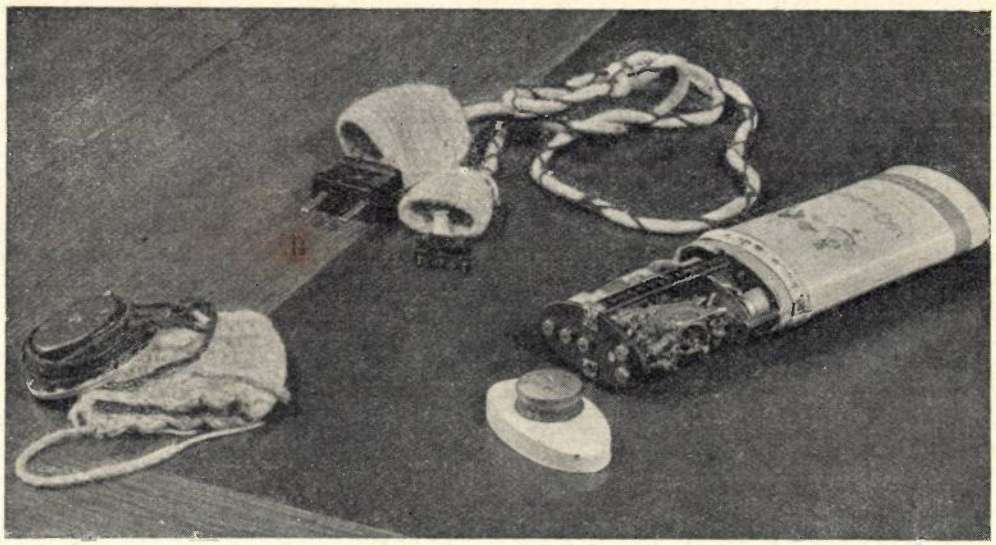


Fig. 4. A wireless set assembled in baby's powder duster. The connecting flex camouflaged by knitted covering was used as girdle on mother's dressing gown. The telephone covered in similar fashion was hung in the cradle as baby's rattle.

voltage being obtained by rectifying the mains voltage either with an "acorn" valve or with a selenium rectifier which could supply a maximum current of 10 mA. Smoothing was effected with an electrolytic condenser specially developed for this purpose comprising 2×8 μF in a cylindrical space 18 mm in diameter and 50 mm long.

The cathode heating current was in many cases furnished by a small transformer. Since, however, these were not always easy to come by, the heating current was in other cases taken directly from the mains, an incandescent lamp or better still a condenser being used as series impedance. A valve like the UCH 21, being designed for a small heater current and a high heater voltage, was specially suitable in this respect. Appropriate condensers were again constructed and manufactured on a fairly large scale solely for this purpose.

It should, however, be noted that the use of a transformer had one definite advantage in that it

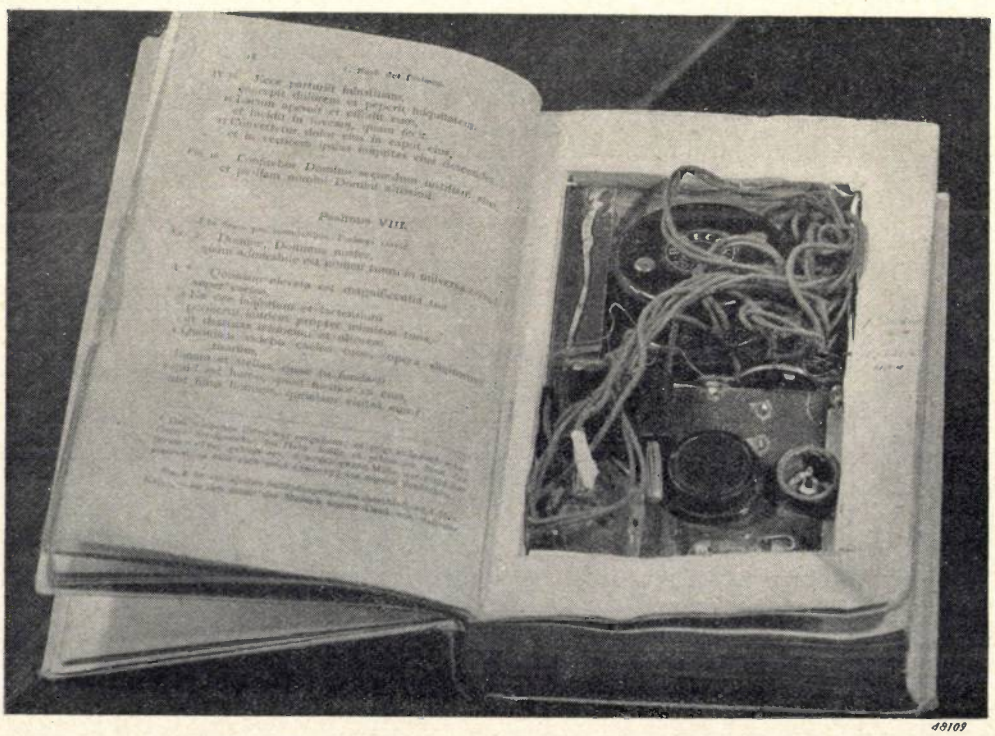


Fig. 5. A wireless set concealed in the cut-out pages of a book.

could also be fed from a bicycle dynamo, connected to the low voltage side. In this case the dynamo furnishes the 6 volts required for feeding the filaments and this low tension is stepped up by the transformer to the 220 volts needed for the anode voltage. Turning the bicycle upside down and mounting the dynamo on the rear wheel fork, the set could be operated by driving the pedals by hand. In this way many have been able to get the news even when for some reason or other the supply of electricity broke down.

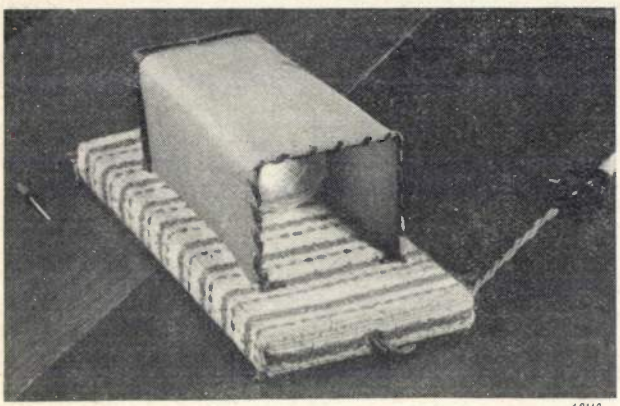


Fig. 6. The base of this lamp contains a superheterodyne receiver with 5 valves and a telephone serving as loud speaker. In addition to its obvious use as reading light the wireless could be put in operation by plugging a nail in an almost invisible hole. The spoke of a bicycle wheel serving as aerial could be slipped right back inside when not in use.

To collect the necessary components was not always an easy matter and frequently the construction of the set will have been prescribed by whatever materials happened to be available. Anyone having a stock of special parts under his custody would sooner or later be obliged to keep it safely under lock and key to prevent it from dwindling in a mysterious manner.

In finding a suitable camouflage individual ingenuity was given free rein. At an exhibition of the resistance movement we noted for instance a wireless receiver assembled inside the headlight of a bicycle lamp; as the glass was covered with the prescribed black-out screen nothing could be seen from outside. Other devices are illustrated in the picture heading this article, while in *figs. 4, 5 and 6* some original "solutions" are shown in greater detail.

The most urgent need for reliable news was, of course, felt by those imprisoned in German camps. In the spring of 1944 a Dutch prisoner of war in Neu Brandenburg succeeded in sending a request to Eindhoven for certain specified parts to be sent to him concealed in a food parcel. So much experience had by that time been gained that it could be decided to send him a complete set instead, camouflaged in a tin of vegetables. In this case the circuit of fig. 3 was adopted and after the set had been assembled the weight of the tin was made up to its proper value, care being taken that the centre of gravity coincided with its normal position. Furthermore the inside of the tin was provided with a special lining so that it would produce the correct sound on being tapped. When eventually a message came through that the "football boots (this was the prearranged code-word) had arrived in safety and were an excellent fit" our joy was great, as will easily be understood.

Needless to say this war-time venture had its dangerous side, and unfortunately some who took part forfeited their lives; in spite of that, however, the work went on, outwitting the enemy with cunning enterprise.

ELECTRICAL DETECTION OF TRACES OF POISONOUS GASES IN THE ATMOSPHERE

by J. BOEKE.

613.632.4 : 544.4

A catalyst of finely divided platinum which catalyzes an oxidation reaction is found to be hindered in its function by many gases which are also poisonous to the human body. Based on this fact, an instrument has been designed which can detect the occurrence of poisonous gases, for instance carbon monoxide in mines, garages or boiler-houses, hydro-cyanic acid gas in galvanizing shops, etc. The instrument contains a platinum catalyst on a platinum wire serving as one of the resistance branches in a Wheatstone bridge and along which a mixture of methylalcohol and atmospheric air is conducted. The heat developed by the oxidation of the alcohol keeps the platinum wire at a temperature of 120-150 °C, at which the bridge is brought into equilibrium. When the catalyst is poisoned the oxidation reaction ceases, the wire cools off and, due to its change in resistance, the bridge is thrown out of equilibrium, thereby setting an alarm arrangement into action. A simple portable model of the instrument is described.

Suffocation and poisoning

The energy for the vital processes in man and animal is furnished by "combustion" of the organic compounds in the food: an oxidation of the hydrogen and the carbon in these compounds by the oxygen in the air. For this oxidation it is necessary that oxygen can be inhaled from the surrounding air through the respiratory organs, eventually being carried by the blood to the various organs involved in the oxidation processes.

If, therefore, the air inhaled by an animal were replaced by a gas containing little or no oxygen, for instance only nitrogen or laughing gas (N_2O), the animal would die for lack of oxygen: it would suffocate, or some of its vital functions would cease; it would become unconscious.

Elihu Thomson showed in 1873, by experiments on the Thomson's family cat, that in the case just described the term suffocation must be used in contrast to the poisoning which occurs, for instance, due to the presence of carbon monoxide in the atmosphere. In the case of suffocation (at least if life is not already extinct), upon oxygen being administered, the organism quickly resumes its normal functions, whereas in the case of the poisoning mentioned the after-effect lasts longer and — in spite of renewed supply of oxygen — death may still set in after some time. In the case of such poisoning as this it is not the external supply of oxygen that ceases, but rather the transfer of the oxygen from the lungs to the compounds to be oxidized. The agents serving for this transfer are the haemoglobin (red pigment) in the blood — which together with oxygen forms a chemical compound, oxyhaemoglobine, that can easily be decomposed again and carries it in that form to the organs — and a number of specific substances called ferments or enzymes. The haemoglobin combines with carbon monoxide more readily than with oxygen, so that where carbon monoxide is present the oxygen-blood compound can no longer be formed and the supply of oxygen through the body stops. In a similar way the ferments in the body can be rendered inactive by certain gases in the air inhaled. Their function as carriers also ceases; they are "poisoned". Since the ferments are present only in very small amounts it is understandable that even only traces of a poisonous gas may be sufficient to block a special ferment function essential in the chain of metabolism of the organism, and thereby cause the entire organism to suffer a "death by starvation". Only 0.03 vol. % of the well-known prussic acid (hydrogen cyanide) in the atmosphere inhaled for a few moments is sufficient to kill an adult. It is less well known that hydrogen sulphide (given off by rotten eggs) is just as poisonous and in fact, because its action is cumulative, even more dangerous. Due to its extremely strong odour, however, we are warned in time and can escape from the poisonous atmosphere.

Unfortunately in many other cases of poisonous gases the nose does not function as an alarm instrument. Our sense of smell very often fails to react to gases which ordinarily do not occur free in nature but which are usually only formed by human agency: carbon monoxide is quite odourless for one, hydrogen cyanide for many. Even with respect to hydrogen sulphide, which does often occur in nature, the nose is not always reliable: after a relatively short time the sense of smell becomes so used to it that the concentration could be increased until a fatal dose is unsuspectingly inhaled.

In practice there are many cases where there is a

risk of coming into contact with these or other poisonous gases. For example, in garages or boiler-houses, where carbon monoxide may be formed owing to incomplete combustion; in mines where carbon monoxide may also occur; in painters' workshops where carbon monoxide is liberated in the drying of paint containing linseed oil; in the fumigation of ships' holds and of houses with hydrogen cyanide preparations, where traces may linger after the treatment; in galvanizing shops where hydrogen cyanide may be developed owing to incorrect handling of cyanide-containing baths, etc. In such cases it may often be desirable to have at one's disposal an objective alarm apparatus sufficiently sensitive to react even to traces of invisible, more or less odourless poisonous gases and vapours.

At first glance this might seem impossible of achievement. How can an instrument distinguish whether a gas is poisonous to human beings or not? The instrument shoud not, of course, react to non-poisonous gases.

A solution of this problem is made possible by the fact that the phenomenon of poisoning in the human body shows a far-reaching parallelism with the phenomenon of the poisoning of a platinum catalyst. An alarm arrangement based upon this will be described here [1]). For a good understanding of its working, however, it is necessary first to say something about the functioning of a catalyst.

Catalysts

Oxidation by transfer in addition to ordinary oxidation is also met with outside the living organism.

Hydrogen is an easily oxidizable gas which on being burnt gives off heat and forms water. At room temperature, however, no combustion takes place, due to the fact that between the two molecules of hydrogen and oxygen approaching each other there is an energy threshold preventing them from combining at room temperature. If the temperature is increased by the flame of a match the energy of motion of a number of molecules is increased sufficiently to cause them to pass over the energy threshold; the heat of combustion then liberated keeps the reaction going.

However, there is yet another possibility of bringing about the oxidation of hydrogen (and of compounds containing hydrogen), namely by means of a preparation of finely divided platinum, so-called platinum black, upon which both hydrogen and oxygen are absorbed. This phenomenon, which was discovered as early as 1823 by Döbereiner, may be explained as follows. In the adsorbed state the molecular structure of the gases is somewhat loosened by the fields of force of the platinum atoms, which means a lowering of the above mentioned energy threshold, so that the latter can already be exceeded at room temperature and reaction takes place. The platinum itself does not take part in the reaction, it only catalyses it. This may be expressed by saying that the oxygen, which in a free state does not react with free hydrogen, is carried over to the hydrogen by the platinum, acting in much the same way as the organic carriers, the ferments, enzymes, etc., which may also be considered as catalysts. Since the platinum does not take part in the reaction and thus is not consumed, and the adsorbed molecules after their mutual reaction leave the surface to make room for the adsorption of new molecules, a small amount of platinum can catalyse the reaction of any desired large amount of gas, as is the case with ferments.

Upon platinum black being heated to a high temperature its catalytic properties are lost. This can be explained by the assumption that the catalysis does not take place uniformly over the whole platinum surface, but at special "active" spots, for instance the edges and corners of the minute crystals of platinum on the surface [2]). This assumption agrees entirely with the explanation of catalysis by adsorption, since the platinum atoms at such spots are surrounded by fewer neighbouring atoms than in the plane face, and therefore have left over, as it were, forces of attraction to attract foreign molecules (see *fig. 1*). At a high temperature

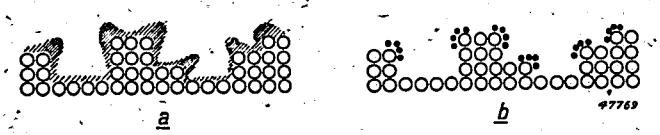


Fig. 1. Diagram of the surface of a catalyst with the fields of attraction shaded (*a*). Edges and corners of crystals in the surface form active spots where foreign molecules are preferentially adsorbed (*b*).

recrystallization takes place, the many small crystals combining to form a few large ones. The number of edges and corners is thereby considerably reduced, until the number of active spots is too small to make any catalytic action perceptible.

Döbereiner, who also knew of this phenomenon, noted that traces of hydrogen sulphide, hydrogen cyanide and carbon monoxide also made platinum black ineffective; the catalyst was "poisoned".

[1]) Netherlands patent no. 54716, applied for 23 Dec. 1939.

[2]) See the survey: E. K. Rideal and H. S. Taylor, Catalysis in Theory and Practice, Mac Millan & Co. London 1926.

The explanation is that the poisonous substances are very easily adsorbed on the platinum, their molecules occupying the active spots and staying there, thus blocking the catalytic function. Since the adsorption is a surface effect, it is also clear that a very small volume of the poison can make the catalyst ineffective, since every active spot need only be occupied by one or a few poison molecules.

For reactions in which different gases are involved, different catalysts are generally used, and the poisons also seem to be more or less specific: Hydrogen sulphide, a strong poison for platinum, is not poisonous for nickel sulphide catalysts, which are used technically in the preparation of synthetic petrol (Am. gasoline). Oxygen, a normal component in reactions accelerated by platinum, is a strong poison for the iron catalysts in the synthesis of ammonia, etc.

There is, however, a remarkably pronounced parallel between the poisoning properties of all kinds of substances for platinum and for the organic catalysts which play a part in the body. Let us compare, for instance, platinum with catalase, an enzyme occurring in most body liquids such as blood and milk. Both substances may function as accelerators of the reaction in the decomposition of hydrogen peroxide into water and oxygen. It has been studied how this catalytic function is hindered by certain "poisons"; in the table below [3]) the number of gram molecules of a series of poisons is given which is necessary to reduce to one half the velocity of the reaction mentioned for a given platinum or catalase preparation. It may be seen that platinum is poisoned by a whole series of substances which are also poisonous to catalase. Carbon monoxide is, it is true, not poisonous for catalase, but it is important that platinum is also sensitive to it, since, as has already been mentioned, carbon monoxide is poisonous for haemoglobin.

It is impossible to state exactly upon what the parallelism between the poisonousness for platinum and that for the organic catalysts is based. It may be noted, however, that they both have the property of catalysing reactions at low temperature where oxygen is involved. The inorganic catalysts which catalyse the same reactions at higher temperatures, such as certain oxides, are insensitive to many of the poisons listed, probably because of the fact that at higher temperature the poison molecules themselves begin to react and therefore do not continue to block the active centres.

Poison	Concentration of poison (in gram molecules) for a certain platinum preparation	Concentration of poison (in gram molecules) for a certain preparation of blood catalase
Sublimate, $HgCl_2$	1/2 000 000	1/2 000 000
Hydrogen sulphide, H_2S	1/300 000	1/1 000 000
Hydrocyanic acid, HCN	1/20 000 000	1/1 000 000
Mercuric Bromide, $HgBr_2$	—	1/300 000
Mercuric Cyanide, $Hg(CN)_2$	1/200 000	1/300 000
Iodine solution, I_2 in KI	1/50 000 000	1/50 000
Hydroxylamine hydrochloride, $NH_2OH.HCl$	1/25 000	1/80 000
Phenyl hydrazine, $C_6H_5\ NHNH_2$	—	1/20 000
Aniline, $C_6H_5NH_2$	1/50 000	1/400
Arsenic trioxide, As_2O_3	1/50	non-poisonous to 1/2000
Carbon monoxide, CO	very poisonous	non-poisonous
Hydrochloric acid, HCl	1/3000	1/100 000

Principle of the alarm apparatus

The fundamental principle of the alarm apparatus, which is based upon the above parallelism, is the following. An oxidation reaction is caused to take place on a platinum catalyst. Heat is thereby liberated. The oxygen for the oxidation is furnished by suction of atmospheric air. If the latter contains a poison (in the form of gas, vapour or mist) the catalyst loses its effect, the reaction stops and thus also the heat development. The check on poisonous substances in the air is thereby reduced to a check on the heat development.

As oxidation reaction we have chosen the oxidation of methylalcohol. A mixture of methyl alcohol vapour and air flows along a layer of platinum black deposited on a platinum wire with an intermediate layer of aluminium oxide. This wire forms one of the arms of a Wheatstone bridge (*fig. 2*). By the heat of oxidation released by

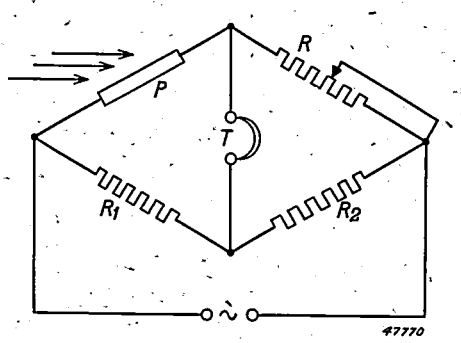


Fig. 2. Diagram showing the principle of the circuit of the alarm apparatus. The platinum wire P bears the catalyst (a layer of platinum black) along which a mixture of methylalcohol and atmospheric air is blown. R variable, $R_{1\,2}$ fixed bridge resistance, T headphone.

[3]) Borrowed from an article by W. D. Bancroft in the book referred to in footnote [2]).

the catalyst the temperature of the platinum wire and thus its resistance is increased. The bridge is brought into equilibrium at the temperature ultimately attained by the wire after the oxidation has proceeded for several moments (120-150 °C).

The presence of a poison will immediately cause the temperature of the catalyst and its carrier to fall, the equilibrium of the bridge is disturbed and a buzz is heard in the head phones shown in the diagram. Of course it is also possible to cause the disturbance of the equilibrium of the bridge to actuate a series of sirens or signal lamps by means of a relay.

After an alarm has been given the apparatus is prepared again for detection by "depoisoning" the catalyst. This is done by raising its temperature for a short time to about 400 °C, which evaporates the poison or causes it to disappear owing to a reaction with the mixture of methyl alcohol and air.

Activity and equilibrium temperature of the catalyst

In order that the catalyst should be hindered in its function by very small amounts of poison, it is desirable that the activity of the catalyst (*i.e.* the number of active spots) should be small compared with the amount of the reaction components passed over it per unit of time. This means that all the active spots available are continually occupied with the reaction. If this were not the case the blocking of a certain number of active spots due to occupation by poison molecules would not necessarily affect the total amount of reaction product formed per unit of time. The heat developed per unit of time would then also remain unchanged, and with certain conditions of cooling (by radiation and convection) the temperature, which is determined by the equilibrium between heat developed and heat dissipated, would not necessarily undergo any change either.

The activity of the catalyst must therefore expressly be reduced until it operates under full load. This is done by heating the catalyst wire for some time at about 700 °C. Obviously, however, one must not go beyond a certain limit, because an entirely inactive catalyst can no longer be neutralised by a poison.

This limitation can be expressed in another way. As the activity of the catalyst is reduced so the temperature at which it works must increase in order that the oxidation reaction may keep itself going. In the extreme case of zero activity, the working temperature must simply be equal to the ordinary flame temperature at which the oxidation can keep itself going without a catalyst. From this it follows that with increasing equilibrium temperature of the catalyst there can be less and less question of "poisoning": the parallelism with the sensitivity to poisons of the enzymes and ferments which act at blood temperature becomes less and less pronounced.

An equilibrium temperature of 120 to 150 °C was found to be the most suitable. In making this choice it must also be taken into account that the working temperature should not be too low, in order to ensure adequate disturbance of the bridge equilibrium when the catalyst becomes poisoned and cools down to room temperature, so that a clear indication can be obtained with a very simple and not highly sensitive apparatus.

The fact that we may count on a cooling to room temperature may be explained as follows. Upon poisoning, *i.e.* reduction in the number of active spots available, less heat will be produced by the reaction. On the other hand, as already stated, to maintain the reaction with a smaller number of active spots a higher temperature is necessary. From this it may be seen that upon a slight poisoning of a fully loaded catalyst no new equilibrium can be established; the temperature falls rapidly to room temperature.

The desired equilibrium temperature of 120-150 °C is realized by annealing the catalyst until with the given glow of the alcohol-air mixture it adjusts itself to that temperature. Of course the velocity of supply of the reaction components must then remain constant, in order that the equilibrium of the bridge shall not be disturbed already without any poisoning. The regulation of the velocity of flow is found not to be very critical, since the influence of the more rapid supply of the reaction mixture as the flow increases is approximately compensated by the deeper cooling then taking place. It is important, however, to keep the concentration of alcohol in the mixture very constant.

When the catalyst is put into action the oxidation reaction does not start spontaneously. The catalyst must first be heated artificially for a moment above the equilibrium temperature to be reached. This is done in the same way as the "depoisoning" previously mentioned.

Construction of a portable instrument

A portable alarm apparatus constructed according to the principles described is shown in *fig. 3*. In addition to great sensitivity, reliability and simplicity of operation, it was necessary — to make it portable — to aim at light weight and small energy consumption. The solution was found in the following details.

The catalyst wire is mounted in an exchangeable tube in the same way as the metal filament in an incandescent lamp.

The air to be tested is supplied by means of a diaphragm pump. This consists of a flat loudspeaker diaphragm with a cover in front of it with a blast opening. The diaphragm is vibrated by a kind of electric bell. This simple arrangement works as a pump with a capacity of about 7 liters air per minute.

Via a needle valve and a capillary methyl alcohol drips from a reservoir into a mixing chamber where the alcohol vapour is carried along by the current of air passing through. A fine wire in the capillary permits a fine regulation of the alcohol supply and a rapid cleaning of the capillary if it becomes stopped up. The mixing chamber is pre-heated by the exhausted oxidation gases; thanks to this measure the working temperature of the catalyst is little affected by temperature variations of the outside air.

In order to obtain a buzzing tone in the head-

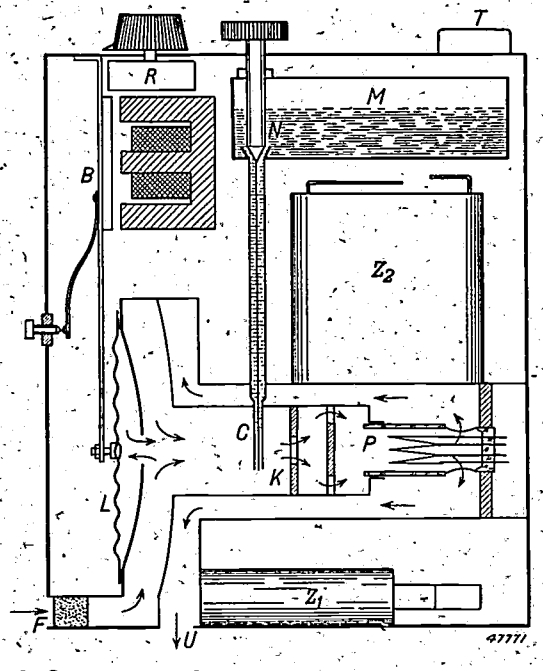


Fig. 3. Cross-section of a portable model of the alarm instrument. *P* exchangeable "lamp" with platinum catalyst, *M* reservoir with methylalcohol which, via the needle valve *N* and the capillary *C*, drips into the mixing chamber *K*. *B* electric bell which sets the loudspeaker diaphragm *L* vibrating by means of which the air entering, possibly via the filter *F*, is blown through the mixing chamber. The exhausted air leaves the apparatus at *U*. *R* variable resistance for balancing the bridge circuit. Z_1 pocket torch battery for normal operation, Z_2 battery for starting and depoisoning the catalyst, *T* connection for the headphone.

phone used as indicator the resistance bridge has to be fed with A.C. voltage or intermittent D.C. voltage. For this purpose the voltage across the winding of the electric bell system driving the pump is used.

The resistances in the arms of the bridge take up energy. Since the energy consumption must be a minimum, the winding of the bell system, in which in any case current must flow, is provided with a middle tap, and the two halves are used as fixed bridge resistances. Moreover with the help of a condenser the winding is brought into electrical resonance with the characteristic frequency of the interruptor of the bell system; see *fig. 4*. By these measures the total consumption of energy of the apparatus is reduced to 0.5 Watt, which is supplied by a pocket-torch battery.

When the variable resistance for setting the bridge in equilibrium is turned completely back, a switch connects the catalyst wire with another battery, so that it can be heated for a few seconds by 3 Watts to about 400 °C, for "starting" and "depoisoning".

In front of the openings through which the pump draws in air from the outside a filter can, if desired, be placed to block the entrance for certain gases. In this way it is possible to demonstrate the presence of certain gases selectively, for instance in garages where petrol vapour is alwasy present and where the instrument is not required to react to that but only to possible carbon monoxide; in this case, therefore, a filter must be used which absorbs petrol vapour but allows carbon monoxide to pass through. The fact that the catalyst is also poisoned by petrol vapours which are only slightly harmful for the body is one of the deviations from the above-mentioned parallelism. However, it is also possible to take advantage of this fact in order to detect petrol or other vapours in places where their presence would lead to the formation of explosive mixtures. An analogous possibility of application is based on the fact that the catalyst is sensitive to Freon and other voltatile halogenated hydrocarbons only slightly harmful for human beings.

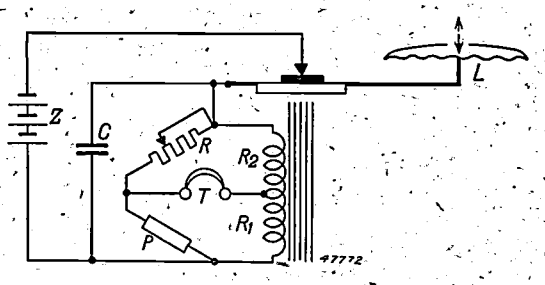


Fig. 4. More detailed diagram of the circuit of the apparatus. The two halves of the exciter coil of the bell system also serve as the fixed resistances R_1, R_2 of the bridge circuit. *Z* battery, *C* condenser, *L* diaphragm pump; other letters as in fig. 2.

These fairly expensive gases are used, for example, in refrigerating plants, and the instrument described would be very well suited for detecting leaks in such plants. It is clear, however, that such applications have nothing to do with the original purpose of the apparatus and that other specific actions or reactions of the gases in question could probably also be used for the purpose.

POTENTIAL DISTRIBUTION AT THE IGNITER OF A RELAY VALVE WITH MERCURY CATHODE

by N. WARMOLTZ.

621.316.546.032.43: 621.317.329

In order to obtain a better insight into the manner in which a cathode spot is formed on mercury by means of a dielectric igniter it was desired to know, among other things, the field strength on the surface of the mercury. This article describes how this was determined with the help of a one-hundred times enlarged model of an igniter in an electrolytic tank. The presence of two dielectrics between the mercury of the cathode and the conducting part of the igniter (namely vacuum and quartz) makes it necessary to use two electrolytes in the electrolytic tank. The liquid-tight partition between these electrolytes must be conductive in the direction perpendicular to its surface, but insulating in the horizontal direction along that surface. A simple solution is given for the problem of constructing such an anisotropic wall.

When assuming that the mercury surface is smooth the results of the measurement lead to field strengths which are too low for cold emission. If, however, it is assumed, with Tonks, that the mercury surface may be drawn out to small sharp points by the electric field, then a satisfactory explanation is found for the emission observed.

A form of discharge much used in technology is the arc. Its wide-spread use is due to its power of conducting large currents with relatively low voltage loss. This is one of the factors that have made it possible to construct rectifiers of high efficiency.

Arc discharges may be divided into several kinds. For rectifiers and related applications the following are mainly used:

1) the thermal arc, where the cathode is raised to such a temperature as to bring about a thermionic emission of the order of the arc current;
2) the field arc, where electrons are liberated from the cathode by an electric field — which must be of a strength at least several times 10^7 V/cm — set-up by the positive ions formed in the arc. By way of contrast with thermionic emission one usually speaks here of cold emission.

It is this cold emission with which we are concerned in mercury-cathode rectifiers. Here the cathode is a pool of mercury and the discharge emanates from the so-called cathode spot, *i.e.* a highly luminescent spot where there is a strong concentration of electrons and ions and where the current density is estimated to average 4000 A/cm² (estimations vary, however, from 1700 to 18 000 A/cm²). This spot moves about over the mercury in a most irregular manner.

Whether the mercury arc is a thermal or a field arc cannot well be decided from direct temperature measurements because of the many difficulties connected with such measurements in this case. The fact that it cannot be a thermal arc, however, is evident when it is realized that for thermionic emission of the current densities just mentioned a temperature of at least 3500 °C would be necessary; the occurrence of such temperatures must be considered out of the question because mercury very readily evaporates at a much lower temperature. From the current density and the amount of mercury that evaporates it follows, moreover, that the temperature of the cathode spot must be between 200 and 300 °C, at which temperature thermionic emission is quite negligible.

Methods of initiating a cathode spot

A cathode spot is not formed spontaneously. In order to make a mercury-cathode rectifier function, a cathode spot must be formed in some way or other. An old but still used method consists in sending an auxiliary current through the mercury cathode and an igniting electrode which reaches into the mercury and is then drawn out of it, for instance by an electromagnet or a bimetal relay. The auxiliary current is then interrupted and a cathode spot is found to have been formed. From this an arc discharge to the main anode(s) can develop.

Another method of ignition originated by Cooper-Hewitt dates from 1901. The glass wall of the vessel containing the cathode mercury is surrounded by a metal band at the level of the meniscus. When a voltage of about 10 kV is applied between the cathode and this "ignition band" — the latter must be positive — a spark occurs at the edge of the meniscus which can pass over into a

[1]) See for example M. J. Druyvesteyn and J. G. W. Mulder, The Physical Basis of Gas-filled Rectifiers with Hot Cathode, Philips Techn. Rev. 2, 122, 1937.

cathode spot [2]). This dielectric ignition, in contrast to the previously mentioned ignition by immersion with its mechanical inertia, has the advantage that the ignition can be repeated at a rapid rate and brought about at accurately determined moments. A discharge tube with dielectric ignition thus has the character of a relay valve [3]), *i.e.* a switch with practically no time lag. Moreover, thanks to the mercury cathode, it can pass very large currents without harm. These properties are of particular importance for switching devices serving to generate current impulses of variable length, interval and amplitude. The technique demands these properties, for instance, for electrical resistance welding [4]) and for stroboscopic purposes. Thus, for example, in the stroboscope [5]) already discussed in this periodical the

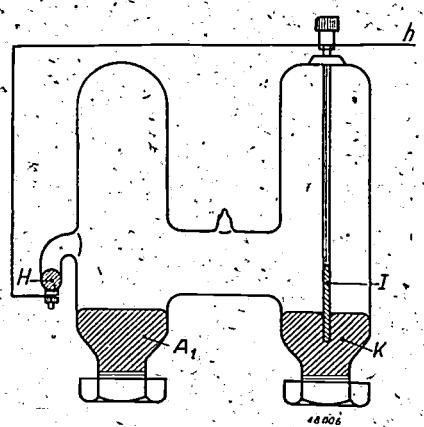


Fig. 1. Relay valve with dielectric igniter. A_1 main anode, K mercury cathode, I dielectric igniter, H auxiliary anode, h point to which a voltage impulse positive with respect to K must be applied to obtain a cathode spot.

tube sketched in *fig. 1* is used as switch for the lamp producing the light flashes. In this case current impulses of up to 2000 A are required, with a flash time of the order of 10^{-5} sec and a frequency of up to 250 flashes per sec. The ignition electrode of the tube shown in fig. 1 is an improved design of Cooper Hewitt's ignition band; it consists of a small quartz tube filled with mercury, which is shown in *fig. 2* enlarged and in a slightly different form. The action is the same as with the ignition band, the improvement consisting mainly in the

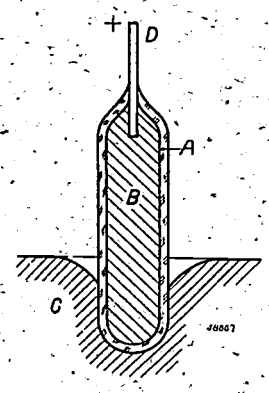


Fig. 2. Dielectric igniter. A insulating wall, B conductor, C mercury cathode, D voltage connction of the igniter.

greater durability of quartz. During the first few moments following the formation of the cathode spot an auxiliary electrode (fig. 1) connected to the same voltage source as the igniter S serves as anode for the discharge, which then develops further toward the main anode. The whole breakdown takes place within a very short time (of the order of 10^{-7} sec. calculated from the moment at which the cathode spot was formed).

When studying the properties of this igniter it was found desirable to know the distribution of potential in the space between the mercury surface and the outside wall of the igniter (respectively C and A in *fig. 3*). This is connected with the above-mentioned cold emission of electrons from

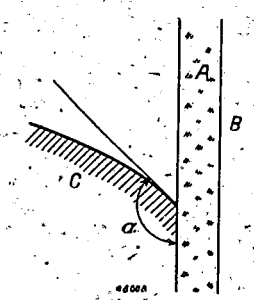


Fig. 3. Position of the mercury meniscus at the igniter. The letters A, B and C have the same significance as in fig. 2. α is the so-called limiting angle whose magnitude differs for different substances and which for the combination mercury-quartz was 136°30′.

the mercury surface. For this, of course, the electric field strength at the surface and the potential difference through which an electron passes on its path through the above-mentioned spacc are of great importance. These two quantities can be

---

[2]) N. Warmoltz, On the Mechanism of the Capacitative Igniter and of the Resistance Igniter in Mercury-Vapour Rectifiers, dissertation Delft 1946 (English summary).

[3]) There is also another group of relay valves, namely those with a control grid, in which the anode current can only flow when the control grid voltage exceeds a certain value. The cathode in this case may be either of mercury or it may be a hot cathode. Applications of relay valves with hot cathodes are described, for instance, in the article referred to in footnote [4]); in principe, however, relay valves with capacitive ignition can also be used there. We cannot here give a comparison of the two systems: relay valves with control grid and those with capacitative ignition.

[4]) D. M. Duinker. Relay Valves as Time Switches for spot welding, Philips Techn. Rev. 1, 11, 1936.

[5]) S. L. de Bruin, An Apparatus for Stroboscopic Observation, Philips Techn. Rev. 8, 25, 1946.

derived from the distribution of potential, which we have determined in the manner described below.

We did not try to calculate this distribution of potential because of the mathematical complexity of the problem. Measurements led more quickly to the goal, and in particular measurements carried out on a similarly shaped model in an electrolytic tank trough, especially since the necessary apparatus was for the main part already available [6]).

The electrolytic tank

Use is made of the fact that the distribution of potential in an electrostatic problem is identical with that in the corresponding conduction problem, that is to say it is identical with the case where the same electrodes with the same mutual potential differences are surrounded by conductors (preferably liquid, such as electrolytes) instead of dielectrics. The specific resistance of these conductors must be high compared with that of the electrodes. Both cases are governed by Laplace's equation and the same boundary conditions.

In a cross-section of the conduction model — which can often advantageously be constructed as a proportional enlargement of the original electrostatic case — the potential is measured at any desired point with the help of a moveable probing electrode. Providing the latter is small enough, it causes no appreciable disturbance of the original potential distribution.

In the practical application a tank of insulating material is used, filled with a weak electrolyte, tap water for instance. The model of the electrodes, made for instance of copper, is placed in this vessel, and voltages are applied between the electrodes (A.C. voltages of not too low frequency in order to avoid polarization at the electrodes). With the moveable probing electrode points are found where the potential has a certain value, ascertained by means of a bridge circuit. A stylus coupled with the probing electrode records the corresponding points on a sheet of paper; the geometrical position of these points is one of the desired equipotential lines.

Design of a tank with two electrolytes

In order to keep the model simple we have considered the distribution of potential in the case of plane igniters. Those actually used are in fact cylindrical, but the ignition mechanism is in principle the same in both shapes.

A possible form of a plane igniter is sketched in *fig. 4*. A model of this was constructed with a hundred-fold linear enlargement for the measurements in the electrolytic tank a top view of which is given diagrammatically in *fig. 5*. The conducting part of the igniter is represented by the flat plate *B*, the mercury surface by the plate *C* bent to correspond to the meniscus, both being of copper. The space indicated by *A* corresponds to the insulating covering of the igniter. Here, however, a complication arises which did not occur in the tank measurements described in the article referred to in footnote [6]).

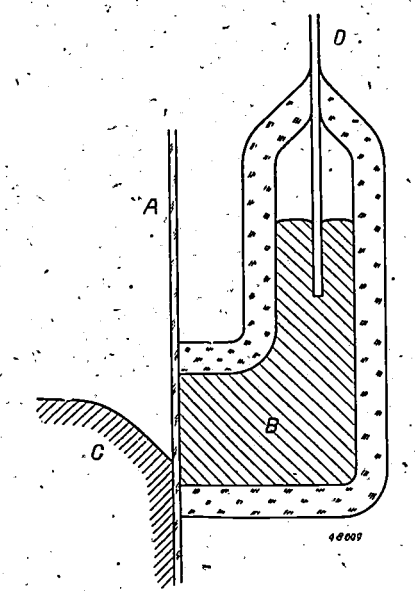


Fig. 4. Cross-section of a plane igniter. *A* quartz plate, on which the ingition occurs, *B* inner conductor (for instance mercury), *C* mercury cathode, *D* voltage connection.

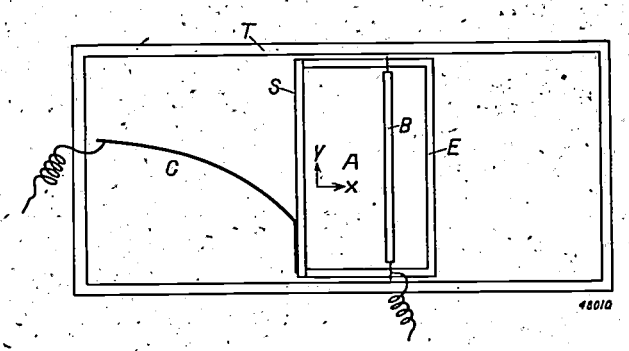


Fig. 5. View from above (diagrammatic) of an electrolytic tank (*T*) for measurement of the distribution of potential on a model of a dielectric igniter. The flat copper plate *B* and the bent copper plate *C* represent (enlarged 100 times) the analogen of the inner conductor and the mercury meniscus respectively (cf. fig. 4). The plate *B* is situated in a vessel *E* filled with a solution of copper sulphate; outside this vessel there is tap water. The space *A* corresponds to the quartz plate *A* of fig. 4. By parallel displacement of the plate the width of *A* can be *B* varied to correspond to different thickness of wall of the igniter. *S* is an anisotr copic wall conductive in the direction *X*, but not in the direction *Y* (see fig. 6). The probing electrode, which is not drawn, is moveable in the spaces *C-S* and *S-B*. The probing electrode is a copper wire of 0.1 mm diameter, projecting 1 to 2 mm out of an insulating covering.

[6]) G. Hepp, Measurements of Potential by means of the Electrolytic Tank, Philips Techn. Rev. 4, 223, 1939.

There, in the electrostatic case, there was only one dielectric (a vacuum), between the electrodes and this could therefore be represented in the tank model by one electrolyte. In our case, however, there are two dielectrics: the space between A and C (fig. 3), in which there is only mercury vapour, and the insulator A consisting of quartz or suchlike material. This means that at the corresponding places in the model different electrolytes must also be used, with conductivities in the same ratio as the dielectric constants of the insulators, *viz* 1 for vacuum (in this respect the mercury vapour present is without significance) and 4.4 for quartz. The place in the tank corresponding to the quartz (A in fig. 5) must therefore be filled with an electrolyte having a conductivity 4.4 times that of the place corresponding to the vacuum. We chose for the latter tap water and for the former a solution of copper sulphate. Since the two electrolytes must of course be kept separate, the copper sulphate is in a separate vessel.

When considering what material should be used for the wall S (fig. 5) of this vessel, one encounters the following peculiar difficulty. Metal cannot be used because conductive walls would form equipotential surfaces for which there is no analogy in the electrostatic case; this would lead to an entirely erroneous picture of the original distribution of potential. If, however, the whole vessel were made of non-conductive material, an entirely different potential distribution would again be obtained: both electrolytes would then function as coatings of a condenser, with the walls of the vessel as dielectric taking up practically the whole potential difference applied. What is desired is a construction in which the wall S (fig. 5) conducts in the direction X but not in the direction Y (its behaviour in the third direction is a matter of indifference), while the other walls must be insulating.

An obvious solution satisfying these apparently contradictory requirements is the following. A large number of metal pins are driven through an insulating plate in such a way that they project on both sides but make no contact with each other. These pins then give the desired conductivity in the X-direction, while no current can flow in directions perpendicular to that. To make such a wall, however, is rather laborious. A more practical solution has been found in the construction illustrated in *fig. 6*, consisting of a plate of insulating material surrounded by a series of spaced, U-shaped pieces of bare copper wire. The latter provide for the necessary conductivity in a direction perpendicular to the surface of the plate (X direction in fig. 5). Since the U pieces are spaced, there is no conductivity in the longitudinal (Y) direction.

It is only in the immediate vicinity of this anisotropic wall that the discontinuities of the "trellis-work" are manifested. Therefore no measurements were taken at distances less than the thickness of the wire of the spacing (both 0.5 mm).

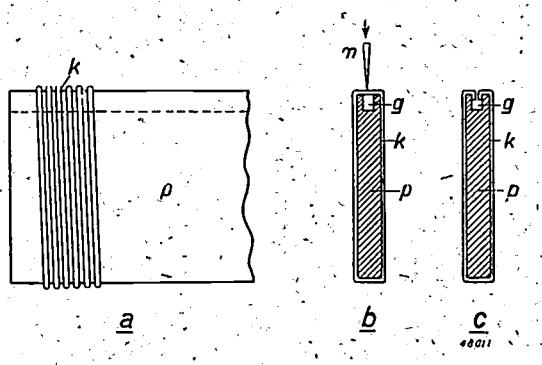


Fig. 6. Anisotropic wall separating the electrolytes (S in fig. 5). A plate p of insulating material is wound at intervals with bare copper wire k, fig. 6*a*. In the longitudinal direction there is a groove g, fig. 6*b*. With a knife m the windings are cut through above this groove and the ends thus formed are bent around the edge of the groove, fig. 6*c*. In this way a plate is formed which is surrounded by a series of spaced U-shaped pieces of bare wire.

The copper plate B (fig. 5) is moveable in the vessel containing the copper sulphate, so that measurements can be made with different widths of the space A (corresponding to the different thicknesses of quartz in the igniter).

Results of measurements

As a check, several potential distributions were first measured with tap water also in vessel E (fig. 5). It is then to be expected that the equipotential surfaces on the partitioning wall (S, fig. 5) will be continuous. *Fig. 7* shows the result of a measurement for two thicknesses of wall (0.35 and 3.0 mm, *i.e.* 3.5 and 30 cm in the model). As may be seen, by a parallel displacement over the thickness of the partitioning wall S (or rather slightly more than that thickness in order to exclude the region where the field disturbance of the individual "bars" is appreciable) the equipotential lines can be made to run almost without interruption.

Fig. 8 gives some of the actual measurements taken on the model of the igniter for three thicknesses of wall (0.35, 1.0 and 3.0 mm). The kink observed in the equipotential lines actually corresponds to the dielectric constant of the igniter wall of about 4.4.

---

[7]) This method has also been applied by R. Stachowiack, E.T.Z. **62**, 441, 1941.

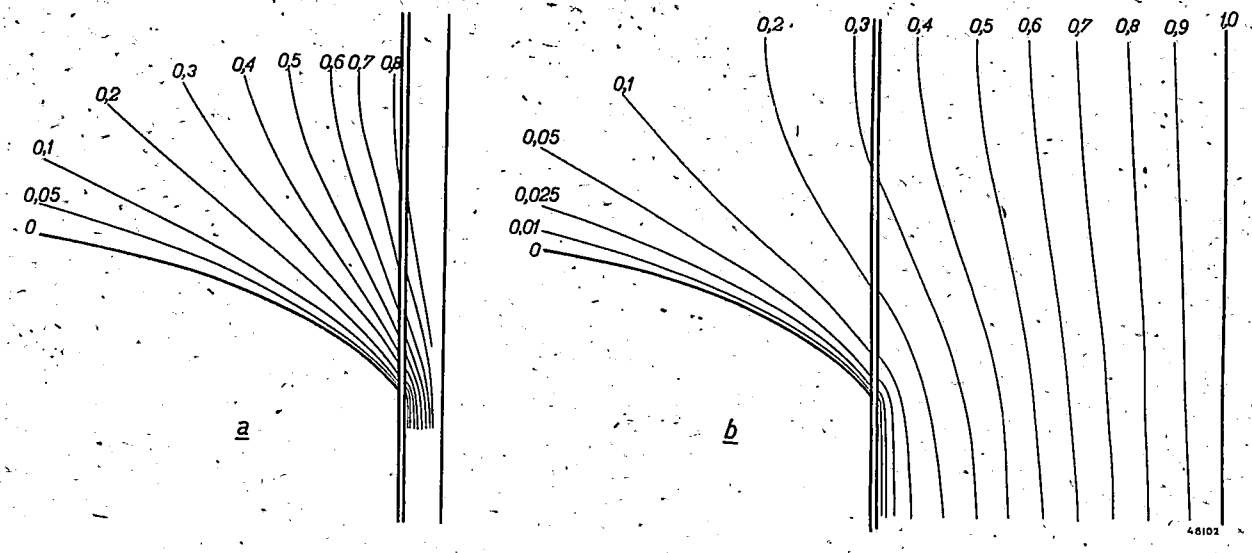


Fig. 7. Result of check measurements in which the vessel E (fig. 5) was also filled with tap water, to correspond to an imaginary igniter wall with a dielectric constant of 1 and a thickness of 0.35 mm (fig. 7*a*) and 3.0 mm respectively (fig. 7*b*). The curves are measured equipotential lines, the numbers indicating the corresponding potential (that of the electrodes C and A being set at 0 and 1 respectively). When the thickness of the separating wall S is imagined to be reduced to zero, the equipotential lines continue without interruption, as was to be expected.

The electrical lines of force undergo a change of direction at the boundary surface between two media of different dielectric constant (ε_1, ε_2), in such a way that the electrical field strength on either side of the boundary surface has the same tangential component, but that the normal components are inversely proportional to the dielectric constants. From this it follows that for the change of direction of the lines of force (*cf. fig. 9*)

$$\frac{\operatorname{tg}\beta_1}{\operatorname{tg}\beta_2} = \frac{N_1Q_1/P_1N_1}{N_2Q_2/P_2N_2} = \frac{1/P_1N_1}{1/P_2N_2} = \frac{\varepsilon_1}{\varepsilon_2}.$$

For the equipotential lines intersecting the lines of force at right angles, the following is therefore valid:

$$\frac{\operatorname{tg}\gamma_1}{\operatorname{tg}\gamma_2} = \frac{\operatorname{cotg}\beta_1}{\operatorname{cotg}\beta_2} = \frac{\varepsilon_2}{\varepsilon_1},$$

when γ_1 and γ_2 are the respective angles which the equipotential lines make with the normal to the boundary surface.

By drawing the lines of force in these figures and measuring along them the distances of the equipotential surfaces the electrical field strength can be determined. In order to find the field strength at the mercury surface several equipotential surfaces running close along the electrode C were accurately measured once more.

It was found, as is indeed well known, that the electrodes must be absolutely clean. If that is so, then the probing electrode takes up the potential of the electrode by just touching it. Otherwise there

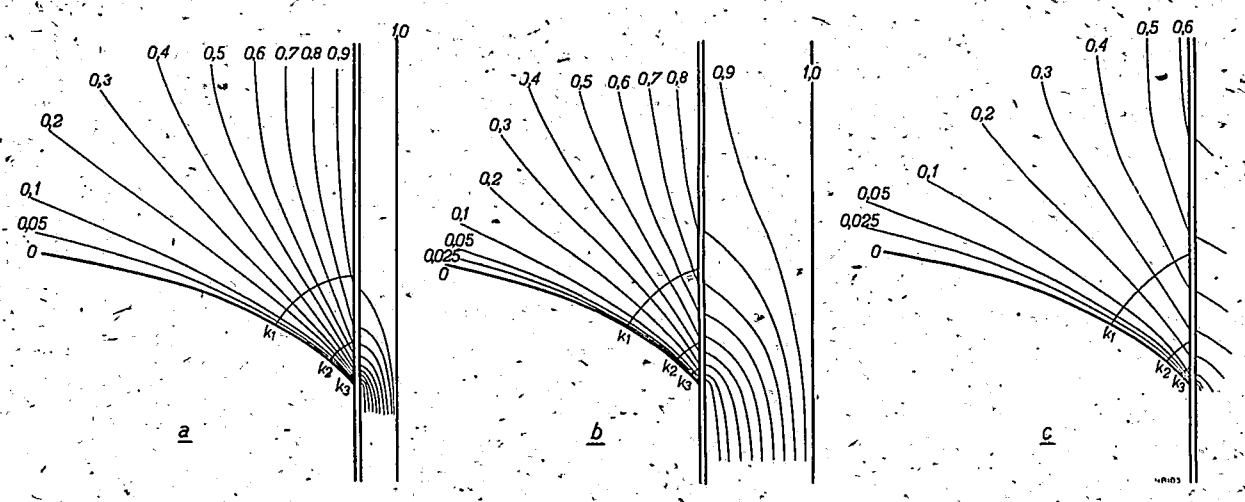


Fig. 8. Several measured distributions of potential on a plane igniter whose wall has a dielectric constant of 4.4 and a thickness (*a*) of 0.35 mm, (*b*) of 1.0 mm, and (*c*) of 3.0 mm; k_1, k_2 and k_3 are lines of force.

always remains a potential difference between that electrode and the probing electrode caused by lower conductivity of the surface layer. It was found necessary to clean the electrodes and the partition wall daily with concentrated sulphuric acid.

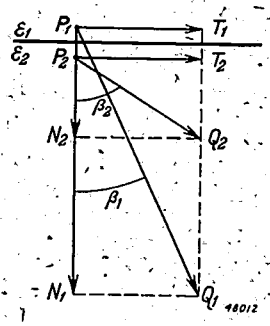


Fig. 9. P_1 and P_2 are points on either side of the boundary surface between media with different dielectric constants (ε_1, ε_2). The electrical field strength (P_1Q_1, P_2Q_2) at these points has the same tangential component ($P_1T_1 = P_2T_2$), but the normal components are inversely proportional to the dielectric constants ($P_1N_1 : P_2N_2 = \varepsilon_2 : \varepsilon_1$). From this follows the change in direction of the lines of force occurring at the boundary surface and also that of the equipotential lines intersecting the lines of force at right angles.

Fig. 10 shows some of the results of the determination of the field strength at the mercury surface with walls 0.35, 1.0 and 3.0 mm thick. Here $1/F$ (F expressed in kV/cm) is plotted as a function of the height above the boundary line between the mercury and the wall of the igniter. Since an infinitely high field strength is expected at this boundary line, due to the discontinuity caused by the limiting angle, the curve for $1/F$ is there drawn to zero.

Mechanism of the ignition

In order to explain this conception of the mechanism of the igniter we shall consider the process of ignition somewhat more closely.

It has been found by direct observation that the spark arises at a height above the mercury igniter wall which, with a given thickness and material of the wall, depends only on the voltage of the igniter. For example, with a wall thickness of 0.35 mm we found the following values for these two quantities:

3.15 kV	0.21 mm
3.55 kV	0.065 mm
4.5 kV	0.025 mm

At these voltages it follows from fig. 10, curve *a*, that the field strength at those places is: 94.5, 255 and 750 kV/cm.

The question now is whether or not cold electron emission can be expected with these field strengths. R. Haefer determined the necessary field strength with tungsten points examined ultramicroscopically for smoothness of surface [8]). In agreement with the theoretically derived values he found that for a current density of 45 A/cm² a field strength of 3×10^4 kV/cm is required, and that for 10^4 kV/cm only 3×10^{-16} A/cm² may be expected. The same applies to mercury, which has almost the same work function as that of tungsten. Consequently the field strengths of 10^2-10^3 kV/cm occurring in our case are unable to cause any appreciable cold emission.

The following two points now require an explanation:

1) How can cold emission occur at the place observed while the field strength deduced from the measurements is too low?
2) Why is it that the spark does not arise deeper in the gap between meniscus and quartz wall, where the field strength is higher?

The answer to the last question would lead us too far afield; suffice it to state that deeper in the gap some discharge can indeed be observed, but this cannot develop into a cathode spot, because the potential difference passed through by the electrons emitted is too small.

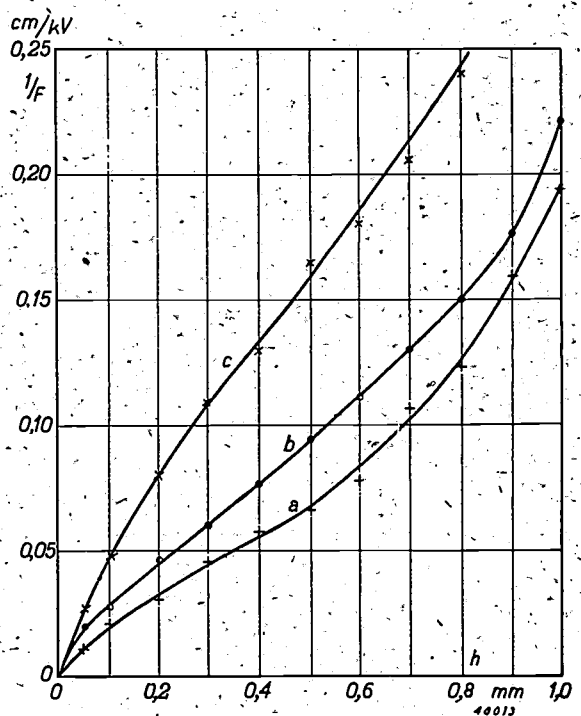


Fig. 10. Field strength F at the mercury surface at a voltage of 1 kV on the igniter. $1/F$ (F in kV/cm) is plotted as a function of the height h above the mercury-quartz boundary line on which the ignition takes place (in order better to extrapolate to small values of h, $1/F$ instead of F is plotted). The curves *a*, *b* and *c* are again valid for wall thicknesses of 0.35 mm, 1.0 mm and 3.0 mm respectively.

[8]) R. Haefer, Z. Phys. **116**, 604, 1940.

As to the first point, it has been found to be very probable, also on the basis of our further investigation of igniters, that a mechanism already considered theoretically by Tonks [9]) provides the explanation. The mercury surface, instead of being absolutely smooth, has microscopically small humps, which according to Tonks can be drawn out by the electric field to such sharp points as to produce a field strength sufficiently high for cold emission. The humps are caused by thermal agitation and in the absence of an electric field are kept relatively flat by the surface tension.

Briefly the whole process resolved itself into the following.

Upon voltage being applied to the igniter the electric field draws out the originally flat humps on the mercury surface to sharp points, where the field strength reaches values at which cold emission occurs. The emerging current of electrons flows during the first moments towards the wall of the igniter and then through ionization of the mercury vapour develops into an arc discharge.

It appears that the times necessary, according to Tonks' calculations, to draw out the humps into points, at the field strengths which follow from the above described measurements in the electrolytic tank, agree well with the ignition lags measured by us. The latter thus go to confirm Tonks' theory.

[9]) L. Tonks, Phys. Rev. 48, 562, 1935.

VOL. 8 No. 12, p. 353-384 DECEMBER 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

NON-METALLIC MAGNETIC MATERIAL FOR HIGH FREQUENCIES

by J. L. SNOEK.

621.318.323.2:538.245

It is a well-known phenomenon that when an a.c. field is generated in metallic coils eddy currents are set up in the core. The resultant loss of energy can be minimized, in the case of low frequencies, by using a laminated core with insulation between the plates. The higher the frequency, however, the finer the distribution of the core laminations has to be, and as a matter of fact for rather high frequencies cores are being used that are made up of wire and even grains (so-called dust cores); in the latter kind the lines of force run in part through the non-magnetic insulation of the grains, so that there is always a fairly large minimum "air-gap".

Entirely new methods have become possible as the result of researchwork with ferrites, substances of the type of MFe_2O_4, with M denoting a bivalent metal. Some of these ferrites have a cubic structure and are capable of forming mixed crystals in any proportions. With a few exceptions these non-metallic substances are ferromagnetic. Their specific resistivity is 10^7-10^{12} times as high as that of iron, so that the problem of eddy currents is relegated entirely to the background, even in the range of radiofrequencies and without laminations. By suitable composition and heat treatment it is also possible to keep the hysteresis and residual losses relatively low, whilst the initial permeability may be of the order of 1000. By means of a pressure or spraying process this material, called "Ferroxcube", can be moulded into various shapes, and coils of improved quality or smaller volume have been made with it for different purposes. Thanks to its being homogeneous it makes a good magnetic screen.

Metallic and non-metallic ferromagnetics

Magnetic properties were most probably discovered in non-metallic substances much sooner than in metals. Anyhow magnetite (magnetic iron ore, Fe_3O_4) was known in ancient times as a non-metallic ferromagnetic mineral.

It was not until the nineteenth century that any real technical use was made of these magnetic properties, except for compasses. And when, following upon the epoch-making discoveries of Faraday and others, an electrotechnical industry sprang up and a demand arose for suitable magnetic materials, it was not these minerals that were sought but rather iron or alloys with iron as basic component; magnetic oxides were considered to be nothing more than a scientific curiosity.

When it came to using metallic cores in magnetic alternating fields one very soon came up against the phenomenon of eddy currents discovered by Foucault, and the losses due thereto. In order to reduce these losses one proceeded to laminate the cores; this offered the additional advantage that the plate-shaped material could be properly purified by annealing, thereby improving its magnetic properties. By simple deduction one can arrive at the following formula for the eddy current losses W_e of a laminated core (assuming the magnetic induction to be equal at all points of a core section):

$$W_e = \frac{4}{3}(af B_{\max})^2 \cdot \frac{d^2}{\varrho} \cdot V \cdot 10^{-16} \text{ watt} \quad . \quad (1)$$

Here f is the frequency in c/sec., $B_{\max}$ the amplitude of the magnetic alternating induction in gauss, d the thickness in cm, ϱ the specific resistance of the core material in ohm. cm, V the volume of iron in cm^3 and a the so-called form factor, which in the case of an induction sinusoidally variable with time amounts to $\pi/2\sqrt{2} = 1.11$.

This formula shows that there are only two ways of reducing eddy current losses for a given frequency and induction, viz:

1) by reducing the plate thickness (d);
2) by increasing the specific resistance (s).

The thinner the sheet is rolled from which the core plates are punched or cut out, the better it will be purified by the subsequent annealing, but on the other hand this means increased cost of manufacture, more time in stacking the plates to form a core of certain dimensions and more loss of space owing to the necessary insulation between the laminations. As a result it is found that there is a certain limit for the optimum plate thickness, which for the frequencies of the electric light and power mains (40-60 c/sec) lies between 0.5 and 0.3 mm.

To a certain extent the building-up of cores out of laminated material has its attractions for the coil-maker, because cores of different sizes can be made with identical parts simply by varying the number of plates going to make up the stack, but this method costs time, a factor that counts particularly in mass manufacture. There is no doubt, therefore, that if it were possible from the point of view of eddy current losses, it would in many cases be preferable to employ solid cores such as can now be made in sufficient purity by the modern methods of refined melting.

As to the second method of reducing eddy current losses, by increasing the specific resistance, in spite of extensive metallurgical research it has not been possible to find magnetic iron alloys with a specific resistance higher than about five times that of pure iron, which is 10^{-5} ohm. cm. Although even this figure was a considerable step forward, it was by no means sufficient to satisfy the requirements of modern de elopments in telephone and radio technology with their so much higher frequencies accentuating the problem of eddy current losses.

Consequently — still holding fast to metallic cores — one was obliged to go farther in the direction of a finer distribution of the core material, and as a result so-called dust cores were introduced, consisting of fine grains of metal enclosed in an insulating envelope and compressed into a conglomerate of the greatest possible density. The cost of manufacturing such a finely divided powder is only a fraction of that of laminations of a thickness equalling the diameter of the grain. To draw a wire of that diameter would be still more expensive.

It was, of course, realized right from the outset that the introduction of powder cores involved a considerable loss of permeability, since the numerous granules are separated by insulating layers acting magnetically as "air-gaps". Every endeavour has been made to minimize this drawback, and not without some degree of success, as will appear from what follows.

Loading coils with a core composed of metal granules 50 microns and less in diameter have been developed for frequencies up to 3000 c/sec, and with an effective permeability of the core material as high as 125. From this it can be calculated that less than $1/125 = 0.008$ of the total length of the lines of force is occupied by non-magnetic material, so that the thickness of the insulating envelope around the grains is less than $0.008 \times 50 = 0.4$ micron (presumably about 0.2 micron). To attain such a result as this an extremely refined technique had to be developed for applying such thin layers to the grains, and it is obvious that any further appreciable improvement in this direction can hardly be expected.

In order to limit eddy current losses sufficiently for the so much higher frequencies used in radiotechnics, the size of grain had to be very drastically reduced, thereby taking into the bargain, of course, a further diminution of effective permeability.

It is not surprising that in the course of these developments the idea frequently occurred to try a different tack altogether and to replace the iron or nickel alloys by non-metallic substances, such as certain magnetic iron oxides having a very high specific resistance. In fact there are two patent specifications of Hilpert dating from 1909 concerning the use of ferrites for ferromagnetic coil cores for high frequencies. The specific resistances of 10^5-10^7 ohm cm. quoted as the maximum values for these ferrites contrast strongly with those of the usual metallic core materials, for which only with great difficulty a figure of $5 \cdot 10^{-5}$ ohm cm. has been reached. The gain to an amount of a factor 10^{10}-10^{12} when using ferrites is quite sufficient to relegate the problem of eddy current losses entirely to the background, but notwithstanding this fact the results do not seem to have been encouraging; at least nothing further has been reported in that respect.

In 1933 Philips laboratories started an investigation into the magnetic properties of ferrites [1]), and it was found that the ferrites mentioned by Hilpert do indeed have rather unsatisfactory properties, *viz*:

1) low initial permeability (not much over 10),
2) inconsistency in results,
3) still rather high total losses notwithstanding the negligible eddy current losses.

Nevertheless Philips' investigations were continued and it has now in fact been found possible to make ferrites with favourable magnetic pro-

[1]) See J. L. Snoek: New developments in magnetic materials. Elsevier Publ. Co., New York and Amsterdam, 1946.

perties. But first of all let us consider what ferrites actually are, so as to make it clear presently in what direction this solution had to be sought.

What are ferrites?

Ferrites are understood to be substances answering to the formula MFe_2O_4, in which M represents a bivalent metal. Ferrous-ferrite Fe_3O_4 (the magnetite previously mentioned), cupri-ferrite, $CuFe_2O_4$, lead-ferrite, $PbFe_2O_4$, etc. are already known, and in the beginning of this century these substances were regarded rather as the salts of an acid, $HFeO_2$, and bivalent metals.

Some of these ferrites are cubic in structure, a point of great importance in view of the fact that a high temperature, 1000-1400 °C, is required for their manufacture, to which we shall revert later. It is in fact only with the cubic structure that shrinkage during cooling is equal in various directions, so that with a substances of a cubic structure cooling need not cause internal stresses. At the time when this investigation into the properties of magnetic oxides was begun, it had just been made apparent from the theories of Kussmann, Becker and Kersten that the absence of stresses was an essential condition for high initial permeability and low hysteresis losses. Consequently investigations were confined to ferrites of the cubic system.

Conditions for obtaining ferrites having favourable properties

Condition for a cubic structure

According to Goldschmidt's crystallographic research the crystal structure depends, *inter alia*, upon the diameters of the ions, and this is borne out by the fact that from cubic Fe_3O_4 other likewise cubic ferrites can be derived when the bivalent ferro-ion is replaced by a bivalent ion of approximately the same diameter as the ferro-ion. This essential condition is not always sufficient, however, to guarantee a structure that is stable also at room temperature. Cupri-ferrite ($CuFe_2O_4$), for instance, has the tendency upon cooling to assume a tetragonal structure; though by rapid cooling from 900 °C downwards it is possible to keep this ferrite in its cubic form at room temperature. Such quenching, however, leads to internal stresses and possibly even to fracture, so that for our purpose cupri-ferrite is of no use.

A table of the radii of ions shows what bivalent ions can be considered for replacing the ferro-ion in Fe_3O_4, and below in *table I* the radii of these ions are given, together with the radius of the ferro-ion for comparison.

Table I

Metal	Radius of the bivalent ion [1]
Iron	0.83
Magnesium	0.78
Manganese	0.91
Cobalt	0.82
Nickel	0.78
Copper	0.83
Zinc	0.83
Cadmium	1.03

For so far as is known, the ferrites of the metals listed in this table form mixed crystals in all proportions, and these mixed crystals can also be considered for our purpose. The ferrites of the two last-named metals, zinc and cadmium, occupy a very special place, as will be explained below.

Details of the structure of ferrites [2])

As shown from a röntgenographical investigation by W. H. Bragg [3]), an elementary cell of a ferrite consists of eight molecules, this being expressed, in the case of ferro-ferrite for instance, by the formula $Fe_8^{2+}Fe_{16}^{3+}O_{32}^{2-}$, in which the charges of the atoms present in the form of ions is indicated at the same time.

Owing to their negative charge the oxygen ions are much larger than the positively charged metal ions. They form a cubic system of spheres. In a part of the interstices of this oxygen lattice are the metal ions, arranged in a manner characteristic of the structure. In ferrites these metal ions are arranged like the magnesium and aluminium ions in the mineral spinel, $MgAl_2O_4$, for which reason one speaks of the spinel structure. Now between the oxygen ions in this structure there appear to be two kinds of positions in which the metal ions may be found, namely so-called tetrahedral interstices, each bounded by four oxygen ions located at the corners of a tetrahedron, and so-called octahedral interstices bounded by eight oxygen ions forming an octahedron. In each elementary cell there are 8 tetrahedral — and 16 octahedral — places occupied by metal ions. In spinel it is most probable that the 8 tetrahedral spaces are all occupied by magnesium ions and the 16 octahedral spaces by aluminium ions.

It might then obviously be assumed in the case

---

[2]) A more extensive article on the structure of ferrites and allied compounds, which are also of electrotechnical interest, will be published shortly in this periodical.

[3]) Phil. Mag. 30, 305, 1915.

of the ferrites that the 8 bivalent metal ions occupy the 8 tetrahedral spaces and the 16 ferric ions the 16 octrahedral spaces, but a closer examination reveals that this is only the case with two of them, zinc ferrite, $ZuFe_2O_4$, and cadmium ferrite, $CdFe_2O_4$; in all other known ferrites the bivalent metal ions and the ferric ions are distributed among both kinds of interstices. This surprising fact is of great importance in more than one respect.

Condition for the occurrence of ferromagnetism; "Ferroxcube".

Zinc ferrite and cadmium ferrite, while being the only known simple ferrites in which the ferric ions are found exclusively in the octahedral interstices, are at the same time the only non-magnetic ones. Obviously, therefore, there must be some connection between these two facts, but it would take us beyond the scope of the present article to go more deeply into this, so that we will merely state that the presence of ferric ions in the tetrahedral spaces seems to be an essential condition for ferromagnetism.

As already remarked, the cubic ferrites in question can form mixed crystals in all proportions. It has been found that under favourable conditions mixed crystals of magnetic with non-magnetic ferrites — particularly zinc ferrite — may have very high values of initial permeability (μ_0). In *table II* the value of μ_0 is given for some simple ferrites, and in *table III* that of some mixed crystals obtained by adding zinc ferrite.

Table II

Ferrite	Initial permeability μ_0
$Fe\,Fe_2O_4$ (normal)	appr. 10
$Fe\,Fe_2O_4$ (stress-free)	appr. 70
$Cu\,Fe_2O_4$ (quenched)	appr. 70
$Mg\,Fe_2O_4$	max. appr. 10
$Ni\,Fe_2O_4$	max. appr. 10
$Co\,Fe_2O_4$	scarcely > 1
$Mn\,Fe_2O_4$	max. appr. 250 but inconsistent

Table III

Mixed crystal	Initial permeability μ_0
$Cu\,Fe_2O_4 + Zn\,Fe_2O_4$	appr. 1500
$Mg\,Fe_2O_4 + Zn\,Fe_2O_4$	appr. 700
$Mn\,Fe_2O_4 + Zn\,Fe_2O_4$	appr. 3000
$Ni\,Fe_2O_4 + Zn\,Fe_2O_4$	appr. 4000

From suitably chosen and treated mixed crystals Philips have manufactured ferrites which are being marketed under the trade name "Ferroxcube", indicating the cubic structure of these substances. By varying their composition or the heat treatment different grades, denoted as "Ferroxcube" I, II, III, etc., are obtained with widely divergent properties. The technical uses of these ferrites will be mentioned farther on.

Special measures to obtain high permeability

High permeability is obtained when there is no more than a slight interaction between the minute elementary magnets and the crystal lattice of the substance. The energy of interaction found in the undisturbed lattice is called crystal anisotropy, and this is known to assume very low values round about the Curie point, *i.e.* the temperature at which the ferromagnetism of a substance disappears and gives place to paramagnetism. This phenomenon is turned to account in the manufacture of "Ferroxcube", the Curie point being lowered, by the addition of the non-magnetic zinc ferrite, from several hundred degrees Celsius to a level approaching room temperature.

As a rule, however, the crystal lattice is disturbed by internal stresses, due in part to magneto-striction. By magneto-striction is understood the spontaneous distortion occurring when the material is magnetised; it is said to be positive when under increasing magnetisation elongation takes place and negative when shrinking occurs.

The materials most suitable for obtaining very high permeability are therefore those showing little magneto-striction, provided care is also taken to keep the internal stresses low (annealing at high temperature; homogeneous cubic structure). It has been found quite simple to get this low value of magneto-striction, especially in the case of "Ferroxcube" III, due in part to the fact that by mixing different ferrites both positive and negative values of magneto-striction are easily obtained. By mixing in the correct proportions magneto-striction can in fact be reduced to almost nil.

Condition for a high specific electric resistivity

Magnetite, Fe_3O_4, has the comparatively low specific electric resistivity of 10^{-1} ohm cm (at room temperature). Other ferrites in which the ferrous ion has been replaced by another bivalent metal ion have a specific resistance of the order of 10^5-10^7 ohm cm. In cases, however, where only a small part of these metal ions are ferrous ions a specific resistivity of 10^2-10^4 ohm cm. is found.

Consequently in order to reach the highest possible resistivity the presence of ferrous ions should be avoided. For all practical purposes it is not necessary that they should be entirely absent, because a value of say 100 ohm cm. is no objection for a core thickness of about 1 cm and frequencies up to about 10^6 c/sec.

The low specific resistivity of Fe_3O_4 has been ascribed to the fact that both ferrous and ferric ions appear in the same kind of interstices in the crystal lattice. Such an explanation is plausible if it is assumed that an electron is taken from a ferrous ion and added to a ferric ion, so that the former becomes ferric and the latter ferrous. Since both kinds of ions are in fact present in similar places such an interchange does not essentially alter the structure. This means that such a displacement of electrons may easily occur, or in other words there is a considerable conductivity.

Losses and quality of coils

Separation of the losses

In the foregoing we have seen that it is possible to compound ferrites having a high permeability and a high specific electric resistivity. Thanks to the latter property the eddy current losses will generally be small, but these are not the only losses occurring, as was apparent from the experiments with Hilpert's materials. And in nearly all the uses to which coils are put the total losses of energy in the coils have some influence upon the properties of the oscillatory circuit or the filter of which the coil forms a part. These losses, therefore, call for some further consideration.

The total losses in a coil through which an alternating current is flowing with an angular frequency $\omega = 2\pi f$ can be accounted for in a replacement diagram by assuming a resistance R to be connected in series with a self-induction L. This loss resistance comprises in the first place the direct current resistance of the coil increased by a factor due to "skin effect", and further a part representing the dielectric losses, whilst finally another part corresponds to the magnetic losses. Since, however, we are only concerned here with the last mentioned losses the other two parts of R can be left out of consideration, so that in the following R represents only the magnetic losses.

The magnetic losses are made up of three components: eddy current losses, hysteresis losscs and residual losses. It is possible that the last mentioned are related to the so-called after-effect phenomena [4]).

It is important to consider the ratio of R to the product of the permeability μ, the frequency f and the self-induction L. This ratio, the significance of which will be explained below, can be agreed with the above-mentioned division of the magnetic losses by writing it as follows:

$$\frac{R}{\mu f L} = c_h B_{\max} + c_e f + c_r(f) \quad . \quad . \quad . \quad (2)$$

Here the hysteresis losses are represented by the hysteresis constant c_h (a material constant) multiplied by the maximum induction $B_{\max}$, whilst the eddy current losses are represented by the product of the frequency f and a factor c_e depending not only on a material property (*e.g.* specific resistance) but also on the shape and dimensions of the core — for instance for a core with circular cross-section (diameter D cm).

$$c_e = \frac{\pi^3}{2} \cdot \frac{D^2}{\varrho} \cdot 10^{-9} \text{ sec.} \quad . \quad . \quad . \quad . \quad (3)$$

The residual losses are represented by the term $c_r(f)$, which depends upon the frequency.

Signification of the quantity $R/\mu fL$

The left-hand member of eq. (2) can also be written in the form:

$$\frac{R}{\mu f L} = \frac{2\pi}{\mu} \cdot \frac{R}{\omega L} = \frac{2\pi}{\mu} \cdot \frac{1}{Q} = 2\pi \cdot \frac{\operatorname{tg}\delta}{\mu}, \quad . \quad (4)$$

in which Q is the so-called quality factor and δ the angle of loss. Thus the expression $R/\mu fL$ appears to differ only by a factor 2 from the quantity $(\tan\delta)/\mu$, which is of importance when determining the effect of an air-gap in the magnetic circuit (the application of an air-gap is an easy means of adjusting the self-induction of a coil). If L is the self-induction of a coil on a closed core of a material with permeability μ, and L' the self-induction obtained by applying an air-gap in the original core, then for the latter core (incl. air-gap) an "effective" permeability μ' can be taken, defined by:

$$\mu' = \frac{L'}{L}\mu.$$

If δ is the loss angle of the original coil and δ' that of the coil with air-gap, it can be deduced with fair approximation that

$$\frac{\operatorname{tg}\delta}{\mu} = \frac{\operatorname{tg}\delta'}{\mu'} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

Strictly speaking $(\tan\delta)/\mu$ is not invariable but

$$\frac{\operatorname{tg}\delta}{\mu(1+\operatorname{tg}^2\delta)-1} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

Since, however, only low values of $\tan\delta$ are of interest (small in comparison to unity), in practical use $\tan^2\delta$ with respect to unity can be ignored. If, moreover, one has to do with values of μ or μ' that are large in comparison to unity, the expression (6) will not differ appreciably from the approximated form (5).

From equation (5) it follows that a reduction of μ to μ' is accompanied by a proportional reduction of $\tan\delta$, that is to say there is a raising of the quality factor (Q). The gain to be obtained in this

[4]) Philips Techn. Review, 8, 57, 1946 (No. 2)

way is all the greater according as the original permeability is higher.

In principle the same applies, it is true, to dust cores, where the high permeability 1 of the insulating layers results in a low effective permeability, but — apart from the more favourable behaviour at high frequencies — the homogeneous "Ferroxcube" material has the advantage over dust cores in three important respects:

1) one is quite free in selecting the air-gap most favourable for each individual case; the dust core, on the other hand, has as it were a "built-in", air-gap, which cannot be made smaller;

highest level at which the material is useful for practical purposes, and for "Ferroxcube" III — the permeability of which is about 1000 — the frequency limit lies at approximately $0.5 \cdot 10^6$ c/sec.

In the case of ferrites the losses are usually formed for the greater part of the residual losses. Investigations carried out with the object of finding materials with a higher frequency limit had therefore to be directed primarily towards lower residual losses. It was found, however, that this could only be attained at the cost of permeability. In many cases it appeared, for instance, that whereas the absence of ferro-ions was favourable in respect

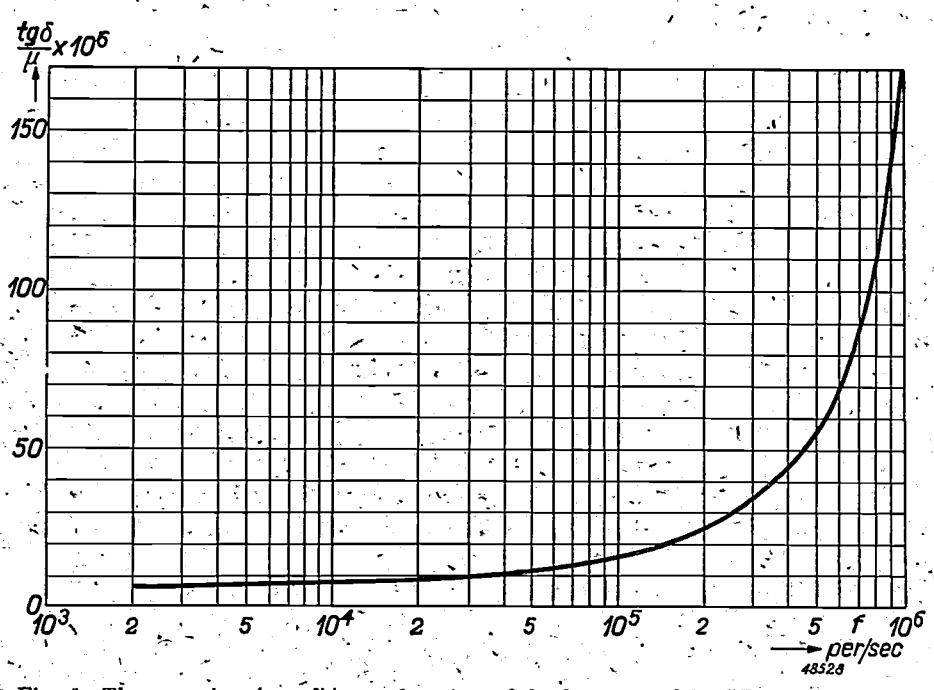


Fig. 1. The quantity $(\tan \delta)/\mu$ as function of the frequency f for "Ferroxcube" III. The permeability of this material is about 1000. The frequency limit where $\tan \delta$ reaches $6 \cdot 10^{-2}$ lies at about $0.5 \cdot 10^6$ c/s.

2) a much more effective magnetic screening can be obtained than with the heterogeneous dust core material (this will be reverted to when dealing with the application of "Ferroxcube");
3) the magnetic induction in the homogeneous "Ferroxcube" is much more uniformly distributed than in the grains of a powder core; this is favourable for arriving at low hysteresis losses.

Loss angle as function of frequency

In *fig. 1* the quantity $(\tan \delta)/\mu$ of "Ferroxcube" III is set out as a function of the frequency. It will be noted that with rising frequency it first increases gradually and later at a faster rate. The frequency at which $\tan \delta$ reaches 6.10^{-2} is regarded as the

to residual losses at high frequencies, for a high permeability their presence was essential. The highest frequency limit so far reached, about $40 \cdot 10^6$ c/sec, was obtained with a nickel-zinc ferrite having an initial permeability of about 50 and containing no ferrous ions.

Shaping and mechanical properties

Before proceeding to deal with some of the applications of "Ferroxcube" we will briefly mention something about its manufacture.

Simple forms such as discs, rings, cubes and the like are made by compressing the powdered base material in the dry state in steel moulds.

Long, cylindrical shapes (rods or tubes of different diameters) are obtained by mixing the powdered

material with a binder to form a plastic mass, which is then forced through an aperture of the size required.

The objects turned out in this manner are then annealed in an electric oven to temperatures between 1000 °C and 1400 °C, the binding agent used for plasticising thereby evaporating and the chemical reaction taking place through diffusion.

Fig. 2. Some specimens showing the shapes in which component parts of "Ferroxcube" can be made. The disc, ring and E-shaped pieces are moulded by dry compression of the basic material in powder form; the rod and tube-shaped pieces are formed by adding a binder to the powder and pressing the mass through a circular opening.

The heated product has a greater hardness but by grinding and polishing can be brought exactly to the desired dimensions. *Fig. 2* shows some samples.

Practical applications

Band-pass coils for carrier wave telephony

One of the most important components of a carrier wave telephony installation is the electric band-pass filter [5]), the purpose of which is to avoid cross-talk. Only very low losses are admissible in the constituent parts of these filters (coils and condensers). "Ferroxcube" III is eminently suitable as magnetic core material for the coils. Furthermore the constancy of the effective permeability with time and under temperature fluctuations has been proved to satisfy the high demands of practice. *Fig. 3* shows a filter coil made of "Ferroxcube". The coil proper is enveloped by a sort of box of "Ferroxcube" consisting of two discs and a ring, whilst inside the coil is a small cylinder of the same material with an air-gap at each end, thus reducing the original permeability from about 1000 to an effective permeability $\mu = 35$. The box-like construction ensures a good magnetic screening, enabling a number of coils to be built together into a compact unit without risk of mutual inductance (cross-talk) - see *fig. 4*. The assembled coil has a quality factor $Q = 600$ (at a frequency of 60 kc/sec) and a volume of 44 cm³, both of which figures compare very favourably with those of a coil of the older construction (*fig. 5*) with $Q = 220$ and a volume of 210 cm³. When comparing these figures it is to be borne in mind that with properly constructed coils Q is proportional to the linear dimensions.

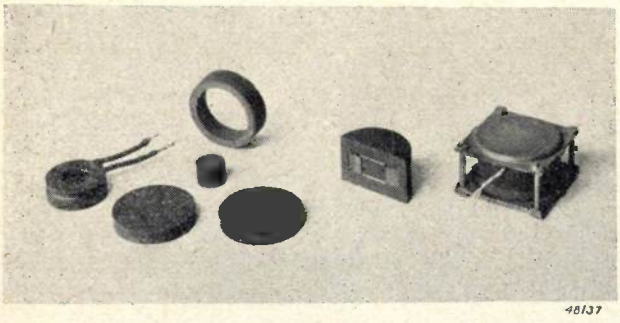


Fig. 3. Band-pass filter coil for carrier wave telephony. Extreme left: the winding and next to that the core and jacket parts made of "Ferroxcube". Farther to the right: a section of the jacket, core and winding: the two air-gaps have been filled with a small white plate to show them up better. Extreme right; the complete coil.

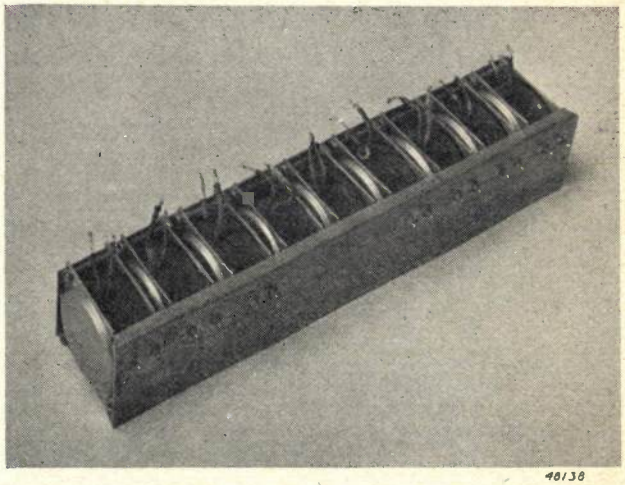


Fig. 4. The coils required for one band-pass filter (here nine) can be built together into a compact whole. Thanks to the good screening provided by the "Ferroxcube" jacket (see fig. 3) there is no perceptible mutual inductance.

Band-pass filter coil for the I.F. part of receiving sets

In radio sets built on the superheterodyne principle a signal of a fixed frequency, called the intermediate frequency, is obtained by mixing the incoming signal with an auxiliary signal generated in the set itself. A filter provides for a band of the desired width to be passed through that lies on either side of this intermediate frequency.

By using "Ferroxcube" the coils for these band-pass filters can be made of much smaller volume, as may be seen from *fig. 6*; this is most welcome to

[5]) Philips Techn. Review, 7, 104, 1942.

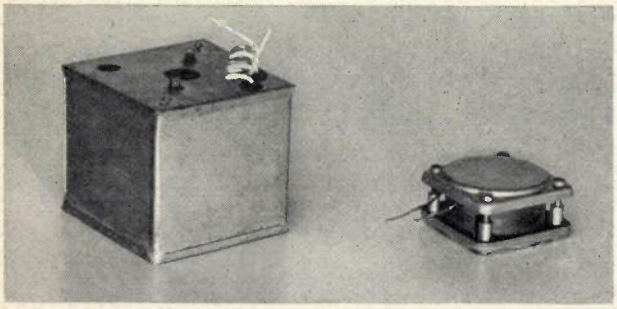


Fig. 5. Band-pass filter coils for carrier wave telephony. Left: screening cage containing a coil of the old construction with the core composed of filaments of nickel-iron and the jacket of a "powder-core" material. The screening is necessary owing to the low effective permeability of the powder core material. Quality factor $Q = 220$ at 60 kc/sec, volume 210 cm³. Right: coil with core and jacket of "Ferroxcube" (*cf.* figs. 3 and 4); $Q = 600$ at 60 kc/sec, volume 44 cm³.

set-makers aiming at producing the smallest possible sets.

Other uses

It may be asked whether "Ferroxcube" is also suitable for the cores of transformers connected to lighting or power mains. Generally speaking the answer is in the negative, because ferrites are saturated already at low inductances, of say 2500 gauss, and consequently the core diameter or the number of windings, for instance, would have to be at least five times as large as that required for the usual transformers with iron core, working with $B_{max} =$ appr. 12 000 gauss.

There are, however, plenty of possibilities in other directions, such as:

1) the tuning of oscillatory circuits, either by sliding a "Ferroxcube" core inside the coil (or coils) or by adjusting the effective permeability by means of d.c. magnetisation;
2) applications in the technics of frequency and impulse modulation;
3) for obtaining a more concentrated magnetic field in an object to be heated by induction in a magnetic alternating field of high frequency.

It should be evident that this new material opens up an extensive field for investigation.

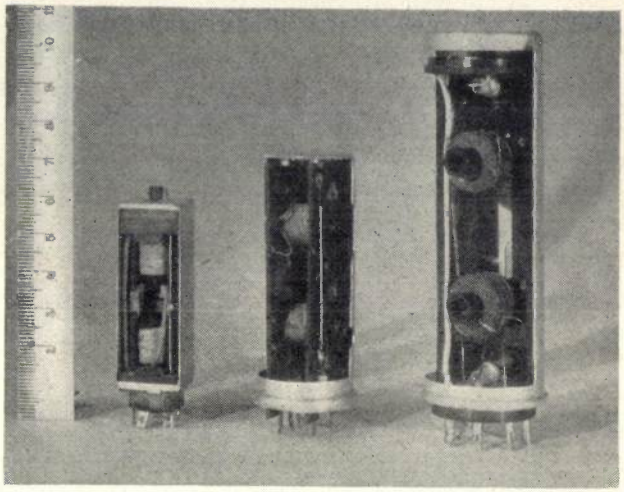


Fig. 6. Coil systems of the I.F. band-pass filter of a radio receiving set. The right-hand coil system contains no ferromagnetic material; the smaller dimensions of the other two are due to the use of "Ferroxcube". The left-hand coil system is a further development of the middle one. All three coil systems are used for the same frequency and are of the same quality. Part of the screening has been cut away to show the inside.

A MAGNETRON FOR D.C. VOLTAGE AMPLIFICATION

by H. B. G. CASIMIR. 621.385.16:621.385.831

In D.C. voltage amplification an amplifier valve with grid control (for instance a triode) often cannot be used because the input and output circuits can never be kept galvanically separated. Instead a magnetron can be used, *i.e.* a diode controlled by a magnetic field. The input circuit of a magnetron, being that circuit of which the solenoid setting up the magnetic field forms a part, is galvanically entirely separate from the anode circuit. In this article the magnetron is discussed as an amplifier valve. It is found that the magnetic sensitivity of a magnetron, which plays a part analogous to the amplification factor of a triode, can be considerably increased by introducing a grid at cathode potential close to the anode. As an example the employment of such a magnetron with grid is described in an X-ray apparatus where a regulator triode, at a very high voltage, had to be fed by a weak D.C. voltage signal given by a regulator at earth potential.

Triode and magnetron as amplifier valves

The simplest method of amplifying an electric signal is based in many cases on the use of a triode. The amplifying action of the triode is manifested in the fact that small changes in the grid voltage caused by the signal result in large changes in the anode voltage. These changes in the anode voltage are the direct result of the changes in the anode current, which is here controlled by an electric field due to the signal. However, for a long time there have been electronic valves in which the anode current is controlled by a magnetic field due to the signal and which can also be used as amplifier valves. While the triode may be considered as a diode with a grid between cathode and anode, the magnetron [1]) is a diode which is placed in the magnetic field of a solenoid whose axis is perpendicular to the direction of the electron current (if the diode is for instance cylindrical the axis of the solenoid coincides with the cathode). Changes in the current in the solenoid then cause changes in the intensity of the magnetic field, which again affects the motion of the electrons between the cathode and anode, whereby with suitably chosen conditions changes in the intensity of the anode current may occur. We shall discuss this in more detail below; it will be found that for a magnetron as well as for a triode quantities such as slope and internal resistance can be defined and that, moreover, a quantity can be introduced which is comparable with the amplification factor in a triode, namely the so-called magnetic sensitivity.

For the amplification of electrical signals the triode (or tetrode or pentode) is nearly always used, and not the magnetron. There are, however, cases where the triode and also every other valve with grid control is inadequate and where the magnetron is the ideal valve. This is connected with the fact that in every valve with grid control the average voltage level to be used for the grid compared with the voltage level of the anode and of the cathode is fixed within rather narrow limits by the characteristics of the valve. In particular the voltage difference between the anode and the grid of a triode in the commonly used types cannot amount to more than a few hundred volts. As long as it is desired to amplify an A. C. voltage this involves no essential limitation of the usefulness of valves with grid control, even when the voltage level of the amplified signal has to lie much higher than that of the original signal. By making use of a transformer or of condensers provision can be made for the voltage difference between the anode and the grid not to exceed the above-mentioned permissible value. In the case of D.C. voltage amplification this is impracticable. It is indeed possible to apply a voltage difference between the input and output circuits which is within the limits permitted by the characteristic, but there is always a galvanic coupling between the two circuits which may be an objection. On the other hand in the case of a magnetron the input circuit (the solenoid) can be kept galvanically absolutely separate from the output circuit (the anode circuit), so that even with D.C. voltage amplification the above assumed difference between the voltage levels of the incoming and outgoing signals presents no difficulty.

A concrete example in which the magnetron has proved serviceable for the above reason occurred in an X-ray apparatus. At the conclusion of this article several particulars of this will be given.

[1]) It should be noted that the word "magnetron" is used in its original significance, namely that of a magnetically controlled diode, while at the present time "magnetron" usually means an oscillator valve for very short waves, where the electrons move in a constant magnetic field. *Cf.* G. Heller, Philips Techn. Rev. 4, 201, 1939.

Principle of the magnetron

The "classical" magnetron (see *fig. 1*) is a cylindrical diode (anode voltage V_a, anode radius r_a), situated in an axial magnetic field H generated by a solenoid.

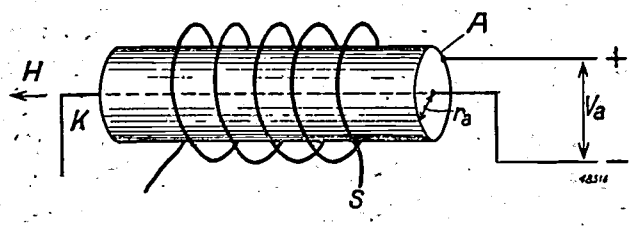


Fig. 1. Diagram showing the principle of a magnetron. *A* anode, *K* cathode, *S* solenoid. The arrow indicates the direction of the magnetic field *H*.

When only the electric field is present the electrons emitted by the cathode will describe radial trajectories to the anode; the electric field is radially directed. The velocity v of an electron at a given point is determined by the voltage V prevailing at that point (calculated with respect to the cathode). If m is the mass and e the charge of the electron

$$\tfrac{1}{2} mv^2 = eV \quad \ldots\ldots\ldots \quad (1)$$

If we consider only the magnetic field, an electron with velocity v experiences a force prependicular to the velocity and to the direction of the field and equal to evH. This results in the trajectory becoming curved; the kinetic energy of the electron undergoes no change due to the magnetic field since the force is perpendicular to its trajectory. The energy equation (1) thus does not change when in addition to the electric field we also introduce a magnetic field.

The force evH is equal to the centrifugal force mv^2/r, where r is the radius of curvature of the trajectory:

$$\frac{mv^2}{r} = evH \quad \ldots\ldots\ldots \quad (2)$$

If now an electric field and a magnetic field are both applied, the electrons describe curved trajectories whose radius of curvature at every point can be calculated as a function of H and V by eliminating the velocity v from (1) and (2):

$$r = \frac{1}{H}\sqrt{\frac{2\,mV}{e}} \quad \ldots\ldots\ldots \quad (3)$$

or, substituting the values of m and e and expressing r in cm, H in gauss and V in volts:

$$r = \frac{3{,}37}{H}\sqrt{V} \quad \ldots\ldots\ldots \quad (3a)$$

With a given electrical voltage V we now vary the magnetic field strength H. At a small value of H the electrons move almost radially along only slightly curved paths from the cathode to the anode (see *fig. 2a*), since with a small value of H the value of r is very large, according to formula (3*a*). An anode current then flows, and as H increases it remains constant as long as the electrons emitted by the cathode reach the anode. As the value of H increases r decreases, and the paths of the electrons become more and more curved until, beginning with a critical field strength H_k, the electrons no longer reach the anode (see fig. 2*b* and 2*c*); there is no longer any anode current and the magnetron is "closed". If the cathode is sufficiently thin this critical field strength H_k can easily be calculated. It may then be assumed that the potential is constant in the larger part of the cylindrical space

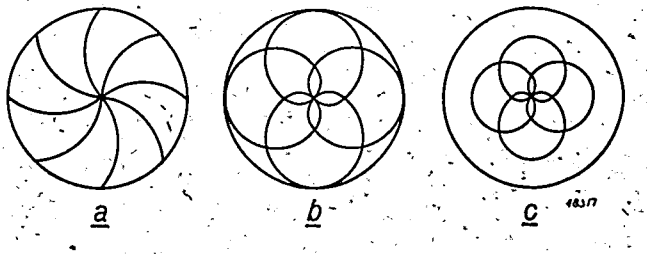


Fig. 2. Electron trajectories in a magnetron at different magnetic fields H: *a*) when $H < H_k$ anode current flows; *b*) $H = H_k$, the limiting case where the paths of the electrons just touch the anode; *c*) when $H > H_k$ the elcctrons describe closed trajectories and do not reach the anode.

enclosed by the anode, and that it is equal to the anode voltage V_a. The decrease in the potential toward the cathode occurs mainly in its immediate vicinity.

The electrons will then, according to (3*a*), describe paths with almost constant radius of curvature, thus approximately circles. The diameter $2r$ of these circles at the critical field strength H_k is equal to the radius r_a of the anode. With (3*a*) this gives

$$H_k = \frac{6{,}74}{r_a}\sqrt{V_a} \quad \ldots\ldots\ldots \quad (4)$$

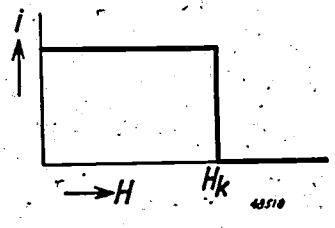


Fig. 3. "Magnetic" characteristic of an ideal magnetron, *i.e.* anode current as a function of the magnetic field strength at a single given value of V_a. At another value of V_a the vertical part of the characteristic is shifted towards the right or left.

The magnetic characteristic $i = i(H)$ at a single given value of V_a is shown in *fig. 3*. For $H > H_k$ the anode current is constant, all electrons reaching the anode; for $H > H_k$ the current is "cut off". According to (4) H_k depends on V_a: if V_a increases, H_k also lies at a higher value.

The magnetic sensitivity of the magnetron

As is known, the slope S and the internal resistance R_i of a triode are defined by the equation

$$\Delta i = S\Delta V_g + (1/R_i)\,\Delta V_a, \quad \ldots \quad (5)$$

where Δi, ΔV_g and ΔV_a are the changes, respectively in anode current i, grid voltage V_g and anode voltage V_a. From equation (5) it follows that

$$S = \left(\frac{\partial i}{\partial V_g}\right)_{V_a} \quad \ldots \quad (6)$$

and

$$R_i = 1\Big/\left(\frac{\partial i}{\partial V_a}\right)_{V_g} \quad \ldots \quad (7)$$

Moreover, the amplification factor μ is defined as

$$\mu = -\left(\frac{\partial V_a}{\partial V_g}\right)_i, \quad \ldots \quad (8)$$

or, because of (5),

$$\mu = S R_i \quad \ldots \quad (9)$$

Since in the magnetron the function of the grid voltage is taken over by the magnetic field, the "magnetic" slope and the internal resistance are here defined by analogy with equations (6) and (7) by

$$S = -\left(\frac{\partial i}{\partial \ldots}\right)_{V_a} \quad \ldots \quad (10)$$

and

$$R_i = 1\Big/\left(\frac{\partial i}{\partial V_a}\right)_H \quad \ldots \quad (11)$$

(In the definition (10) of S the minus sign is taken because the anode current i decreases as H increases.)

Equation (9) then defines for the magnetron a quantity which is analogous to the amplification factor in the case of a triode. This quantity will be called the "magnetic sensitivity" of the magnetron. The magnetic sensitivity, however, in contrast to the amplification factor of a triode, is not a dimensionless factor. This is seen immediately when the magnetic sensitivity is represented by the formula analogous to (8):

$$\mu = \left(\frac{\partial V_a}{\partial H}\right)_i \quad \ldots \quad (12)$$

It is now found that equation (9) fails us in the case of an ideal magnetron with a discontinuous magnetic characteristic (fig. 3). Then the electrical characteristic (*fig. 4*) of the magnetron, *i.e.*

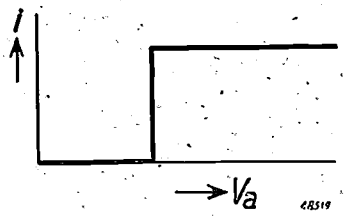


Fig. 4. "Electrical" characteristics of an ideal magnetron, *i.e.* anode current as a function of the anode voltage at a single given value of the magnetic field H.

$i = i(V_a)$, with a given magnetic field H is also discontinuous; at small values of V_a the influence of the magnetic field is predominant, the electrons describe closed trajectories and $i = 0$; at large values of V_a which — see equation (4) — are larger than $(Hr_a/6.74)^2$, the anode draws all the electrons. By reference to figures 3 and 4 it may be seen that in the horizontal part of the characteristics $S = 0$ and $R_i = \infty$, while in the vertical parts $S = \infty$ and $R_i = 0$, so that in the ideal case the magnetic sensitivity defined by (9) cannot be determined from the characteristics.

A similar consideration shows that also equation (12) cannot be used in this ideal case.

In order to calculate the magnetic sensitivity of an ideal magnetron, however, there are two indirect methods. One is to consider a non-ideal magnetron for which the characteristics show a gradual change. The quantities S and R_i are then always finite, so that the product SR_i is no longer indeterminate. It is then possible to calculate $\mu = SR_i$ and ascertain the limiting value of the magnetic sensitivity (thus calculated) when the shape of the characteristics no longer approaches

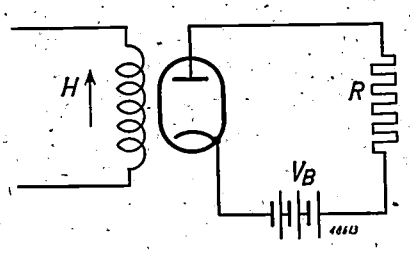


Fig. 5. The magnetron in a circuit for resistance amplification.

the ideal. The other possibility is based upon the consideration of an ideal magnetron included in a resistance amplifier. We shall first follow this second method, since it will also lead us to a knowledge

of the behaviour of the magnetron in its practical applications.

Magnetron in a resistance amplifier

Let us consider the simple resistance amplifier with an ideal magnetron as amplifier valve (*fig. 5*). We wish to study the relation between the voltage V_R over the anode resistance R and the magnetic field H. For that purpose we draw the electrical

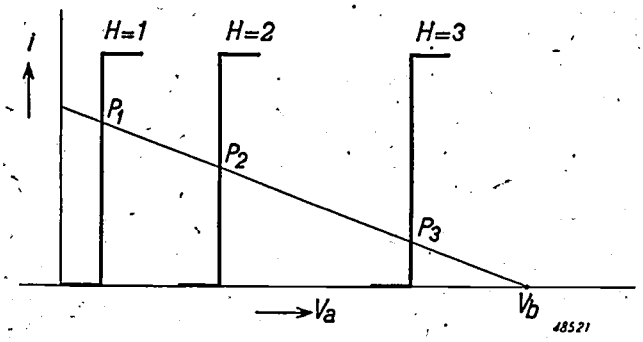


Fig. 6. Vertical part of the electrical characteristic (*cf.* fig. 4) for different values of the magnetic field H (in arbitrary units). The oblique straight line represents the "loading line", *i.e.* the relation between the anode current i and the voltage V_a when the magnetron is included in a circuit of a resistance amplifier with given battery voltage V_b and anode resistance R. The slope of the loading line is inversely proportional to R. The points of intersection P_1, P_2 .... of the loading line with the characteristics determine the current flowing in the anode circuit at given values of V_b and R and different values of H.

characteristics of the magnetron, *i.e.* i as a function of V_a for different values of the magnetic field (*fig. 6*) (this is analogous to the i-V_a diagram with V_g as parameter in the case of a triode). It is hereby assumed that the diode is ideal, *i.e.* that for $H = 0$ the saturation current already flows at very small positive anode voltage. We allow the characteristics to be cut by the "loading line", *i.e.* by the straight line

$$i = \frac{V_b - V_a}{R},$$

which gives the relation between i and V_a at a given battery voltage V_b and resistance R. The points of intersection of the loading line with the characteristic determine the current flowing in the anode circuit at given values of H, V_b and R.

By choosing suitable values of R and V_b, *i.e.* of the slope and position of the loading line, the current can be made to adjust itself on the vertical sections of the characteristics. We also assume that this is actually the case. Then the relation between V_a and H (see equation (4)) is given by:

$$V_a = \left(\frac{r_a}{6{,}74}\right)^2 H^2 \quad . \quad . \quad . \quad . \quad . \quad (13)$$

By elimination of V_a we obtain the following from the last two equations:

$$i = \frac{1}{R}\left[V_b - \left(\frac{r_a}{6{,}74}\right)^2 H^2\right] \quad . \quad . \quad . \quad . \quad (14)$$

The desired relation between V_R and H is thus

$$V_R = V_b - \left(\frac{r_a}{6{,}74}\right)^2 H^2 \quad . \quad . \quad . \quad . \quad (15)$$

The parabola given by equation (15) is reproduced in *fig. 7*. It may be seen that V_R decreases with increasing H and for $H_b = 6.74\sqrt{V_b}/r_a$ becomes zero. The slope of the parabola may serve as a measure of the amplifying effect of the resistance amplifier considered and is analogous to what in a triode is called the "amplification" of a resistance amplifier.

In the case of a triode — as is known — the amplification becomes equal to the amplification factor when the internal resistance R_i is small compared with the anode resistance R. By analogy to this the slope in question becomes equal to the above-defined magnetic sensitivity, since for the ideal magnetron, in the vertical parts of the characteristics of fig. 6, $R_i = 0$. One thus obtains

$$\mu = -\frac{dV_R}{dH} \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

(since dV_R/dH is always negative — compare fig. 7 — in equation (16) the minus sign is taken). From this it now follows that, according to (15) and (13)

$$\mu = 2\left(\frac{r_a}{6{,}74}\right)^2 H = \frac{r_a}{3{,}37}\sqrt{V_a} \quad . \quad . \quad (17)$$

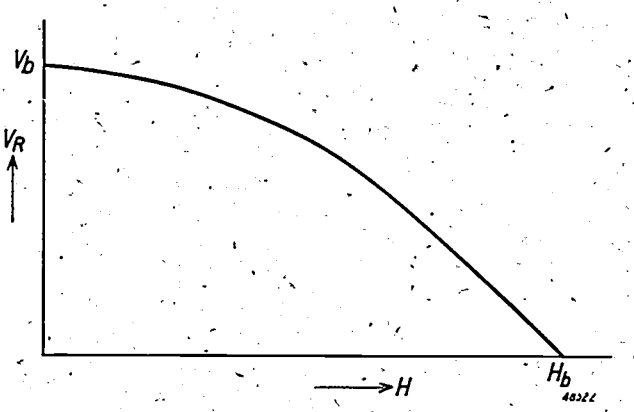


Fig. 7. Voltage V_R over the anode resistance as a function of the magnetic field H for a resistance amplifier with an ideal magnetron. For $H = 0$, V_R is equal to the battery voltage V_b; for $H = H_b = 6.74\sqrt{V_b}/r_a$ (r_a radius of the anode). $V_R = 0$. The curve is a parabola with slope equal to the magnetic sensitivity μ of the magnetron.

We shall find this result again by the other method mentioned above. From equestion (17) it is found that the magnetic sensitivity increases proportionally with the strength of the magnetic field.

The maximum value of the magnetic sensitivity at a given battery voltage is

$$\mu_{max} = \frac{r_a}{3{,}37}\sqrt{V_b}.$$

Magnetron with gradually varying characteristic

Actually a magnetron is never ideal. Since the electric field is not strictly radially and axially symmetrical, and also due to the fact that the vacuum in the valve is not perfect, the characteristics are not discontinuous (*fig. 8*). Thus, for

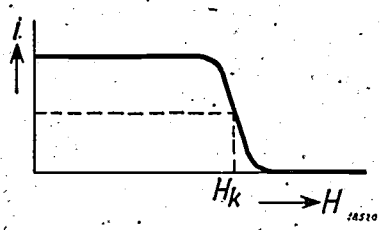


Fig. 8. Gradually varying magnetic characteristic (compare with fig. 3). H_k is here defined as that value of H where $i(H_k) = {}^1/_2\, i(0)$.

example, the electric field at the ends of the anode may be too weak to draw the electrons to the anode, while in the middle elcctrons *do* strike the anode. The shape of the characteristics is thus less sharp, since the anode current does not become zero at one definite value of V or H.

In the case of a magnetron with gradually changing characteristics the product $\mu = SR_i$ is no longer indeterminate. We shall now calculate it on the assumption that the characteristics, while not discontinuous, still have a steep slope.

We may then define H_k as the field strength at which the current has decreased to one half: $i(H_k) = {}^1/_2\, i(0)$. For H_k thus defined equation (4) will remain approximately valid. For the neighbourhood of H_k we may write, in approximation, according to equation (10):

$$i(H) = i(H_k) - (H - H_k)\, S(H_k),$$

and from this, since

$$1/R_i = \left(\frac{\partial i}{\partial V_a}\right)_{II} = \left(\frac{\partial i}{\partial H_k}\right)_{II} \frac{dH_k}{dV_a},$$

it follows that

$$1/R_i = \left\{\left(\frac{\partial i(H_k)}{\partial H_k}\right)_{II} + S(H_k) - (H - H_k)\left(\frac{\partial S}{\partial H_k}\right)_{II}\right\} \frac{dH_k}{dV_a}.$$

In general the first and last terms of the polynomial between the brackets are small compared with $S(H_k)$.

An estimation of the first term can be made on the basis of the following consideration. If we are working in the saturation region, $\partial i(0)/\partial H_k = 0$. If we are not working in the saturation region, since according to Langmuir's formula $i(0) \sim V_a^{3/2}$, $i(H_k) \sim V_k^{3/2}$ also, and with (4) this gives: $i(\ \ _k) \sim 1.\, k^3$, from which it follows that

$$\left(\frac{\partial i(H_k)}{\partial H_k}\right)_{II} = \frac{3\, i(H_k)}{H_k}.$$

If now the slope is so large that

$$\left|\left(\frac{\partial i}{\partial H}\right)_{V_a}\right| \gg \frac{3\, i(H_k)}{H_k},$$

or, according to *fig. 9* that $\tan\gamma \gg \tan\beta$ we may indeed ignore the first term between the brackets. From fig. 9 it appears that only at large values of β (*i.e.* small values of H_k and V_a) is this not the case.

The last term between the brackets is zero for $H = H_k$ and remains small in the vicinity of that point.

For not too small values of V_a and H_k it thus follows that

$$1/R_i = S(H_k)\frac{dH_k}{dV_a} \quad . \quad . \quad . \quad . \quad (18)$$

and the magnetic sensitivity is then (see equation (9)):

$$\mu = S(H_k)\, R_i = \frac{dV_a}{dH_k} \quad . \quad . \quad . \quad . \quad (19)$$

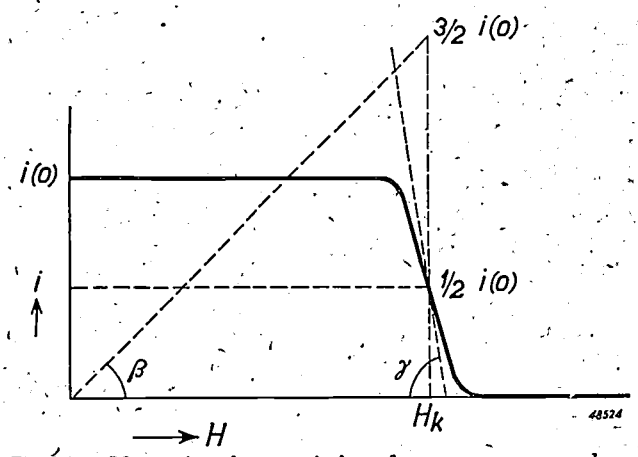


Fig. 9. Magnetic characteristic of a magnetron; when $\tan\beta \ll \tan\gamma$ formule (20) for the magnetic sensitivity is valid.

Upon comparison of (19) with (8) it is found that dV_a/dH_k in the case of the magnetron is similar in signifiance to $-(\delta V_a/\delta V_g)_i$ in the case of the triode.

With the help of equation (4) it is now possible to calculate dV_a/dH_k, so that one finally obtains

$$\mu = 2\left(\frac{r_a}{6{,}74}\right)^2 H_k = \frac{r_a}{3{,}37}\sqrt{V_a} \quad . \quad . \quad . \quad (20)$$

The magnetic sensitivity is thus independent of the slope of the magnetic characteristic. From this it follows in particular that equation (20) is also valid for the limiting case of an ideal magnetron with a discontinuous characteristic. This formula is indeed identical with the formula which we previously derived for the ideal magnetron in a resistance amplifier (see equation (17)). The only advantage of a very steep magnetic characteristic is that then, according to (9), P_i is small. Since in a resistance amplifier $R_i \ll R_a$ (R_a anode resistance) in order to reach the maximum amplification a relatively small value of R_a will be sufficient.

Magnetron with grid

If it is desired to increase the theoretical magnetic sensitivity we must, according to equation

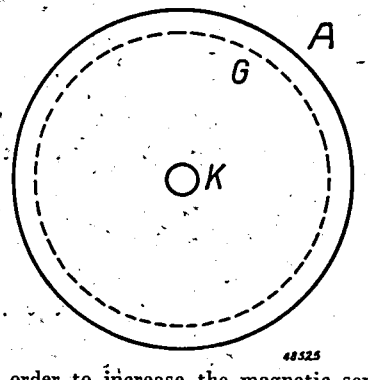


Fig. 10. In order to increase the magnetic sensitivity μ, a grid is introduced into the cylindrical magnetron close to the anode. K cathode, G grid, A anode.

(2), either increase the anode voltage or the radius. The first has the objection that the magnetic field has also to be taken larger: one must work with a large constant field upon which the variations are superposed. The increase in the radius quickly leads to unwieldy dimensions.

It is possible to attain a much higher sensitivity, without encountering the two difficulties mentioned, by placing close to the anode (*fig. 10*) a grid which is at approximately cathode potential. The influence of the anode voltage on the motion of the electrons between the cathode and the grid is then much smaller, so that the magnetic field has relatively more effect. Instead of equation (4), the following then holds:

$$H_k = \frac{6{,}74}{r_a}\sqrt{V_{eff}}, \quad \ldots \quad (21)$$

where V_{eff} is an effective voltage which — entirely analogous to the so-called effective grid voltage in a triode — can be represented by

$$V_{eff} = \frac{V_a}{G} + V_g; \quad \ldots \quad (22)$$

the constant G in the triode with screen grid being equal to the amplification factor. For the sake of simplicity we set $V_g = 0$. Equation (21), because of (22), then becomes

$$H_k = \frac{1}{\sqrt{G}}\,\frac{6{,}74}{r_a}\sqrt{V_a} \quad \ldots \quad (23)$$

We shall now consider the magnetron with grid included in a resistance amplifier. Proceeding in the same way as in the case of an ideal magnetron with no grid, we arrive at the following relation between V_R and H:

$$V_R = V_b - G\left(\frac{r_a}{6{,}74}\right)^2 H^2. \quad \ldots \quad (24)$$

From *fig. 11* we see that the curve giving this relation is steeper than the analoguous curve for the magnetron with no grid, which corresponds to equation (15). The magnetic sensitivity has thus indeed become larger, namely by a factor G with the same magnetic field and the same radius of the anode, as appears upon differentiation of equation (24):

$$\mu = -\frac{dV_R}{dH} = 2G\left(\frac{r_a}{6{,}74}\right)^2 H. \quad \ldots \quad (25)$$

The anode voltage must now be higher, but that is often no objection. The value of G is limited by the requirement that the value of V_a/G must be large enough to draw a reasonable current from the cathode. As to the optimum dimensions, these depend upon the employment of the valve. Compared with the simple magnetron, however, considerable gain can nearly always be obtained for instance with a factor $G = 20$.

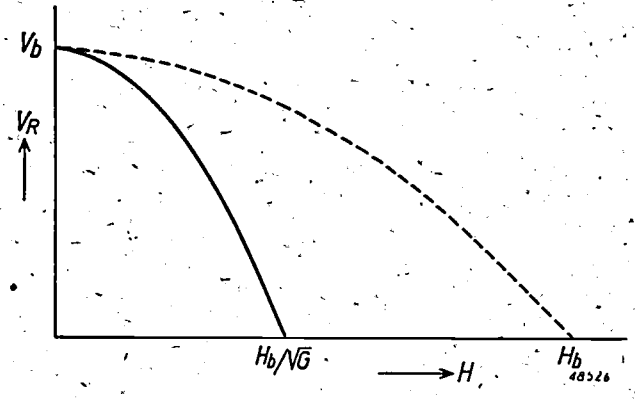


Fig. 11. Voltage V_R across the anode resistance as a function of the magnetic field H for a resistance amplifier, *1*) with a grid-magnetron (continuous line), *2*) with an ideal magnetron with no grid (dotted line, identical with the curve of fig. 4). Both curves are parabolas. Since the magnetic sensitivity μ is equal to the slope of the parabola, at equal values of H, μ is larger in the first case than in the second.

It may still be noted that equation (25) can also be derived on the basis of the consideration of a magnetron with gradually (but steeply) changing characteristics.

Magnetron with grid as D.C. voltage amplifier in an X-ray apparatus

Such a magnetron with grid has been employed in the above-mentioned X-ray apparatus. The problem encountered in designing the apparatus in question came down to the question of how a switching and regulating arrangement at a very high voltage (about 100 kV) could be operated by a weak D.C. voltage signal given by a switching and regulating arrangement at earth potential. Various solutions to this problem are of course possible, but a discussion of all these is outside the scope of this article.

It need only be noted that the use of an amplifier valve with grid control was out of the question because of the reasons mentioned above, at least if it was desired to avoid the conversion of D.C. into A.C. and *vice versa*. On the other hand the employment of the grid-magnetron was found to be very suitable. A simplified circuit diagram of the installation is reproduced in *fig. 12*. It may be seen that the voltage level of the solenoid S is completely separated from the very high voltage level of the rest of the installation; the latter can be varied at will without affecting the circuit in which the solenoid is included. Nevertheless, the grid voltage of the triode T changes as soon as a change occurs in the current in the solenoid. The grid voltage in question is equal to the voltage over the resistance R, and the latter voltage depends finally on the current in the solenoid.

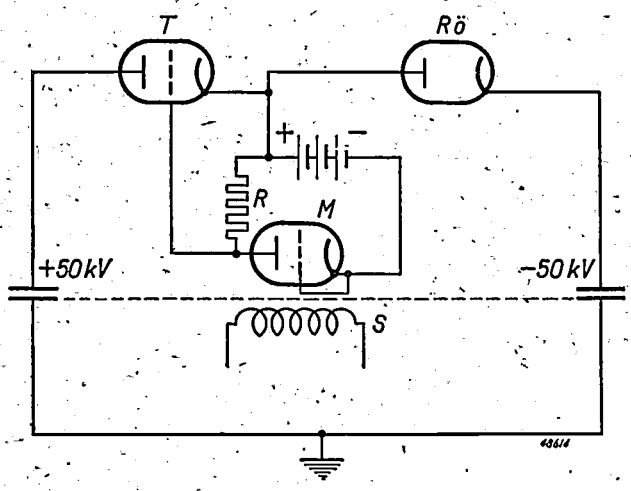


Fig. 12. Simplified circuit diagram of an X-ray apparatus in which the grid magnetron serves as amplifier valve. T triode regulating the voltage on the X-ray tube *Rö*. The voltage difference between grid and cathode of T is equal to the voltage across the resistance R; the variations in this latter voltage are due to the amplification of the voltage variations in the circuit in which the solenoid causing the magnetic field is included. This magnetic field regulates the current in the grid magnetron M. When no current flows through the solenoid, the magnetron is "open" and the triode "closed". The horizontal dotted line indicates symbolically that the voltage level below this line (*i.e.* the voltage level of the circuit of the solenoid) is entirely independent of the voltage level above it.

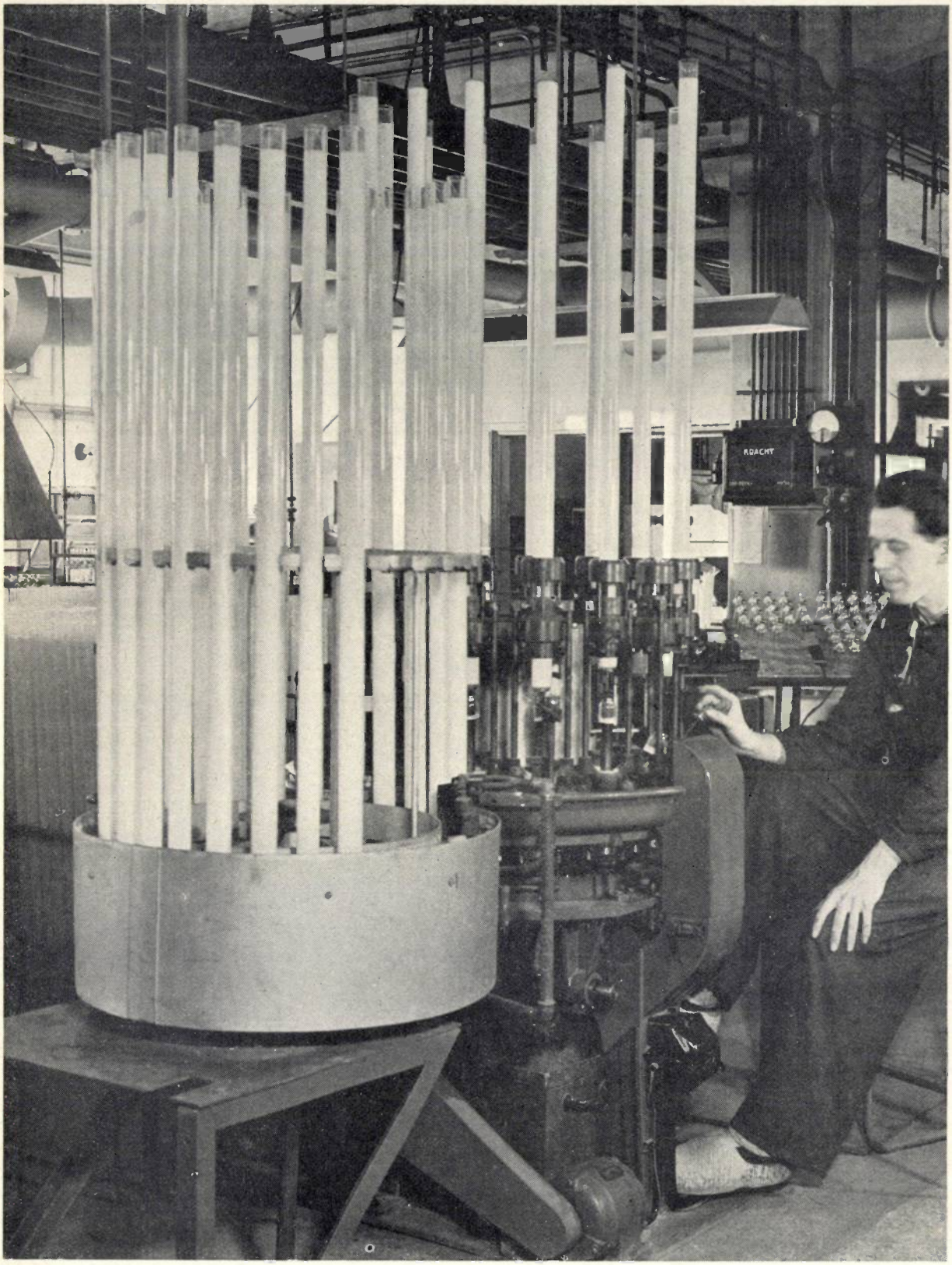
We may regard the triode T as a variable resistance whose value R_T is determined in part by the current in the solenoid. When no current flows through the solenoid the magnetic field is zero, the magnetron is "open" and the triode "closed", *i.e.* $R_T = \infty$, because the grid voltage is then equal to the full (negative) battery voltage V_b. If on the other hand the current through the solenoid is so large that the strength of the magnetic field lies above the critical value $H_b/\sqrt{G}$ (*cf.* fig. 11), then the magnetron is "closed"; as a result the voltage over R, and thus also the grid voltage, is zero: the triode is entirely "open", *i.e.* $R_T = 0$. At currents in the solenoid which lie between the limiting values mentioned, $0 < R_T < \infty$.

If the triode is used as relay it means that only the values $R_T = 0$ and $R_T = \infty$ are used. In the X-ray apparatus of fig. 12, however, the triode also served as regulator valve with which the voltage on the X-ray tube *Rö* was regulated; R_T then followed the changes in the current in the solenoid, between the limits $R_T = 0$ and $R_T = \infty$ [2]).

The advantage of a magnetron with grid was thereby manifested in the fact that in both cases — triode as regulator valve and as relay — much weaker currents were needed in the solenoid to reach a given change in R_T than if a magnetron with no grid had been used. If the triode plays only the part of a relay, the solenoid current necessary to "open" the triode completely (change from $R_T = \infty$ to $R_T = 0$) is $\sqrt{G}$ times as small as in the case of the magnetron with no grid.

[2]) If the saturation current flows through the X-ray tube, not too large changes in R_T will have no influence on the current in the X-ray tube, only the voltage being affected.

AUTOMATIC SEALING- IN OF THE ELECTRODES IN TUBILAR FLUORESCENT LAMPS



The tube is placed over a preheated, mounted electrode on a rotating machine (centre) which automatically seals the two together by gradually increased heating of the rims. The tubes are then transferred to a cooling drum (left) in which the realed ends are gradually cooled down. The same process is repeated with the tubes inverted, for sealing in a mounted electrode at the other end, after which the tubes are ready for evacuation.

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

Reprints of the majority of these papers can be obtained on application to the Administration of the Research Laboratory, Kastanjelaan, Eindhoven, Netherlands. Those papers of which no reprints are available in sufficient number, are marked with an asterisk.

1697: N. G. de Bruyn: A combinatorial problem (Proc. Kon. Ned. Akad. Wetenschappen Amsterdam **49**, 758-764, 1946).

A P_n-cycle is defined as an ordered cycle of 2^n digits 0 or 1 (i.e. a series of such digits placed on the circumference of a circle), such that the 2^n possible sets of n consecutive digits of that cycle are all different (as a consequence, any ordered set of n digits 0 or 1 occurs exactly once in that cycle). Posthumus, studying these cycles in connection with a practical problem of telecommunication, was led to the conjecture, that the number of P_n cycles be equal to 2 to the power $2^{n-1}-n$. In this paper this conjecture is proved to be correct As a consequence a theorem concerning a special type of networks is stated and proved and another application of the theorem is mentioned.

1698: D. Polder and J. H. van Santen: The effective permeability of mixtures of solids (Physica, 's-Grav. **12**, 257-270, 1946).

For several purposes it is of importance to know how the dielectric or magnetic permeability of a mixture of substances, which show mutually different permeabilities in pure state, depends on the composition of the mixture. In this paper the case is dealt with, where the different substances are powders of solids, packed in a medium like oil or wax or air. The individual particles or holes are assumed to be ellipsoidal. The calculation is based on an approximation, which in the case of spherical particles proves to give results identical with those obtained by Böttcher's method. The effect of the shape of the ellipsoids is discussed. Considerations are given concerning the nature of the approximations used in this paper and in other theories on this subject.

1699: M. J. O. Strutt and A. van der Ziel: Signal-noise ratio at v.h.f. (Wireless Engeneer **23**, 241-249, 1946).

The two basic concepts of correllation between fluctuation currents and voltages and of the decompositions of fluctuating quantities into singly periodic components are discussed briefly. The noise ratio of a grounded-cathode amplifier is evaluated and the conditions for minimum noise ratio are stated. Experimental evidence is shown to confirm the theoretical results. Further grounded-grid amplifier stages and velocity modulation valves are considered and the requirements as to valve construction (uniform electron paths and low dielectric and other losses) are dealt with. It is shown that the purposed reduction of noise ratio applies also to wideband reception. In order to compensate the loss or gain incurred by the reduction of noise ratio suitable feed-back may be applied. The interrelation of noise figures introduced by various authors and the present noise ratio are discussed.

1700: B. D. H. Tellegen: Het bepalen van de integratieconstanten bij de berekening van in- en uitschakelverschijnselen (T. Ned. Radiogenootschap II, 173-188, 1946). (The evaluation of the integration constants occuring in the computation of transient phenomena).

A network is considered containing a source of pressure v, under the influence of which a current i flows in a certain branch. A method is given to calculate from the differential equation connecting i and v the discontinuities in i and its derivatives resulting from discontinuities in v. and its derivatives.

1701: H. A. Klasens: Transfer of energy between centres in zinc sulphide phosphors (Nature, **158**, 306, 1946).

The transfer of energy between centres in zinc sulphide phosphors containing two kinds of centres (*e.g.* blue and green) in explained by the supposition that after bringing an electron from one centre (*e.g.* blue) to an excited state (by lifting it in the empty so-called conduction band of energy), the hole in the blue centre may be transfered to a filled band and thus may travel to a green centre. The reverse process (green to blue) is supposed to have a negligible probability. The above hypothesis is expressed in the form of two simultaneous differential equations giving the number of excited centres as a function of time.

From these equations formulae may be derived for the intensity ratio of the blue and the green emission during irradiation as a function of the irradiation intensity and the temperature and for the temperature — dependence of the emission of

e.g. the blue centres. The effect of "killers" on the after glow (green emission) is explained on similar lines.

1702: J. Boeke: Ontwikkeling van de electrische meetapparaten in de chemie gedurende de oorlog. (Chem. Weekblad 42, 230-273, 1946) (Development of electrical measuring instruments for chemical purposes during the war).

This article contains a survey of recent improvements in electrical methods as used in chemistry, especially in the following fields: Electrochemistry, dielectric constants, measurement of moisture, spectrometry (absorption and emission), chemical analysis, industrial apparatus.

1703: W. de Groot: The influence of irradiation with light on the dielectric properties of Zn S phosphors (Physica 12, 402-404, 1946).

Experiments are described, whereby the influence of irradiation with light on the light emission and the dielectric behaviour of ZnS phosphors are observed simultaneously. It is proved that the change in capacity ΔC of a condenser containing ZnS, and the decay of this change after stopping irradiation are much slower than the rise and decay of the light emission. Experiments are described, which make probable that the change in dielectric behaviour is due to free electrons and not to trapped electrons. Some phenomena, such as a steep initial decay of the ΔC-effects with a phosphor containing a killer, remain unexplained.

1704*: D. J. Bouma: Kleuren en kleurindrukken (Philips technische bibliotheek; uitg. Meulenhoff & Co N.V. Amsterdam 1946, 320 pages, 113 fig.) (Colours and colour sensations).

In this book (an English translation of which is forthcoming) a survey in simple language is given of the fundamentals of colour science and the methods of colorimetry. In the exposition of the basic facts ample use is made of colour space. Additional chapters deal with colour vision, the historical development of colour science, thresholds, the estimation of colour differences and related problems, technical and scientific applications. The appendix comprises 15 tables and an extensive reference list.

1705*: H. A. Klasens and M. E. Wise: Decay of zincsulfide type phosphors (Nature, London, 158, 483, 1946).

The bimolecular decay of zincsulfide phosphors, assuming retrapping in the metastable phosphorescense centra, is described by a second order non-linear differential equation. In the special case where the two unknown constants (α = retrapping coefficient, β = bimolecular recombination constant) are equal, the equation can be solved and curves may be drawn for the intensity *vs.* time at different temperatures and for different intensities of excitation. These curves generally display a steep initial decay due to the bimolecular recombination and a more or less prolonged tail due to the phosphorescence mechanism.

1706: C. J. Bouwkamp: A note on singularities occuring at sharp edges in electromagnetic diffraction theory. (Physica, 's-Grav. 12, 467-474, 1946).

Wave fuctions u describing the diffraction by plane screens can be divided in two classes according as u or $\partial u/\partial n$ vanishes at the surface of the screen. Differentiation with respect to the coordinate in the n(ormal)-direction alters the character, whilst tangential differention does not change the class. Another effect (Rayleigh) of differentation is the appearance of singularities at the edge of the screen. This is demonstrated explicitly in case of Sommerfeld's solution of the electromagnetic diffraction by a half-plane. Typical difficulties in electromagnetic diffraction theory are mentioned. Möglich's solution of the electromagnetic diffraction by a circular screen is shown to be erroneous.

1707: W. Elenbaas: The hypothesis of minimum voltage in the theory of the arc. (Physica, 's-Grav. 12, 491-498, 1946).

The author criticizes the explanation of the contraction of the high pressure discharge in the vicinity of the cathode making use of the theory of minimum voltage. Although this theory and the exact theory may be put in such a form, that they yield very similar fomulae, there is a discrepancy as regards the cross section of the arc by a factor which at ordinary current densities may be as high as 50 and which moreover depends on the temperature. At very high current densities (almost complete ionisation) the principle is considered to give good results.

SURVEY OF SUBJECTS VOLUMES 1–8

The number in heavy print beside each subject indicates the volume, and the following number the page. Volumes 1—7 correspond to the years 1936—1942. From July 1942 to December 1945 the publication of this journal had to be stopped. The first number of volume 8 appeared in January 1946.

REGISTER OF ARTICLES ARRANGED ACCORDING TO AUTHORS VOLUMES 1 TO 8

The figures in heavy type indicate the volume and those in normal type the pages on which the articles are to be found.